Analysis of bacterial Mur amide ligase enzymes for the identification of inhibitory compounds by *in silico* methods

---

A thesis submitted in partial fulfilment of the requirements for the degree of:

MASTER OF SCIENCE

by

Coursework/Thesis

in

Bioinformatics and Computational Molecular Biology

Department of Biochemistry, Microbiology and Biotechnology

Rhodes University

By: Chiratidzo Respina Chamboko

December 2019

Declaration

I, Chiratidzo Respina Chamboko, declare that the thesis submitted to Rhodes University is my own work and has not previously been submitted for a degree or diploma at this or any other institution.

Chiratidzo Chamboko 04/12/2019

Signature: ..... **Date:** .....

Dedication

I humbly dedicate this thesis to my guardian angel, my lovely grandmother
Kupurai Respina Chamboko

~ 01/06/1930 - 19/07/2003 ~

“Today's tears water tomorrow's gardens.”

Abstract

An increased emergence of resistant pathogenic bacterial strains over the years has resulted in many people dying of untreatable infections. This has become one of the most critical global public health problems, as resistant strains are complicating treatment of infectious diseases, increasing human morbidity, mortality, and health care costs. A very limited amount of effective antibiotics is currently available, but the development of novel classes of antibacterial agents is becoming a priority. Mur amide ligases are enzymes that have been identified as potentially good targets for antibiotics, as they are uniquely found in bacteria. They are responsible for the formation of peptide bonds in a growing peptidoglycan structure for bacterial cell walls. The current work presented here focused on characterizing these Mur amide ligase enzymes and obtaining inhibitory compounds that could potentially be of use in drug discovery of antibacterial agents. To do this, multiple sequence alignment, motif analysis and phylogenetic tree constructions were carried out, followed by docking studies and molecular dynamic simulations. Prior to docking, homology modelling of missing residues in the MurF structure (PDB 1GG4) was performed. Characterization results revealed the Mur amide ligase enzymes contained defined conservation in limited regions, that ultimately mapped towards the central domain responsible for ATP binding (presence of a conserved GKT motif). Further analysis of results further unraveled the unique patterns observed within each group of the family of enzymes. As a result of these findings, docking studies were carried out on each Mur amide ligase structure. At most, two ligands were identified to be sufficiently inhibiting each Mur amide ligase. The ligands obtained were SANC00574 and SANC00575 for MurC, SANC00290 and SANC00438 for MurD, SANC00290 and SANC00525 for MurE and SANC00290 and SANC00434 for MurF. The two best ligands identified for each enzyme had docked in the active site of their respective proteins, passed Lipinski's rule of five and had substantially low binding energies. Molecular dynamic simulations were then performed to analyze the behavior of the proteins and protein-ligand complexes, to confirm the lead compounds as good inhibitors of the Mur amide ligases. In the case of MurC, MurD and MurE complexes, the identified ligands clearly impacted the behavior of the protein, as the ligand bound proteins became more compact and stable, while flexibility decreased. There was however an opposite effect on MurF complexes, that resulted in identified inhibitors being discarded. As a potential next step, *in*

vivo and *in vitro* experiments can be performed with identified ligands from this research, to further support the information presented.

Table of Contents

Table of Figures

List of Tables

List of Web Servers and Software

1. AutoDock-Vina
2. Discovery Studio Visualizer 4
3. GROMACS
4. Jalview
5. MAFFT
6. MATLAB
7. MEGAX
8. MEME
9. MODELLER
10. NCBI- BLAST
11. PROCHECK
12. PROMALS3D
13. ProSA
14. PyMOL
15. T-COFFEE
16. UniProt
17. VERIFY3D
18. VMD
19. Xmgrace

List of Abbreviations

AIC	Akaike information criterion
ATP	Adenosine tri-phosphate
ADP	Adenosine di-phosphate
BIC	Bayesian information criterion
DNA	Deoxyribonucleic acid
DOPE	Discrete Optimized Protein Energy
DOTS	Directly observed treatment, short-course
GROMACS	GROningen MAchine for Chemical Simulations
HIV	Human Immunodeficiency Virus
MD	Molecular dynamics
MSA	Multiple Sequence Alignment
MW	Molecular weight
PDB	Protein Data Bank
PG	Peptidoglycan
Pi	Phosphate
PSSM	Position-Specific-Scoring-Matrix
Rg	Ratio of gyration
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SANCDB	South African Natural Compounds database
TB	Tuberculosis

WHO	World Health Organization
3D	Three-dimensional

Acknowledgements

I would wholeheartedly like to thank my supervisor, Dr Vuyani Moses, for all the assistance and guidance throughout this research. Thank you for believing in me always and helping me see direction where I could not. I am eternally grateful.

To Prof. Özlem Tastan Bishop and the whole RUBi team, thank you for accepting me as one of your own, and leading me in the right path throughout the course of this research. All your input and kind assistance during progress presentations really did help shape this thesis, and for that I am so thankful.

To my friends and significant others, thank you so much for all the emotional support throughout the year. Indeed, I learnt masters is not easy, but you were always there to catch me every time I fell and continued to feed my soul with positivity. Truly love and appreciate all you did and continue to do for me.

A special thank you to my lovely parents, (Mrs.) Gloria Miriam Changundega Chamboko and (Dr) Tafireyi Leo Chamboko. Your unwavering support in my education has been amazing, from the first day I stepped into Dominican Convent Primary School for grade one (with an oversized uniform I was supposed to grow into), until now! Thank you for teaching me hard work pays off and loving me unconditionally through all my ups and downs. You have sacrificed so much to see me happy and well, and I would be so lucky to become even half of the people you are today. Thank you so much.

To Almighty God above, none of this would have been possible without You.

“For I know the plans I have for you," declares the LORD, "plans to prosper you and not to harm you, plans to give you hope and a future.”

Jeremiah 29:11

Chapter 1: Introduction

1.1 Antibiotic rise and resistance formed by bacteria

Bacterial species for many years, have invaded the earth and caused death in humans. As a result of these pathogenic creatures, antibiotics were discovered, in hopes of eradicating bacteria that cause infectious diseases and therefore save humanity. The first antibiotic ever to be discovered was penicillin, by a Scottish researcher named Sir Alexander Flemming, in a peculiar way (Nikaido, 2009). He was initially investigating the influenza virus in his laboratory, and after returning from a two-week vacation, he found a mold that had developed on an accidentally contaminated culture plate growing staphylococcus (Nikaido, 2009). When he examined the plate further, he discovered the culture prevented staphylococci growth and this is how penicillin came to be (Nikaido, 2009). This discovery by Sir Flemming paved way for a more commercial production of many other antibiotics, and their use had a profound impact on bacterial life here on earth (Coates and Hu, 2007). In the modern-day world, antibiotics are being manufactured worldwide, at an estimated scale of about 100 000 tons annually (Nikaido, 2009). They are being used to cure diseases such as bacterial endocarditis, meningitis, gonorrhea, syphilis, pneumococcal pneumonia and many more (Coates and Hu, 2007). It is however unfortunate that antibiotics are being overused, and we take for granted that any infectious disease can be cured by antibiotic therapy (Coates and Hu, 2007). This has in turn created antibiotic resistant strains over the years, rendering the discovered antibiotics useless in combating many infectious diseases (Coates and Hu, 2007).

Bacterial resistance has been in existence from the beginning of the antibiotic era, but within the last twenty years, the emergence of dangerous, resistant strains has occurred (Fair and Tor, 2014). Many bacterial species have become resistant to antibiotic treatment, and it has been reported that more than 60% of the world's infections are antibiotic-resistant (Fair and Tor, 2014). This rapid emergence of resistant bacteria has now been described by many health organizations as a crisis

that could have catastrophic consequences (Fair and Tor, 2014). Resistance of bacteria is caused by various factors, including overuse, incorrect prescribing, extensive agricultural use and the declining amount of new antibiotics being manufactured (Ventola,2015).

Over the years, antibiotics have been overused or inappropriately prescribed, and this is the main driver of resistance in bacteria (Ventola, 2015). In the United States alone, over $1,1 billion dollars have been spent on unnecessarily prescribed antibiotics (Ventola, 2015). Overuse and incorrect prescribing of antibiotics can include using them to treat eczema instead of lotions and creams specific for eczema patients, using antibiotics to treat sinusitis even when it is caused by a virus, using oral antibiotics to treat swimmer's ear instead of antibiotic ear drops and many more instances (Ventola, 2015). Unregulated antibiotics also contribute to overuse of these drugs, and therefore lead to resistance (Ventola, 2015). Many unregulated antibiotics are available over the counter without a prescription, making them easily accessible, cheap, plentiful and available online for purchase (Fair and Tor, 2014). This overuse of antibiotics is dangerous and can often lead to mutations of genes within bacteria (Ventola, 2015). The mutated genes can then be inherited from relatives and developed resistance can be passed from one generation of bacteria to the next (Fair and Tor, 2014). Antibiotics then remove all drug-sensitive competitors and leave behind resistant bacteria that can reproduce as a result of natural selection (Fair and Tor, 2014). Although there have been warnings regarding overuse, antibiotics are still being over prescribed worldwide (Fair and Tor, 2014).

Another major cause of antibiotic resistance is the use of these drugs in agriculture (Ventola, 2015). In both developing and developed countries, antibiotics are used as growth supplements for livestock, to promote growth and prevent infection (Ventola, 2015). This improves the health of animals and produce larger yields and high-quality products (Michael et al., 2014). Though this may be an advantage, the use of antibiotics in livestock has detrimental effects (Michael et al., 2014). When humans consume the livestock as food, resistant bacteria are transferred to them (Golkar et al., 2014). This transfer was first noted over 35 years ago, when high rates of antibiotic resistance were found in both farmers and farm animal intestinal flora (Barlett et al., 2013). More

recent studies using molecular detection methods, have shown that resistant bacteria in farm animals reach humans through meat products (Barlett et al., 2013). The resistant bacteria acquired by humans therefore cause infections that may lead to adverse health consequences or death in extreme cases (Ventola, 2015). Utilization of antibiotics also affects the environment microbiome (Ventola, 2015). More than 80% of the antibiotics given to livestock are excreted into the environment as urine and stool, then dispersed through ground water, surface runoff and fertilizers (Ventola, 2015; Barlett et al., 2013). Other antibiotics such as streptomycin and tetracycline are used as pesticides on fruit trees, mainly in western and southern America, and although this application only applies to a much smaller division of antibiotic use, the resultant geographical spread is considerable (Golkar et al., 2014).

Newly developed antibiotics used to be an effective way of combating resistant bacteria, but for the past few years antibiotic research has not been a major priority in the pharmaceutical industries (Fair and Tor, 2014). This has led to resistance in bacteria substantially increasing and no new antibiotics to fight new infections that they cause (Ventola, 2015). So, without new antibiotics, many bacteria form more resistant strains and this in turn increases health care costs worldwide (Fair and Tor, 2014). With this escalating evolution of resistance and a declining antibiotic manufacturing process, a post- antibiotic era may be eminent (Fair and Tor, 2014).

1.2 Infectious diseases and tuberculosis

Infectious diseases were ranked in the top ten leading causes of death by the World Health Organization as of 2016, particularly in the low-income countries (Bloom and Cadarette, 2019). Many of the infectious diseases include cholera, tuberculosis, leprosy, anthrax, syphilis and many more (Bloom and Cadarette, 2019). They have become the top leading causes of death due to resistant strains emerging, with very few antibiotics available to treat such infections (Bloom and Cadarette, 2019). The top three infectious diseases causing high population deaths are lower respiratory infections, diarrheal diseases and tuberculosis (Bloom and Cadarette, 2019).

Altogether, they cause about 38.4% fatalities annually, with tuberculosis being the major contributor (Bloom and Cadarette, 2019). With the continuous rise in resistant strains, novel approaches to combat these pathogenic organisms is becoming a great need, especially in tuberculosis patients.

Tuberculosis (TB) is a disease that claims the lives of millions of people each year, mainly in less developed countries (Watson and Gill, 1990). Tuberculosis does occur in every part of the world, but in 2017, the largest number of new TB cases was found mainly in South-East Asia and the Western Pacific regions (62% of the new cases) followed by the African region (25% of the new cases) (World Health Organization, 2017). About 87% of new TB cases were found in the 30 high TB burden countries, with only eight countries accounting for at least two thirds of the new cases (World Health Organization, 2017). These eight countries include India, Nigeria, China, Bangladesh, Pakistan, Indonesia, South Africa and the Philippines (World Health Organization, 2017). Tuberculosis can infect anyone, but people living with HIV are 20 to 30 times more likely to develop active TB than people without this virus (World Health Organization, 2017). It has been estimated that in 2017, at least 0.3 million people living with HIV died of TB, and 0.9 million new cases of TB were discovered among people living with HIV, of which 72% of those people were living in Africa (World Health Organization, 2017). This has given rise to an HIV-driven crisis in Africa, where the worst tuberculosis epidemic has occurred since the advent of the antibiotic era (Chaisson and Martinson, 2008). This continues to be fueled by the rise in the population of HIV infected people, weak health care systems within the many underdeveloped countries in Africa, inadequate laboratories and conditions that promote transmission of TB disease (Chaisson and Martinson, 2008). More recently, problems of multi-drug resistant strains of tuberculosis have arisen, therefore worsening the situation and requiring urgent need for development of new drugs (Chaisson and Martinson, 2008).

1.2.1 Causes, development and treatment of tuberculosis

Tuberculosis is a chronic infectious disease that is mainly caused by tubercle bacillus, which is a gram-positive bacterium found in the genus *Mycobacterium tuberculosis,* and mainly infects the lungs of humans (Cole et al*.,* 1998). There are two main types of Tuberculosis, latent and active tuberculosis, where active tuberculosis can develop from an onset infection or from latent tuberculosis after a certain period, or when reactivated (Barry et al*.,* 2009). The principal mechanism for the spread of TB begins with an individual inhaling aerosolized droplet from an infected person, leading to an infection developing in the lungs of the uninfected individual (Ijaz et al*.,* 2002). Once an infection has developed and is established in the lungs, the body may respond by clearing the organism from the body immediately (Jilani and Siddiqui, 2019). This is mainly dependent on the newly infected individuals' immune status, genetics of the individual and if the exposure was primary or secondary (Jilani and Siddiqui, 2019). Virulent factors of *M. tuberculosis* may however make this elimination process difficult for the innate immune systems' phagocytes, the alveolar macrophages (Chowdhury et al*.,* 2018). Some of these factors include the high mycolic acid content found in the outer capsule of the bacteria that make phagocytosis difficult for the phagocytes, and cord factors that could directly damage the phagocytes (Jilani and Siddiqui, 2019). Catalase-peroxidase is another virulent factor that helps resist the oxidative response of the host cell, and lipoarabinomannan helps induce cytokines and resist the host oxidative stress (Chowdhury et al*.,* 2018). Other studies have also revealed that *M. tuberculosis* also inhibits the formation of an effective phagolysosome, which in turn limits and even prevents the elimination of the pathogenic organism (Jilani and Siddiqui, 2019).

When an individual fails to eliminate the TB organism from their system after infection, the development of primary TB occurs, but from observations, only 5 – 10 % of the exposed individuals develop this (Chowdhury et al*.,* 2018). Primary TB is normally located in the middle part of the lungs and is referred to as the Ghon focus, and this can become active, causing active TB straight after a primary infection, but usually becomes latent, causing latent TB, which is the first type of tuberculosis (Jilani and Siddiqui, 2019). Some individuals with latent TB do not develop overt disease, they have no symptoms, and their chest x-ray may be normal (Barry et al*.,*

2009). However, if there is immunosuppression in the host, maybe caused by HIV or other medications which compromise the immune system, latent TB is capable of being reactivated into active TB, which is the second type of tuberculosis (Barry et al*.,* 2009). Individuals with active TB have the bacteria rapidly multiplying and invading different organs of the body, and may experience symptoms such as chest pain, weakness, fever, chills, sweating at night and coughing (Barry et al*.,* 2009). With this active TB disease, an individual may become infectious, and this is caused by the tubercle bacilli which would have spread into the airways mechanically, via erosion of the caseating lesions, and in the absence of treatment, the mortality rate is about 80% (Ijaz et al*.,* 2002). The transmission of TB to other uninfected individuals is airborne, where an infected person in their most infectious stage of the disease coughs into the air and an uninfected individual inhale the aerosolized droplets (Ijaz et al*.,* 2002).

Once infected, an individual develops active TB and requires treatment immediately, before infecting others and evidently perishing from the disease. The first ever drug to be used in treating TB was streptomycin, an aminocyclitol glycoside antibiotic (Jnawali and Ryoo, 2013). Streptomycin is active in killing growing tubercle bacilli but is inactive against the intracellular or non-growing bacteria (Jnawali and Ryoo, 2013). When resistance to this drug emerged, newer drugs needed to be discovered to limit the number of patients dying from the resistant strains of *M. tuberculosis* (Jnawali and Ryoo, 2013). Directly observed treatment, short-course (DOTS) is an internationally recommended control strategy for TB that was initiated after streptomycin, that includes a short course of drugs, mainly the short course drug treatment, comprising of first-line drugs such as Isoniazid (H/Inh), Rifampicin (R/Rif), Pyrazinamide (Z/Pza) and Ethambutol (E/Emb) (Ginsberg, 2008). The treatment with these drugs usually lasts between 6-8 months, depending on the type of patient (Munro et al*.,* 2007). The drugs are normally accompanied by direct observation of therapy (DOT), that can either be carried out by a health worker or by someone nominated by the health worker (Munro et al*.,* 2007).

The first-line drugs may fail to treat tuberculosis patients, and this may be because of numerous reasons (Munro et al*.,* 2007). Patients with TB may not have completed the treatment, and this

could be because the side effects took a toll on them or there may be problems with the delivery of the drugs to the patients, especially in less developed countries that may have limited access to the anti-tuberculosis drugs sufficient for the population with TB (Munro et al*.,* 2007). Another reason the first-line treatment may fail is because of doctors prescribing incorrect regimes, or patients taking the drugs incorrectly (Munro et al*.,* 2007). However, if patients already have drug resistant TB, the first-line treatment will fail even when the patient takes the drugs correctly (Munro et al*.,* 2007). When patients fail to take the treatment correctly or do not complete treatment, this could result in prolonged infectiousness, relapse, drug resistance and even death (Munro et al*.,* 2007). The need for better drugs to combat the multi-resistant TB strains has increased, and drugs other than the first-line drugs must be administered to patients (Jnawali and Ryoo, 2013). These drugs are known as the second-line drugs and there are some that have been developed and some still in the process of being fully approved for use as anti-tuberculosis drugs (Jnawali and Ryoo, 2013). Examples of second line drugs discovered thus far include fluoroquinolones such as ofloxacin (OFX), levofloxacin (LEV), moxifloxacin (MOX) and ciprofloxacin (CIP) (Jnawali and Ryoo, 2013). The injectable anti-tuberculosis second-line drugs include amikacin (AMK), kanamycin (KAN) and capreomycin (CAP) (Jnawali and Ryoo, 2013). The less-effective second-line anti-tuberculosis drugs discovered however are still used in treatment and these include ethionamide (ETH)/prothionamide (PTH), cycloserine (CS)/terizidone and P-aminosalicylic acid (PAS) (Jnawali and Ryoo, 2013). These second-line drugs are however difficult to obtain and are more toxic and expensive (Jnawali and Ryoo, 2013).

A part of the pharmaceutical industry has thus begun to discover and develop new and improved TB drugs that may have a shorter treatment duration of 2 months or less, and have safety comparable to, or better than that of the first-line TB drugs (Ginsberg, 2008). The new drugs must have a novel mechanism of action, oral bioavailability, a pharmacokinetic–pharmacodynamics profile that is consistent with once-daily or less frequent dosing, less expensive, and low or no interactions with the P450 enzymes (Ginsberg, 2008). Although research has begun for better and improved TB drugs, only a small percentage of research worldwide is taking place, and this is the case with many other infectious diseases. Novel antibacterial drug discovery is limited, with very few new mechanisms of action for these drugs. In this research, the peptidoglycan synthetic

pathway will be researched as a multi-resistant antibacterial drug target source, in order to obtain novel methods of action against many resistant bacteria.

1.3 The peptidoglycan biosynthesis pathway and Mur amide ligases

The rise in multiple drug resistant strains of *M. tuberculosis,* and other known resistant bacteria causing infectious diseases, has given rise to the need for more treatments to be discovered and more drugs to be developed (Flynn and Chan, 2001). Novel drugs have been sought after, with scientists venturing into bacteria's biochemical pathway for peptidoglycan biosynthesis (PG) as a source of drug targets (Flynn and Chan, 2001).

The PG biosynthesis pathway is very crucial for a bacteria's survival (Kouidmi et al*.,* 2014). Peptidoglycan is a structure uniquely found in bacteria, and is a major component of their cell walls (Barreteau et al*.,* 2008). It is responsible for the bacteria's structural integrity and aids in balancing the internal osmotic pressure (Barreteau et al*.,* 2008). It is made up entirely of long glycan chains that are cross-linked by short peptides (Barreteau et al*.,* 2008). The cross linking of the glycan chains mainly occurs between an amino group of the diamino acid at position 3 and the carboxyl group of a d-Ala at position 4 (Barreteau et al*.,* 2008). The long glycan chains however, are formed from alternating *N*-acetylglucosamine (GlcNAc) and *N*-acetylMuramic acid (MurNAc) residues, that are linked by β1→4 bonds (Barreteau et al*.,* 2008). The PG biosynthesis pathway with all enzymes involved is shown in figure 1-1.

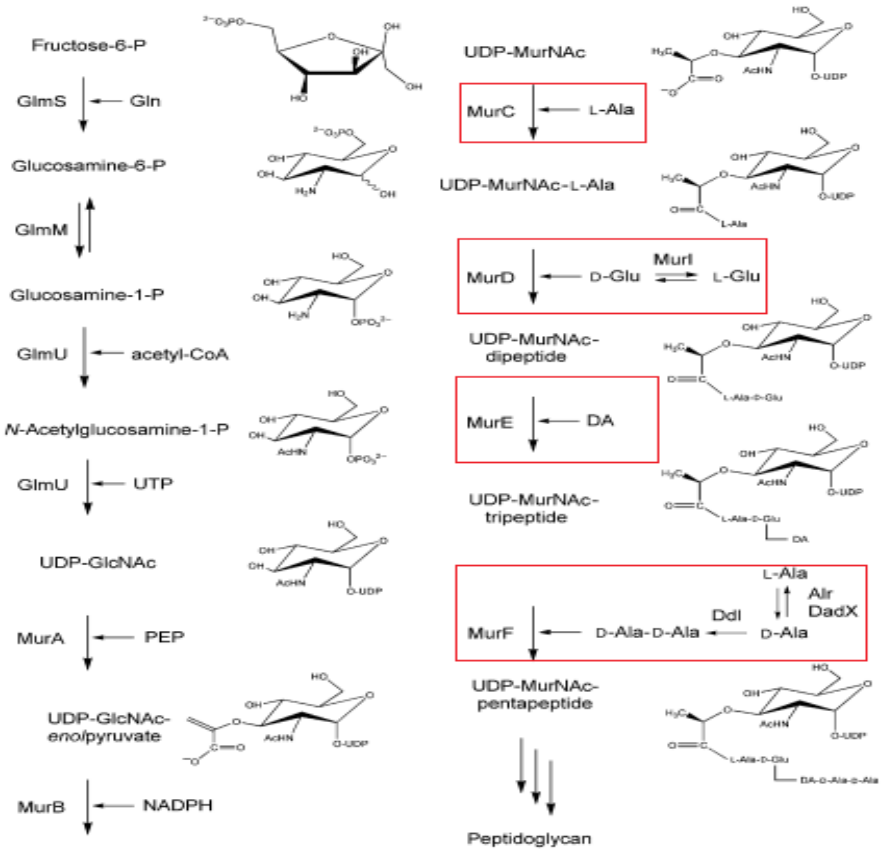


Figure 1-1: Steps of peptidoglycan biosynthesis with Mur amide ligases as potential drug targets for antibiotics. Ligases highlighted in red boxes within pathway (Barreteau et al., 2008).

The PG biosynthesis pathway can be divided into three steps. The first step in the PG biosynthesis pathway is the cytoplasmic step, where MurA and MurB first synthesize UDP- GlcNAc-enol pyruvate and UDP-MurNAc respectively from the UDP-GlcNAc (Kouidmi et al*.,* 2014). The Mur ligase enzymes which are MurC, MurD, MurE and MurF then add the pentapeptide moiety (mainly the sequential addition of L-Ala, D-Glu, meso diaminopimelate (meso -A 2 pm) and D-Ala-D-Ala, respectively) to form the cytoplasmic PG precursor UDP-MurNAc-pentapeptide (also known as UDP-MurNAc-L-Ala- c -D-Glu- meso -A 2 pm-D-Ala-D-Ala) (Kouidmi et al*.,* 2014). These are the main focus of this research, and they are used in the addition of amino acids to the D-lactoyl group of UDP-MurNAc (Kouidmi et al*.,* 2014).

The second step then commences in the PG biosynthesis pathway, and this is the membrane step where a lipid I is formed by MraY transferring a phospho-N-acetylMuramoyl-pentapeptide from UDP-MurNAc-pentapeptide to the lipid carrier undecaprenyl-P (Und-P) (Kouidmi et al*.,* 2014). Lipid II is then formed by MurG then adding GlcNAc from UDP-GlcNAc to lipid I (Liu and Breukink, 2016). Lipid II is then transported by a flippase from the inner side of the membrane to the outer side during the last step of the PG biosynthesis pathway, where it is polymerized, giving a peptidoglycan layer (Liu and Breukink, 2016).

The PG biosynthesis pathway provides the bacteria with a rigid, flexible and strong peptidoglycan layer that is mainly only found in bacteria, with no known forms in humans, making enzymes within this pathway attractive as drug targets for antibiotics (Barreteau et al*.,* 2008). Some clinically approved antibiotics such as beta-lactams, glycopeptides and fosfomycin andcycloserine, interfere with some steps within the PG biosynthesis pathway (Liu and Breukink, 2016). There are however no antibacterial agents that target the amide ligases such as MurC, MurD, MurE and MurF (Kouidmi et al*.,* 2014). These enzymes are excellent drug targets not only because they are crucial in the survival of bacteria but also because they are conserved among all medically relevant bacteria with no equivalent structures in eukaryotic cells (Kouidmi et al*.,* 2014).

1.3.1 Mur Ligases

The Mur ligases catalyze the formation of a peptide bond during the addition of the amino acids, while forming ADP and Pi from ATP simultaneously, and requiring cations such as Mg^{2+} or Mn^{2+} for the reaction (Kouidmi et al*.,* 2014). The mechanism of action for these Mur ligases is the same, and this involves the use of ATP to activate a carboxyl group of the UDP-precursor, forming an acyl phosphate intermediate (Pi) and ADP (Barreteau et al*.,* 2008). The acyl phosphate undergoes nucleophilic attack of the dipeptide, and this leads to the formation of a high-energy tetrahedral intermediate, that later breaks down into Pi and a peptide (Barreteau et al*.,* 2008). The Mur amide ligases defined a new family of enzymes by containing an ATP-binding consensus sequence in each of them and a series of six invariant residues (Barreteau et al*.,* 2008). Other enzymes found

in this family of enzymes do not belong to the PG biosynthesis pathway but include enzymes such as folylpoly-g-L-glutamate synthetase (FolC), the C-terminal region of cyanophycin synthetase (CphA) and the poly-g-glutamate synthetase (CapB) from Bacilli (Barreteau et al., 2008). Mur ligases have been observed to have the same three-dimensional structures, with three domains, mainly the N- terminal domain, the C-terminal domain and the Mn coordinates site as seen in figure 1-2 (Barreteau et al., 2008). The N-terminal domain binds the substrate, that is, the UDP-precursor, and it is the central domain in the binding of ATP (Barreteau et al., 2008). The C-terminal domain binds the amino acid or dipeptide (Barreteau et al., 2008). These enzymes can exist in an open or closed conformation, and the closed conformation has been observed to be caused by ligand or substrate binding (Barreteau et al., 2008).

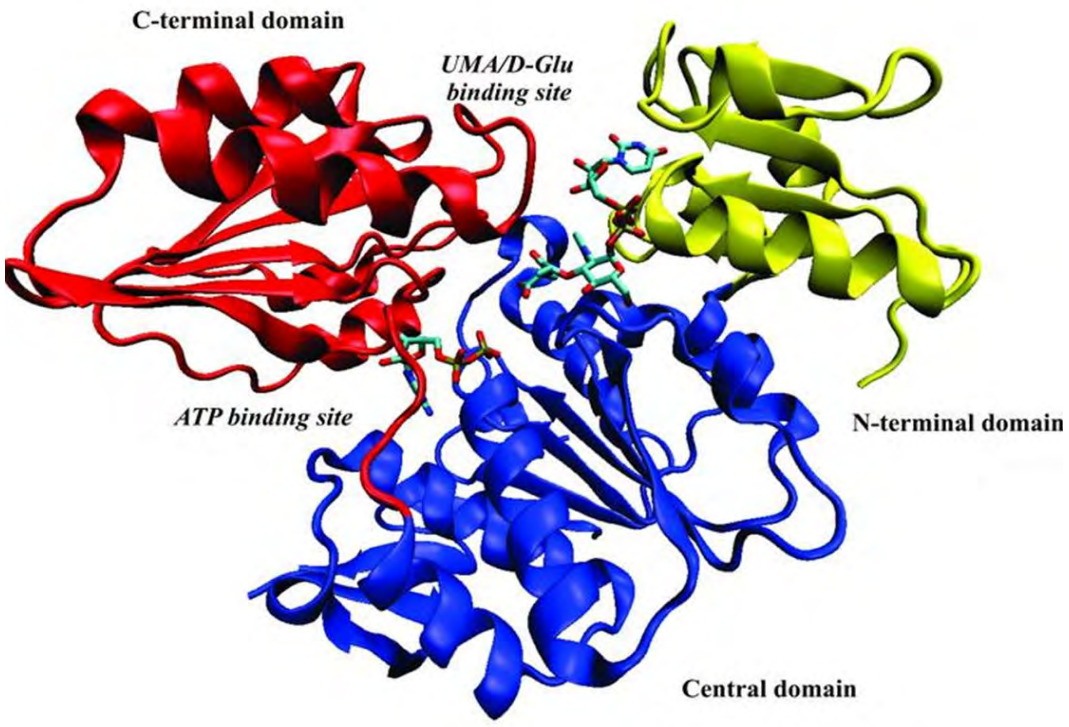


Figure 1-2: General structure of a Mur amide ligase. Protein containing three domains, C-terminal domain (red), N-terminal domain (yellow) and Central domain (blue). The ATP binding site in the central domain, the substrate binding site in the N-terminal domain and the amino acid binding site in the C-terminal domain. Structure in closed conformation due to substrate binding (Barreteau et al., 2008).

The Mur ligase MurC, is a cytosolic enzyme that aids in the addition of the first amino acid of the peptide stem (Barreteau et al., 2008). The amino acid in most bacterial species added is L-alanine, but in some rare instances, L-serine or glycine is added in place of L-alanine (Barreteau et al., 2008). This then means that this enzyme's preferred substrate is L-alanine, but glycine and L-serine can be used with lower efficiencies (Emanuele et al., 1996). The reversible reaction that involves MurC is phosphate and ADP-depend, and begins with the binding of ATP, then UDP-MurNAc and finally L-Ala (Emanuele et al., 1997). MurC apo structures were seen to be in a closed conformation, different from other Mur ligases, and this then suggested that MurC may exist in different conformations in solution when no ligand is bound to them (Barreteau et al., 2008).

MurD is the second ligase acting upon UDP-MurNAc, catalyzing the addition of the second amino acid residue D-glutamic acid (Barreteau et al., 2008). This enzyme showed great specificity for D-Glu and only a few closely related compounds such as homocysteic acid, 3- or 4-methyl-D-Glu, and cyclopentane or cyclohexane analogues of D-Glu were also good substrates for this enzyme (Barreteau et al., 2008). The reaction MurD catalyzes is reversible, but unlike MurC, it is not ADP-dependent (Barreteau et al., 2008). There are two magnesium ions within the crystal structure of MurD, with the first 'classical' one being used in ATP binding, and the second one involved in acyl phosphate formation (Barreteau et al., 2008). The second magnesium ion is coordinated with two water molecules bonded to a carbamoy-lated lysine residue by hydrogen bonding (Barreteau et al., 2008).

The third amino acid in the peptide stem is either meso-diaminopimelic acid or L-lysine, and this is added by the enzyme MurE (Barreteau et al., 2008). Other amino acids have been known to be added in place of meso-diaminopimelic acid or L-lysine, and these include L-ornithine, LL-A2pm, meso-lanthionine, L-diaminobutyric acid and L-homo-serine (Schleifer & Kandler, 1972). The MurE is highly specific, and addition of the wrong amino acid by the MurE in a specific species could result in cell lysis (Mengin-Lecreulx et al., 1999).

MurF is the next Mur ligase that is responsible for the addition of a dipeptide, D-Ala-D-Ala, on positions 4 and 5 of the peptide stem (Barreteau et al., 2008). Studies have shown that MurF has high specificity for the C-terminal amino acid (Barreteau et al., 2008). The enzyme is an open structure that undergoes domain closure when bound to a ligand or substrate (Barreteau et al., 2008).

1.3.2 Inhibition of Mur ligases

Many inhibitors of amino acid Mur ligases have been studied and some have been identified over the years to combat bacterial resistance (Toma i et al., 2012). The best-known inhibitors and those that have shown to be more potent than others discovered are the phosphinate inhibitors that are derivatives of phosphinic acid (Toma i et al., 2012). The phosphinic acid is a known dipeptide analogue that is linked to uridine diphosphate by a hydrophobic spacer (El Zoeiby et al., 2003). It inhibits the Mur ligases by acting as slow-binding inactivators, through a process where the enzyme promotes the transfer of γ-phosphate from ATP to the phosphinate anion, so as to produce ADP and a phosphorylated inhibitor (El Zoeiby et al., 2003). The produced phosphoryl phosphinate moiety has been observed to imitate the tetrahedral intermediate that is formed in the normal PG biosynthesis pathway of reactions involving the MurC to MurF ligases (El Zoeiby et al., 2003). The inhibition with the tetrahedral intermediate-like phosphoryl phosphinate moiety has been seen to bind tightly to the enzymes, but no antibacterial activity has been observed thus far (El Zoeiby et al., 2003). This could be because there is inefficient transport of these inhibitory molecules into the cytoplasm, or a failure for them to accumulate up to more fatal inhibitory concentrations (El Zoeiby et al., 2003). Because of this, the search for other inhibitors for the Mur ligases has continued and these include several substrate analogues that can either be amino acid analogues or analogues of the UDP-sugar substrates (El Zoeiby et al., 2003). Some inhibitors of the Mur ligases have been shown in table 1-1.Inhibitors of Mur ligases that act as multi-target inhibitors are considered to be the best inhibitors that can be designed for multi-resistant strains of *M. tuberculosis* and bacteria in general. To do this, it is important to know and recognize the conserved binding motifs and the common kinetic mechanism in all the Mur amide ligases (Toma i et al., 2012).

Table 1-1: Inhibitors of Mur ligases MurC to MurF

Enzyme	Inhibitor	Extent of inhibition
MurC	Phosphinate	IC_{50} = 49 μM
	l-Alanine analogues:	
	β-Alanine	K_{is} = 110 mM
	β-CN-l-Alanine	K_{is} = 3.3 mM
	l-Vinylglycine	K_{is} = 5.8 mM
	β-Chloro-l-alanine	32.8%
	l-Cysteine	19%
	β-Chloro-l-alanine	76%
	β-Cyano-l-alanine	88%
	β-Fluoro-l-alanine	94%
MurD	Phosphinate	IC_{50} = 680 nM
	Phosphinate	IC_{50} nM
	d-Glutamic acid analogues:	
	dl-Homocysteic acid	58%
	d-*erythro*-3-Methylglutamic acid	47%
	d-*erythro*-4-Methylglutamic acid	26%
MurE	Phosphinate	IC_{50} = 1.1 μM
	A_2pm[a] analogues:	
	(2*S*,3*R*,6*S*)-3-Fluoro-A_2pm	IC_{50} = 2.3 mM
	N-Hydroxy-Apm	IC_{50} = 0.56 mM
MurF	Aminoalkyl phosphinates:	
	N-Acylphosphinate	K_i = 700 μM
	N-Glutarylphosphinate	K_i = 200 μM
	Pseudo-tetrapeptide phosphinate	K_i = 200 μM
	ATP analogue	K_{is} = 33.6 μM

Knowing this, an inhibitor can be designed to recognize the homologous binding motifs, and therefore have the ability to inhibit more than one enzyme, creating a more efficient inhibition of the PG pathway (Toma i et al., 2012). These would also be considered as multi-target inhibitors, and resistance to these inhibitors would be much lower because mutations with resistance would need to occur in more than one target gene simultaneously (Toma i et al., 2012).

Considering new knowledge and equipment that has become available for manipulation, this *in silico* analysis is aimed at possibly finding new and better inhibitory compounds for the Mur amide ligases than those already discovered, using the PG biosynthetic pathway found in *Escherichia coli*.

1.4 Problem statement and motivation

The World Health Organization has framed multi-drug resistant bacteria as a very pressing, global public health problem, for it is one of the leading causes of death annually (World Health Organization, 2017). The rise in bacterial resistance has continued to increase with a decreasing production of antibiotics worldwide (Fair and Tor, 2014). Out of the worlds' 15 largest pharmaceutical companies, only 1.6% of drugs manufactured were antibiotics (Fair and Tor, 2014). This has raised red flags, alarming the world of a post-antibiotic era to come, if action against these pathogenic microbes is not taken (Fair and Tor, 2014). The decrease in production of antibiotics by the pharmaceutical industry has been caused by many factors, but the loss in profitability of antibiotics is a major motivation for this decline (Fair and Tor, 2014). Antibiotics are administered to patients for very limited durations, while chronic treatments are required for much longer time spans, making them more profitable than antibiotics (Fair and Tor, 2014). With this, research into novel antibiotic drug discovery has therefore deteriorated, and allowed for the rise of more resistant strains of bacteria (Engel, 2012). The need for better antibiotics with novel mechanisms of action have risen, and some research has begun to try and resolve this pressing issue, by looking at different parts of bacteria for antibiotic target sources. A venture into the

peptidoglycan biosynthetic pathway has allowed for more drug targets for antibiotics to be identified (Engel, 2012). This pathway is unique to bacteria, making it an excellent source of drug targets against bacteria (Engel, 2012). Some antibiotics are being discovered using enzymes from this pathway, but Mur amide ligases have not quite been used as potential drug targets. In this work, we seek to explore these Mur amide ligases by obtaining inhibitory compounds that could potentially be used as antibacterial agents against multi resistant bacteria.

1.5 Aims and objectives

The main aim of this research is to study the Mur amide ligases found in the Peptidoglycan biosynthesis pathway, specifically as potential drug targets for drug discovery of antibacterial antibiotics, by searching for inhibitory compounds specific to them, perform lead optimization by docking different compounds from databases such as the South African Natural Compounds Database (SANCDB), then performing MD simulations on them as apo proteins, and in complex form with inhibitory compounds identified. The MD simulations will test the effects of the identified compounds from protein dynamics. This aim will be accomplished by means of the following objectives:

1. Characterize Mur amide ligases to observe and understand structural and functional relationships within the family of enzymes.
2. Search for inhibitory compounds specific to the Mur amide ligase enzymes from the South African Natural Compounds Database (SANCDB) (https://sancdb.rubi.ru.ac.za/).
3. Perform lead optimization by docking compounds into Mur amide ligase enzymes using Autodock Vina, to possibly obtain any hit compounds against the ligases.
4. Perform MD simulations on the Mur amide ligases enzymes to test the effect of the identified hit compounds on protein dynamics.

Chapter 2: Characterization of Mur amide ligases by sequence analysis and motif discovery

2.1 Introduction

This chapter is composed of sequence analysis, motif discovery and phylogenetic tree construction experiments undertaken to reveal the relationships between Mur amide ligases. We began this project by performing multiple sequence analysis using different open source tools, then retrieved motifs unique and common within the Mur amide ligase family. The acquired motifs were then mapped to the ligases, revealing similarities and differences between the enzymes. Final characterization of the family of enzymes was then further analyzed by phylogenetic trees. A brief background into the techniques used in this chapter are outlined, followed by a detailed methodology stating how the experiments were performed, then finally a results and discussion section for the observed results.

2.2 UniProt and BLAST

To provide oneself with information regarding relationships between proteins, sequence analysis is normally a starting point. This is a technique widely used in Bioinformatics to compare proteins and derive common features and differences between them. Some techniques used to compare protein sequences include multiple sequence analysis, motif discovery and phylogenetic analysis, and these techniques have been used in this work. Sequences can be retrieved first from different online tools such as UniProt, then non-redundant databases such as BLAST (Basic Local Alignment Search tool) can be used to obtain similar sequences, also known as homologs, for the input sequence.

UniProt is a large database for protein sequences and their annotations, that has continued to grow over the years (UniProt Consortium, 2014). It is a resource composed of several component parts

that can be utilized to obtain protein sequences for research purposes (UniProt Consortium, 2014). The UniProtKB/Swiss-Prot is a section of UniProt containing about half a million sequences, that can be used to obtain curated and reviewed entries (UniProt Consortium, 2014). This sections' growth is dependent on new proteins being experimentally characterized (UniProt Consortium, 2014). The other part of UniProt known as UniProtKB/TrEMBL, contains about 80 million sequences that are unreviewed (UniProt Consortium, 2014). Although entries in this section are not curated, automatically generated annotations are used as supplementation (UniProt Consortium, 2014). The UniProt knowledgebase is a central hub for organization of protein information and has cross-references to more than 150 databases (UniProt Consortium, 2014). Accession numbers are used as tags for protein sequences within UniProt, making it easier to search and find sequences (UniProt Consortium, 2014). In this research, protein sequences for the Mur amide ligases of *Escherichia coli* strain k12 were obtained from the UniProtKB/Swiss-Prot section, and these were used in BLAST to obtain seven homologs for each group of Mur amide ligases.

The BLAST tool is a sequence similarity search tool which can be accessed through a web interface or as a stand-alone tool, that is used to compare DNA or protein query sequences to chosen database of other sequences (Altschul et al*.,* 1997; Johnson et al*.,* 2008). There are several variants of BLAST that can be used, and these include megaBLAST, BLASTP, BLASTX, TBLASTN, PSI-BLAST, RPSBLAST and DELTA_BLAST (Johnson et al*.,* 2008). MegaBLAST is mainly used to compare very similar nucleotide-nucleotide sequences, while BLASTN is used to compare more distantly related nucleotide-nucleotide sequences (Johnson et al*.,* 2008). BLASTP compares protein-protein sequences, and its algorithm is the basis of other types of BLAST searches such as BLASTX and TBLASTN (Johnson et al*.,* 2008). With BLASTX, a nucleotide query is translated and searched against a protein database, while in TBLASTN, a protein query is searched against a translated nucleotide database (Johnson et al*.,* 2008). For PSI-BLAST, a BLASTP is first performed to gather information that is then used to produce a Position-Specific-Scoring-Matrix (PSSM) (Johnson et al*.,* 2008). This PSSM is formed by the query of length N x 20, with each of the N columns corresponding to a letter in the query sequence, and each column containing 20 rows (Johnson et al*.,* 2008). Each row in the matrix corresponds to a specific residue, and describes

the likelihood of other related sequences having the same residue at that specific position (Johnson et al., 2008). With this matrix, PSI-BLAST then searches protein sequence databases to obtain results (Johnson et al., 2008). For Reverse-Position-Specific BLAST (RPSBLAST), a quick search using a protein query sequence is performed against a PSSM database produced by PSI-BLAST, while DELTA_BLAST produces a PSSM from a RPSBLAST search of the query, then uses this PSSM to search against a protein sequence database (Johnson et al., 2008).

Although there are many variants of BLAST, they all employ the same method of function. BLAST is known to work as a heuristic that obtains short matches between a query and a database sequence, then attempts to start alignments from those matches (Altschul et al., 1990). The process BLAST follows to obtain results can be divided into three phases; setup, preliminary search and traceback (Johnson et al., 2008). Beginning in the setup phase, the query sequence, search parameters and type of database required, are read in by BLAST, then the query sequence is checked for repeats or low-complexity before moving on to produce a set of short fixed-length sequences based on this query sequence, also known as "words" (Johnson et al., 2008; Altschul et al., 1990). These words are used to obtain matches in the database sequences (Johnson et al., 2008). In the preliminary search phase, every sequence in the database is searched first, for matches to the query words created, then these are used to commence a gap-free extension (Johnson et al., 2008). The gap-free extensions are scored, and those that reach a certain score are used to seed gapped extensions which are scored as well (Johnson et al., 2008). The gapped extensions that reach a specific score are saved and used in the final traceback phase (Johnson et al., 2008). In this last phase the saved gapped extensions are used as seeds for more gapped extensions that also calculate insertions and deletions, but using more sensitive parameters (Johnson et al., 2008). After BLAST has completed the three phases, a results page with "hit" sequence information is displayed. The results will include the query sequence and all the reference sequences found within the database that match closely to the query sequence. Percentage identities and similarities will also be shown, to distinguish which sequences are more related to the query sequence given. Statistical information such as the "expect" value are also shown, that estimate how many matches may have occurred at a given score by chance, allowing the user to judge how much confidence to have in a hit sequence. In this research, BLASTP was used to search for homologs for the Mur

amide ligases in seven different bacteria. Their sequences were then aligned with each other in multiple sequence alignment, to assess sequence conservation and discover structural and functional information.

2.3 Multiple Sequence alignment

By definition, Multiple sequence alignment (MSA) is the alignment of three or more sequences that share an evolutionary history (Corpet, 1988). To maximize scores and obtain alignments, different alignment methods are used, but in this work, progressive alignment methods are mainly used with programs such as PROMALS3D, T-COFFEE and MAFT. This method begins by aligning two of the most similar sequences within the set, then progressively combining more pairwise sequence alignments, resulting in a final MSA (Corpet, 1988). A guide tree is constructed in the first stage of the MSA progressive method process, that represents the relationship between sequences (Corpet, 1988). This guide tree is then employed as a reference when adding more sequences sequentially to the growing MSA in the second stage of the process (Corpet, 1988). Gaps are introduced into sequences and extended during the process, to optimize alignments, but extension is preferred because it results in a less penalizing score that reduces the overall alignment score as compared to the introduction of gaps (Corpet, 1988). Success of this alignment method depends greatly on the input sequences, as errors occurring at the beginning of the alignment are carried through to the end. Nevertheless, the progressive alignment method is a fast and efficient way of aligning large sets of sequences, and it is readily available in many publicly accessible web servers. Results of the MSA can be viewed using tools such as Jalview, Bioedit or Cinema.

2.4 Motif discovery

Here, once the MSA was complete, the discovery of motifs common throughout the set of sequences was observed. Motifs are a short sequence patterns that reveal structural and functional information associated with the sequences (Bailey et al*.,* 2015). They are thought to be conserved

by evolution and can be observed by many different tools, with the MEME suite software being one example used in this work. An unaligned set of sequences is submitted into MEME, which utilizes statistical techniques to discover *de novo* motifs (Bailey et al*.,* 2015). The results show motif sequence, length and position within queried sequences (Bailey et al*.,* 2015). This paves way for these motifs to be mapped onto crystal structures, to observe location and possibly function of the motif (Bailey et al*.,* 2015).

2.5 Phylogenetic analysis

To further analyze a group of sequences, phylogenetic trees can be constructed. Phylogenetics is the study of genetically related groups of organisms, that reveals the level of evolutionary distance between them (Liao et al*.,* 2006). Aligned sequences are used to construct phylogenetic trees, and based on the amount of sequences and homology, models are constructed to best fit the data (Liao et al*.,* 2006). These models are then utilized to build phylogenetic trees. The trees display evolutionary distances in a more understandable format, and can be rooted or unrooted, depending on whether an ancestral root is available or unavailable (Liao et al*.,* 2006). The root is mainly the start of the phylogenetic tree, with branches extending from it that show evolutionary relationships between sequences (Liao et al*.,* 2006). The branches are of different lengths, signifying the evolutionary time (Liao et al*.,* 2006). Various methods can be employed to construct phylogenetic trees, but the three most commonly used methods that are thought to be more robust are the maximum likelihood, maximum parsimony, and minimum distance methods (Liao et al*.,* 2006). In this work, the maximum likelihood method was utilized to create phylogenetic trees in a computer software called Molecular Evolutionary Genetics Analysis (MEGA).

Although the maximum likelihood method is considered slow and computationally expensive, it is more accurate (Felsenstein, 1981). It is based on the concept that each residue site evolves independently, allowing phylogenetic relationships to be analyzed at every site (Felsenstein, 1981). For each residue position in a sequence, the maximum likelihood method uses algorithms that estimate the probability of that residue at a specific position, depending on whether the ancestral sequences contain that specific residue (Felsenstein, 1981). Once probabilities are

calculated, a bifurcating tree begins to form, which leads to a constructed phylogenetic tree (Felsenstein, 1981).

2.6 Chapter objectives

1. Retrieve *E.coli* Mur amide ligase sequences and homologs from seven other organisms.
2. Conduct a multiple sequence alignment on all sequences, and within individual groups of Mur amide ligases, and search for conserved and non-conserved regions.
3. Discover motifs within the family of Mur amide ligases, then map these motifs to corresponding ligase structures.
4. Construct phylogenetic trees using multiple sequence analysis results

2.7 Methodology

2.7.1 Sequence retrieval

Four curated *E.coli* (strain k12) Mur amide ligase sequences were acquired from UniProt, with each sequence representing each of the amide ligase groups; MurC, MurD, MurE and MurF (UniProt Consortium, 2014). These sequences were then used in NCBI's BLAST to obtain homologous sequences for each group (Pruitt et al., 2011). Seven homologs from different organisms were chosen for each group, based on high sequence identity and similarity (30% and above), and a low E value. The seven different organisms were kept constant in all Mur amide ligase groups, and a combined total of 32 sequences were stored in FASTA format, to be used in MSA. Sequences obtained can be viewed in Appendix A-1.

2.7.2 Multiple sequence alignment

All 32 sequences contained in a FASTA file were submitted to the online MSA tools, T-COFFEE (Notredame et al., 2000), PROMALS3D (Pei et al., 2008) and MAFT (Katoh and Standley, 2013). An alignment was produced with each software, using their default parameters. Four separate FASTA files for the different Mur amide ligase groups were used in the second round of MSA that was carried out. Within each group, 8 sequences were aligned using the same tools. The resultant alignments were viewed in Jalview, and conserved regions were observed using the Clustal X Color Scheme. The best alignment observed for the 36 sequences was then utilized in phylogenetic analysis.

2.7.3 Motif discovery and mapping

An unaligned FASTA file containing the 32 sequences was submitted to MEME v4.11, to discover motifs within the Mur amide ligase family (Bailey et al., 2015). Default parameters were used except for the following changes; motif width was between 6-20 residues, minimum number of motifs to be found was set to 100 and motifs can occur any number of times per sequence. Four FASTA files for each group of the Mur amide ligases were submitted to MEME and the same parameters were used. Motifs discovered were then mapped to *E.coli* sequence from UniProt and structures obtained from the PDB database (MurC-2FOO, MurD-1UAG MurE-1E8C and MurF-1GG4) using PyMOL.

2.7.4 Phylogenetic analysis

Phylogenetic trees were constructed for the Mur amide ligase family using Molecular Evolutionary Genetic Analysis (MEGA) vX tool (Kumar et al., 1994). The trees were constructed using T-COFFEE sequence alignment results, and the maximum likelihood statistical methods were employed to deduce evolutionary relationships. Best-fit substitution models for the data were initially searched for, using a Neighbor-joining tree and a strong branch filter. The best three models were chosen based on the lowest Bayesian information criterion (BIC) and Akaike

information criterion (AIC) scores. Three phylogenetic trees were then constructed using the bootstrap method as the test of phylogeny and 1000 bootstrap replications. Resultant trees were compared to their respective bootstrap consensus trees to ensure branching patterns were consistent and accurate, and from this the best phylogenetic tree was chosen. An all vs all sequence heatmap was then produced to further display the sequence relationships shown on the phylogenetic tree, using Python and MATLAB.

2.8 Results and Discussion

Over the years, antibiotic research has depleted due to many factors, causing many resistant strains of bacteria to emerge (Ventola, 2015). The battle against these growing pathogens has begun, with research for novel methods of action for antibiotics increasing (Ventola, 2015). The hope of the research conducted here is to add to the knowledge for the new antibiotics to be discovered, by looking at a pathway rich in enzymes for potential antibacterial targets. To begin this work, characterization of the Mur amide ligase family of enzymes was conducted, to better understand relationships within this family, and possibly find any distinguishable factors within each group. The Mur amide ligase family comprises of MurC, MurD, MurE and MurF. These enzymes are known to be similar in structure and function. The main aim in this chapter was to observe these similarities and possibly relate them to their shared function. This was initiated by sequence analysis of these enzymes.

2.8.1 Sequence retrieval

The organism chosen for this work was *E.coli*, as it is a universal bacteria that is known to cause a number of infections, and data pertaining sequences and crystal structures have been resolved for this organism. Results from this research using *E.coli* can then be related to all other types of bacteria. Approximately 32 sequences shown in Appendix A-1, were selected from UniProt and BLAST. This included the four E.coli sequences for each group of the Mur amide ligase family (MurC, MurD, MurE and MurF) and seven other sequences in each group for different organisms.

The seven other organisms with Mur amide ligases used were *Mycobacterium tuberculosis*, *Williamsia faeni*, *Rhodococcus kroppenstedtii*, *Nocardia seriolae*, *Actinoalloteichus hoggarensis*, *Mycolicibacterium peregrinum* and *Gordonia terrae.* They were used to diversify the data, so the results obtained can be relatable to other bacterial species, and not just *E.coli*. Human homologs of the Mur amide ligases were searched for, but none were found. This result supported the statement that the Mur amide ligases are unique to bacteria, and are a good source of antibacterial targets, because there are no human homologs known (Barreteau et al*.,* 2008). The sequences obtained were then applied in MSA.

2.8.2 Multiple sequence alignment

Alignments of all the 32 sequences were conducted with MAFT, T-COFFEE and PROMALS3D. Several tools were used so as to compare the results and observe the best alignment, to be used further in this work. Once results were produced from each software, analysis began with searching for known functionally important residues such as residues contributing to the structure of the active site, and residues important for the catalytic function of the enzymes. According to Fiuza et al., 2008, after comparing amino acid sequences of all the Mur amide ligases (MurC - MurF) from different bacterial genera, they were able to identify an ATP-binding motif with a conserved sequence (GXXGK(T/S)), also known as the GKT motif, and several other conserved residues within the family. Taking an *E.coli* MurC enzyme as a model, they were able to identify Asp50, Lys130, Glu174, and Asp351 as important residues for the catalytic process and functioning of these enzymes, whereas His199, Asn293, Asn296, and Arg327 were involved in the active site structure (Fiuza et al., 2008). Conservation of the GKT motif and the different residues established as important within the Mur amide ligase family were searched for within the MSA results obtained, and as shown in figure 2-1, the T-Coffee MSA results showed maximum conservation for most of these residues, making this result a candidate for the best MSA obtained. Further analysis was undertaken, to reveal if the loop regions within a crystal structure of the Mur amide ligases were falling within regions that contained gaps/inserts, and this was true for the T-COFFEE MSA results, making it the best MSA result obtained from the three different software. Other

results from PROMALS3D and MAFT (Appendix A-2) were quite similar to T-COFFEE, but T-COFFEE results proved to be slightly more reliable and therefore used for the rest of the research.

Figure 2-1: Multiple sequence alignment produced by T-COFFEE. A total of 32 sequences aligned to show conservation within the Mur amide ligase family. Sequences share at least 30% sequence identity. Results viewed in Jalview and conserved regions colored by ClustalX. The GKL motif (GXXGK(T/S)) and other important residues used in the selection of the T-COFFEE results shown in black boxes.

Overall the conservation observed after aligning the 32 ligase sequences was limited, with most conservation occurring around active site residues and ATP binding site residues. This aligned

with the fact that all the sequences only share between 30-40 % sequence identity. Although there was little conservation observed throughout the family of enzymes, there were some residues that were conserved within groups of the Mur amide ligases. Analysis of the loop regions within groups of the Mur amide ligases however, revealed that these sections occur in different areas of the sequences, for example, loop regions observed in MurC are slightly different from loop regions observed in MurE in figure 1-2. This could be a factor that may affect the structures of the Mur amide ligases. It had been stated by Barreteau et al., 2008, that the Mur amide ligases have the same structure and mechanism of action, as they all partake in an ATP-dependent reaction that adds an amino acids to a growing peptidoglycan structure, but the MSA results shown imply that there may be some diversity within the family. Yes, they may all share similar important residues within their structure, but there are slight variations that hint a uniqueness within each group of the Mur amide ligases.

The difference could come from the reality that each Mur amide ligase adds different types of amino acids to the growing peptidoglycan structure. MurC begins the sequence of reactions by catalyzing the addition of L-Ala to a UDP precursor, followed by the addition of D-Glu by MurD (Barreteau et al., 2008). Mur E and F then add meso diaminopimelate (meso -A 2 pm) and D-Ala-D-Ala respectively (Barreteau et al., 2008). To further investigate this hypothesis, motif analysis needed to be carried out, to observe the functional motifs, then map them to Mur amide ligase structures and sequences respectively, so as to identify the differences observed in the MSA results.

2.8.3 Motif analysis

To better understand the multiple sequence analysis results, motif analysis was carried out using MEME. Motifs are short, highly conserved residues that are present within sequences, that can be related to structural significance of the protein (Toma i et al., 2012). Motif search was carried out on all 32 Mur amide ligase sequences, and within each group of the family. Results for each group (MurC - MurF) are shown in Appendix A-3. A large number of motifs was searched first

(maximum of 100 motifs discovered), then analysis of e-values was done to reduce the number of motifs so as to only include statistically significant motifs. A group of motifs with the lowest e-values was then obtained for the 32 sequences as shown in table 2-1, to be analyzed further.

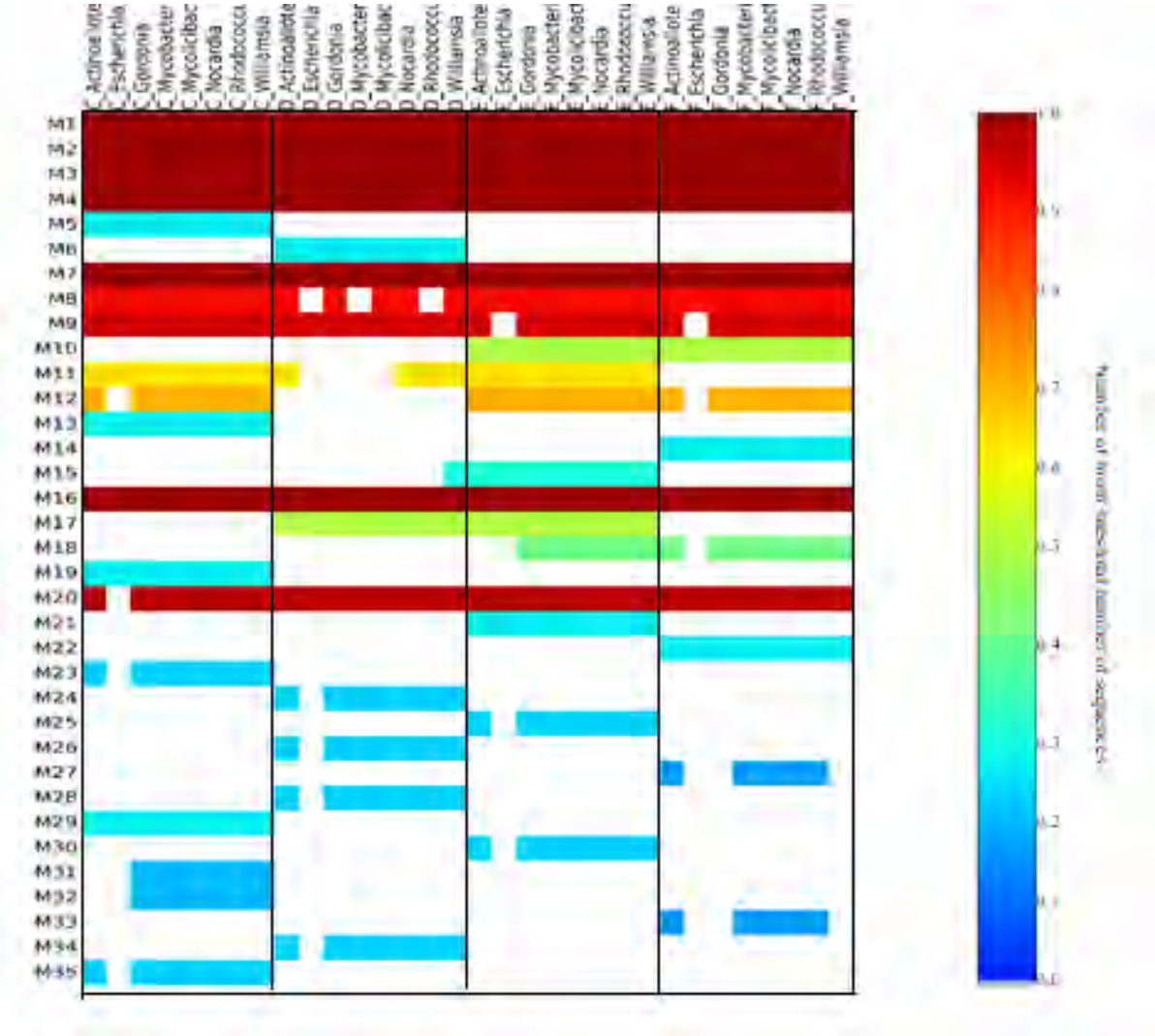


Figure 2-2: Motifs discovered within the Mur amide ligase family. MEME v4.11 used to create heat map. A total of 32 sequences observed to have 35 significant motifs discovered. Conservation increased from blue to red.

Table 2-1: MEME Motifs discovered with their respective e-values and number of sequences containing each motif.

Motif	E-value	Number of sequences with motif
1	3.9e-287	32
2	3.3e-194	32
3	2.3e-186	32
4	2.2e-109	32
5	7.3e-059	8
6	3.0e-058	8
7	4.3e-054	32
8	2.1e-079	29
9	6.1e-073	30
10	1.5e-055	16
11	3.9e-072	20
12	1.1e-055	22
13	1.7e-051	8
14	1.8e-048	8
15	9.8e-043	9
16	6.8e-037	32
17	1.1e-040	16
18	2.5e-042	13
19	2.2e-036	8
20	2.5e-040	31
21	7.8e-032	8
22	1.1e-028	8
23	8.1e-033	7
24	2.3e-031	7

25	7.7e-025	7
26	2.2e-024	7
27	4.8e-016	5
28	1.0e-015	7
29	9.4e-014	8
30	5.1e-013	7
31	3.0e-015	6
32	1.5e-011	6
33	1.8e-011	5
34	5.8e-011	7
35	7.7e-009	7

There were 35 motifs discovered and regarded as statistically significant for the 32 sequences submitted to MEME. As seen in table 2-1, the motifs were not occurring in all the sequences. A total of nine motifs were observed to be reflecting in all groups of the Mur amide ligase family, while the rest of the motifs discovered had varying number of sequences, they reflected in. This observation can also be seen in Figure 2-2, where the nine conserved motifs are visualized by a continuous red color across the sequences, indicating their presence in most of the sequences submitted to MEME. The nine motifs conserved in all Mur amide ligase groups include motifs 1, 2, 3, 4, 7, 8, 9, 16 and 20. In table 2-1, it can also be observed that some motifs only occur in seven and eight sequences. This could indicate that some groups within the Mur amide ligase family have unique motifs not found in the other groups, as each group had eight sequences in total. Analyzing figure 2-2, it is observed that indeed there are some unique motifs occurring in each of the Mur amide ligase groups. MurC contains motifs 5,13,19, 23, 29, 31 and 32 that are unique to this group, while motif 6, 24, 26, and 28 are unique to MurD. In MurE, the unique motifs observed include motifs 15, 21, 25 and 30, while in MurF, motifs 14, 22, 27 and 33 were unique within the group. There were motifs observed to only occur within two or three groups of the Mur amide ligases, but not conserved throughout the whole family. Motif 10 and motif 18 were only occurring in MurE and MurF sequences, while motif 12 occurred in MurC, MurE and MurF. Motif 11 only

occurred in Mur C and MurE, and motif 17 only occurred in MurD and MurE. The motifs observed to be common throughout the different Mur amide groups could signify conserved structural features, whereas the unique motifs discovered for each group could relate to differences in structure at those specific points. Motif analysis was also carried out within the subgroups of the Mur amide ligase family and the results (Appendix A-3), show high conservation throughout each group. The conserved and unique motifs identified were then mapped to Mur amide ligase structures to observe their location. Crystal structures of MurC, MurD, MurE and MurF were obtained from the PDB database with PDB identities 2F00, 1UAG, 1E8C and 1GG4 respectively. Each of the conserved motifs were mapped to the structures using PyMOL, and results can be seen in Figure 2-3.

Most of the motifs mapped to the central domain as observed in figure 2-3, in all the Mur amide ligases. The central domain is responsible for the binding of ATP during reactions, and the GKT motif (GXXGKT/S) is known as the binding motif for ATP (Fiuza et al., 2008). This was observed to be one of the motifs that were conserved throughout the Mur amide ligases, as motif 1 ("GLTVLGVTGTSGKTTTTSLL") contained this GKT motif. The other eight conserved motifs mapping to the central domain may belong to conserved regions within the protein that aid in catalytic processes, and are responsible for characterization of these enzymes as Mur amide ligases. Similar results to figure 2-3 were observed in figure 2-4, after mapping the motifs to the 32 ligase sequences. Most of the motifs were mapping more in the center of the sequences, which represents the central domain.

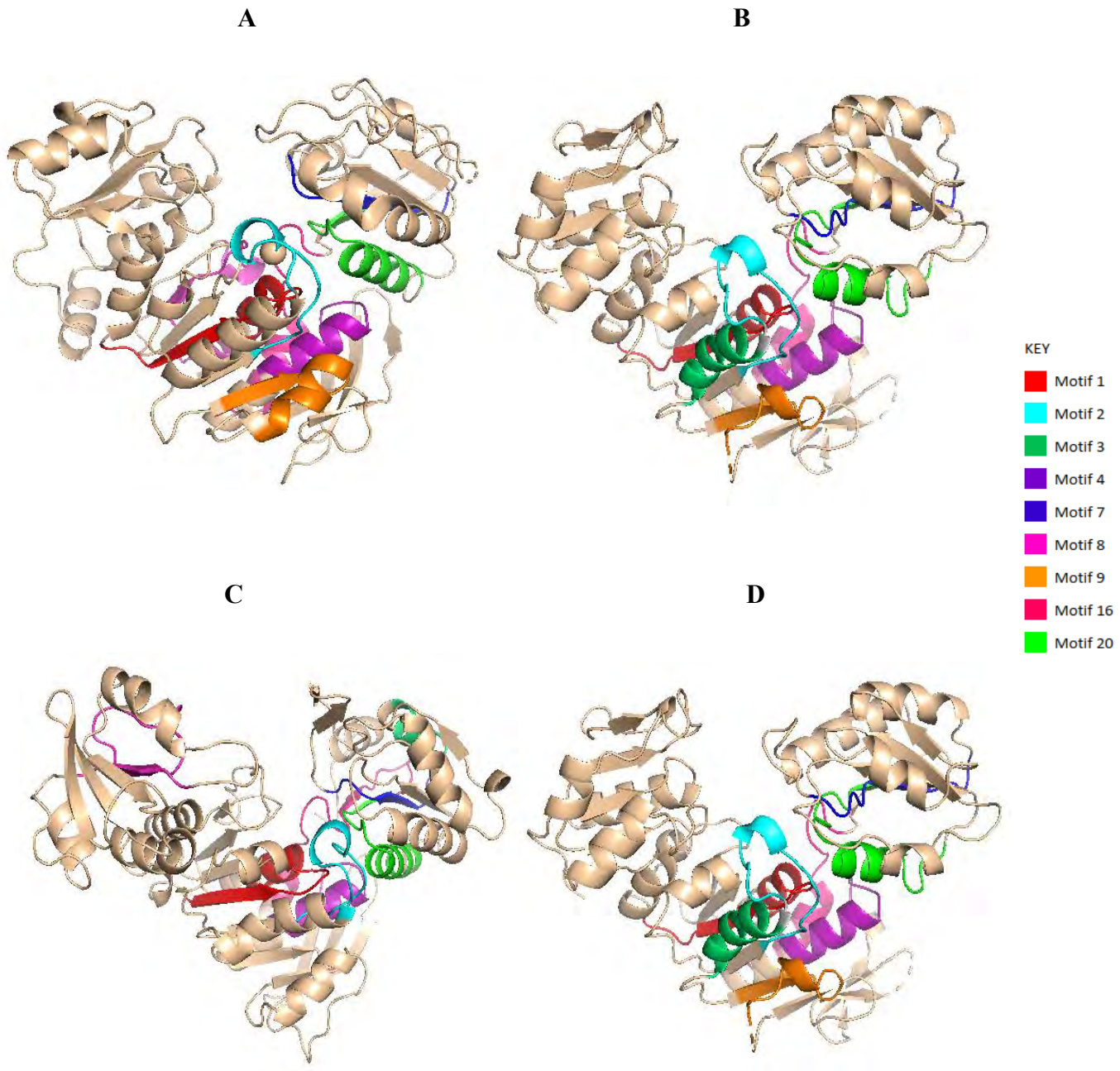


Figure 2-3: Conserved motifs mapped to Mur amide ligase structures. Conserved motifs 1, 2, 3, 4, 7, 8, 9, 16 and 20 mapped onto MurC (A), MurD (B), MurE (C) and MurF (D). PDB identities 2F00, 1UAG, 1E8C and 1GG4 used respectively for MurC, D, E and F. Motifs distinguished by color and a key to link the motif to the color on the structure.

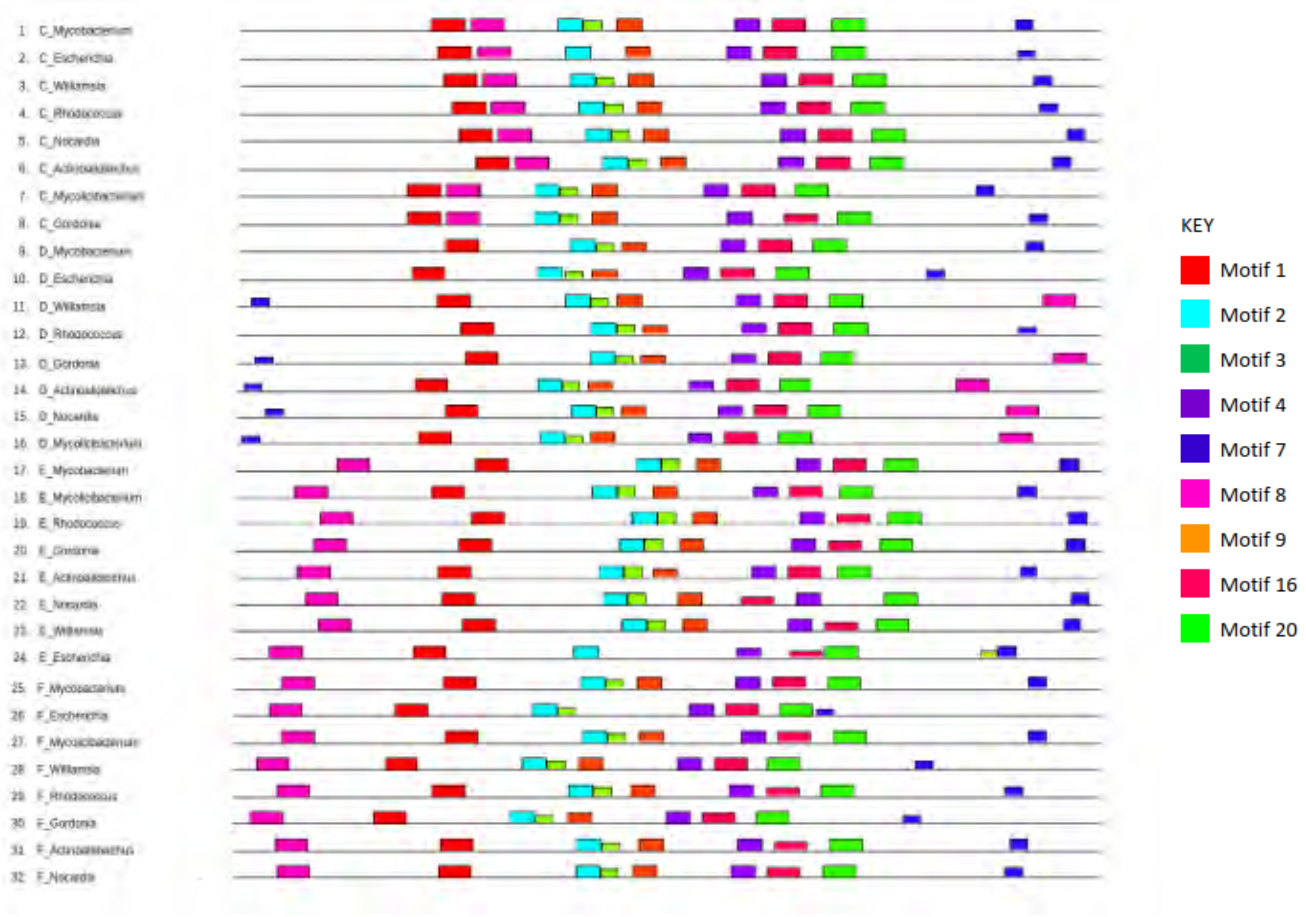


Figure 2-4: Conserved Motifs mapped to Mur amide ligase sequences. Nine conserved motifs mapped to 32 Mur amide ligase sequences. Key with colors to link motif name to location on sequences.

It was interesting to note that motif 7 and 8 were however changing their positions within groups. This could be as a result of the enzymes' specificity. Motif 7 was mapping after motif 20 in MurC, MurE and MurF, but mapping before motif 1 in MurD. Motif 8 was mapping next to motif 1 in MurC, MurE and MurF, but in MurD it was mapping after motif 20. This difference could be a result of rearrangement of sequences within the Mur amide ligase family to best fit the structure and functionality of each enzyme. The results shown by the conserved motifs mapped to structures and sequences coincide with the MSA results shown in figure 2-1, as the GKT motif and other

catalytically important residues were conserved throughout the Mur amide ligase family. This was expected as all the Mur amide ligase enzymes partake in ATP dependent reactions, and an absence of the conserved ATP binding site would result in no reactions taking place (Kouidmi et al*.,* 2014).

Unique motifs were then mapped to structures of Mur amide ligases, to observe if they can be related to the MSA results obtained. Figure 2-5 shows motif 5, 13, 19, 23, 29, 31 and 32 mapped to MurC as these were motifs uniquely found in this group. MurD had unique motifs 6, 24, 26 and 28 mapped to its structure, while MurE had motifs 15, 21, 25 and 30 mapped uniquely to its structure. Finally, MurF had unique motif 14, 22, 27 and 33 mapped to its structure. Motifs shared between ligases were also mapped to their respective structures. Mur C had motif 11 mapped to its structure, which was also common in MurE enzymes, and motif 12 mapped, that was observed to map MurE and MurF. Motif 17 was mapped to both MurD and MurE, and motif 18 was mapped to MurE and MurF. These unique motifs mapped to each group in Figure 2-5 could be evidence of differences within the Mur amide ligase structures. As stated before in MSA results, the sequences only shared <30% sequence identity, and their alignment revealed very few conserved regions within the whole family, therefore the unique motifs could be a major contributor to this observation. In figure 2-5, the unique motifs were mapping towards the C and N terminal domains, which are known variable regions within Mur amide ligases. The N terminal domain is known for binding substrates, while the C-terminal binds the amino acids, and because each Mur amide ligase enzyme catalyzes the addition of different amino acids onto slightly different substrates, the variation noticed within the N-terminal and C-terminal domain was expected (Barreteau et al., 2008). This was also visible in the MSA results (figure 2-1), as no conservation was observed in the N and C-terminals of the sequences.

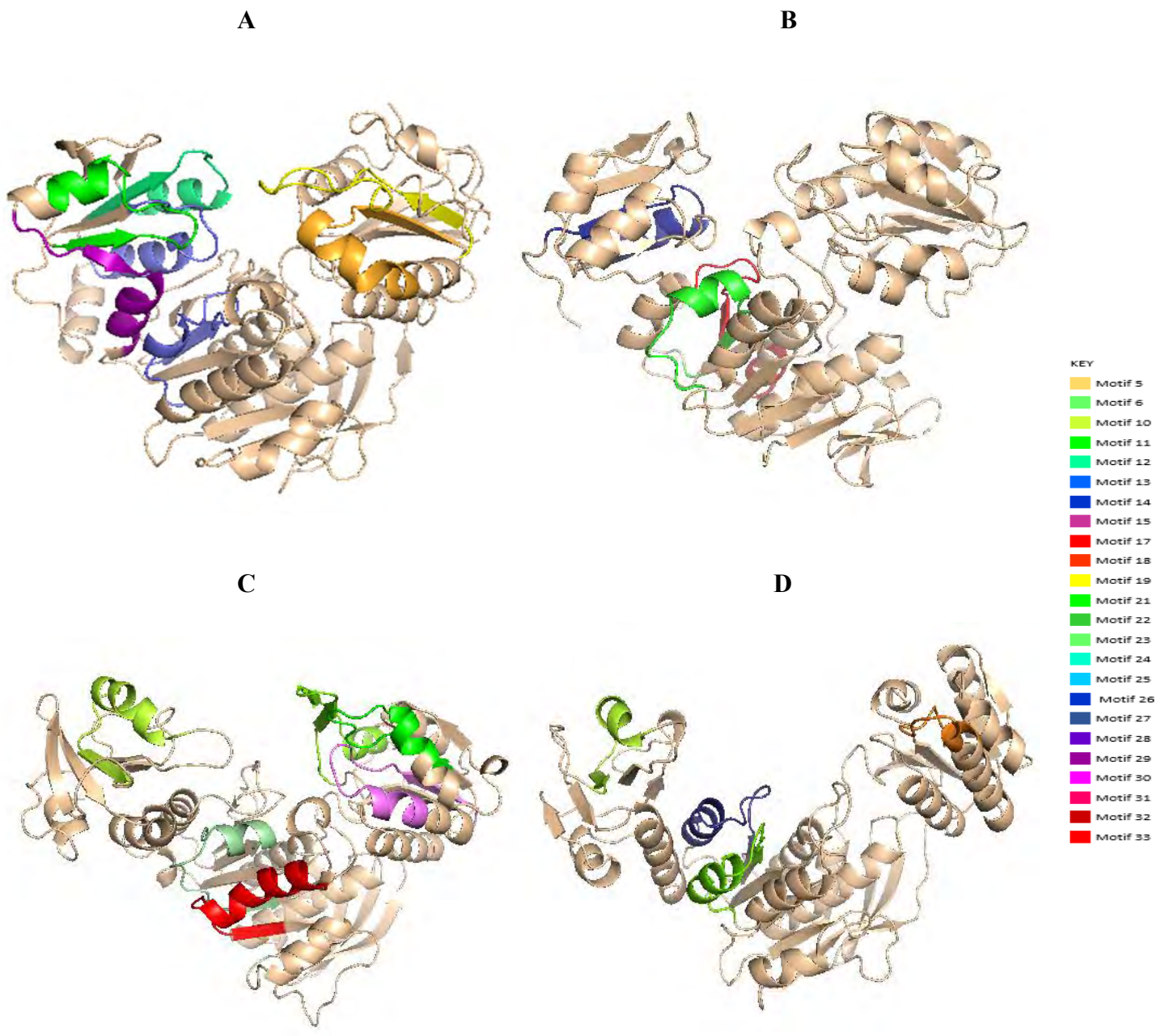


Figure 2-5: Unique motifs mapped to Mur amide ligase structures. Motifs 5, 11, 12, 13, 19, 23, 29, 31 and 32 mapped onto MurC (A). Motifs 6, 17, 24, 26 and 28 mapped to MurD (B). Motifs 10, 11, 12, 15, 17, 18, 21, 25 and 30 mapped to MurE (C). Motifs 12, 14, 18, 22, 27 and 33 mapped to MurF (D). PDB identities 2F00, 1UAG, 1E8C and 1GG4 used respectively for MurC, D, E and F. Motifs distinguished by color and a key to link the motif to the color on the structure.

The discovered unique motifs were also mapped to the 32 Mur amide ligase sequences as shown in Figure 2-6. Just as observed in figure 2-5, the motifs are mapping towards the N and C terminals of the sequences, that are regarded as variable regions within Mur amide ligases. Only two motifs were observed to be mapping within the central domain of MurD and F.

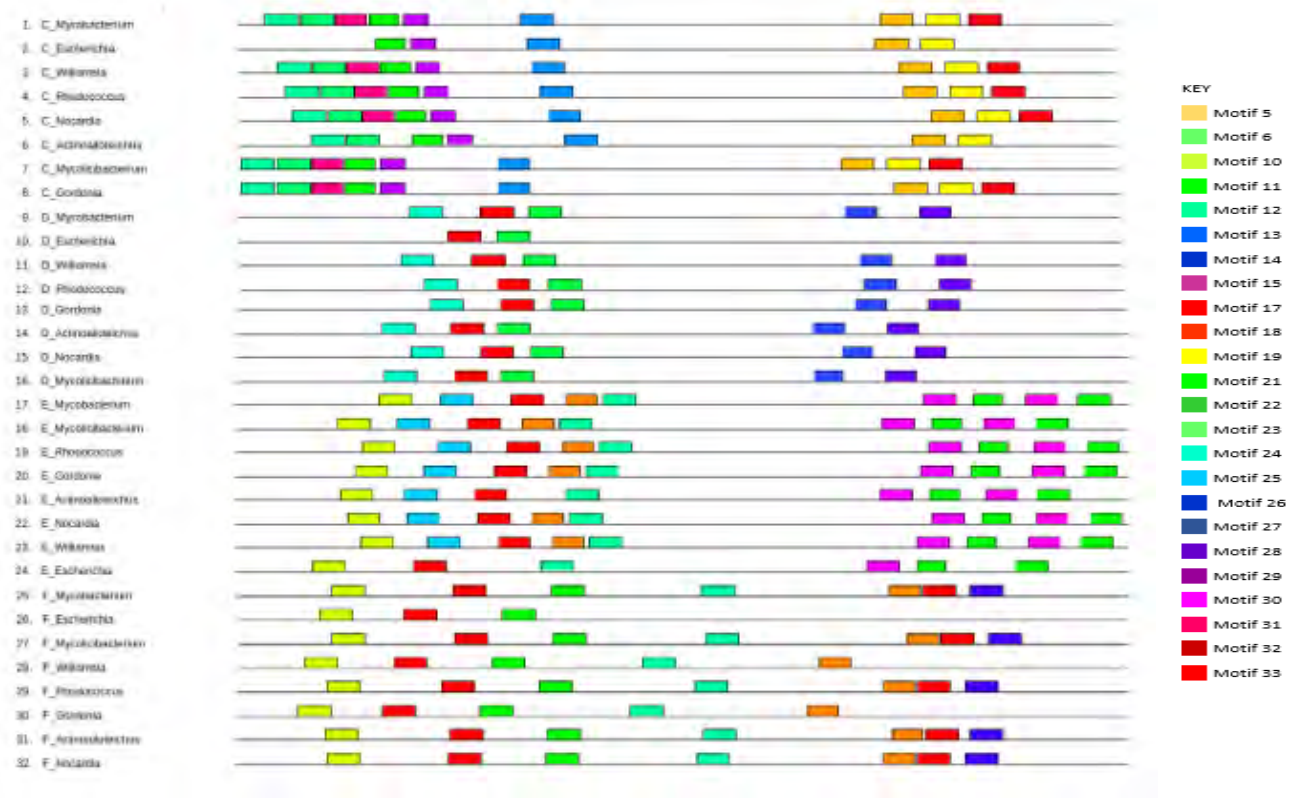


Figure 2-6: Unique motifs mapped to Mur amide ligase sequences. Motifs 5, 11, 12, 13, 19, 23, 29, 31 and 32 mapped onto MurC sequences. Motifs 6, 17, 24, 26 and 28 mapped to MurD sequences. Motifs 10, 11, 12, 15, 17, 18, 21, 25 and 30 mapped to MurE sequences and motifs 12, 14, 18, 22, 27 and 33 mapped to MurF sequences. Key with colors to link motif name to location on sequences.

The differences observed when MSA and motif analysis was conducted, have revealed that the Mur amide family of enzymes may share some conservation, but within each group of the ligases, there is a uniqueness observed. These results are consistent with findings by Patin et al., 2010, as this uniqueness of each Mur amide ligase enzyme could relate to its specificity for different

substrates. There are more type specific motifs within each group, that can be related to their structure and function.

2.8.4 Phylogenetic analysis

To conduct further type specificity within the Mur amide ligase family, phylogenetic tree analysis was carried out. This would further describe the relationships observed within the Mur amide ligases, and explain the groupings observed in motif discovery. Phylogenetic trees were constructed using MSA results in MEGAX, and a more reliable tree was chosen after comparisons were carried out with their respective bootstrap trees. The bootstrap tree for the phylogenetic tree chosen for this study is located in Appendix A-4. Prior to tree construction, models that best fit the data provided in the multiple sequence alignment of 32 Mur amide ligase sequences were obtained. Three models with the lowest BIC and AIC scores were chosen. The first model was WAG+G+F model, with a BIC score of 52957.55 and an AIC score of 52336.08. The second model was WAG+G+I+F with a BIC score of 52961.22 and an AIC score of 52332.10. The last model chosen was LG+G+F with a BIC score of 53109.20 and an AIC score of 52487.73. Three phylogenetic trees were then constructed and the model that yielded the more reliable tree was the WAG+G+I+F model. An all versus all sequence identity heatmap was also constructed with python and MATLAB, that further illustrated the information represented in the phylogenetic tree. Figure 2-7 shows the phylogenetic tree and identity heatmap, and the numbering of sequences in the heatmap was based on the branching of the phylogenetic tree.

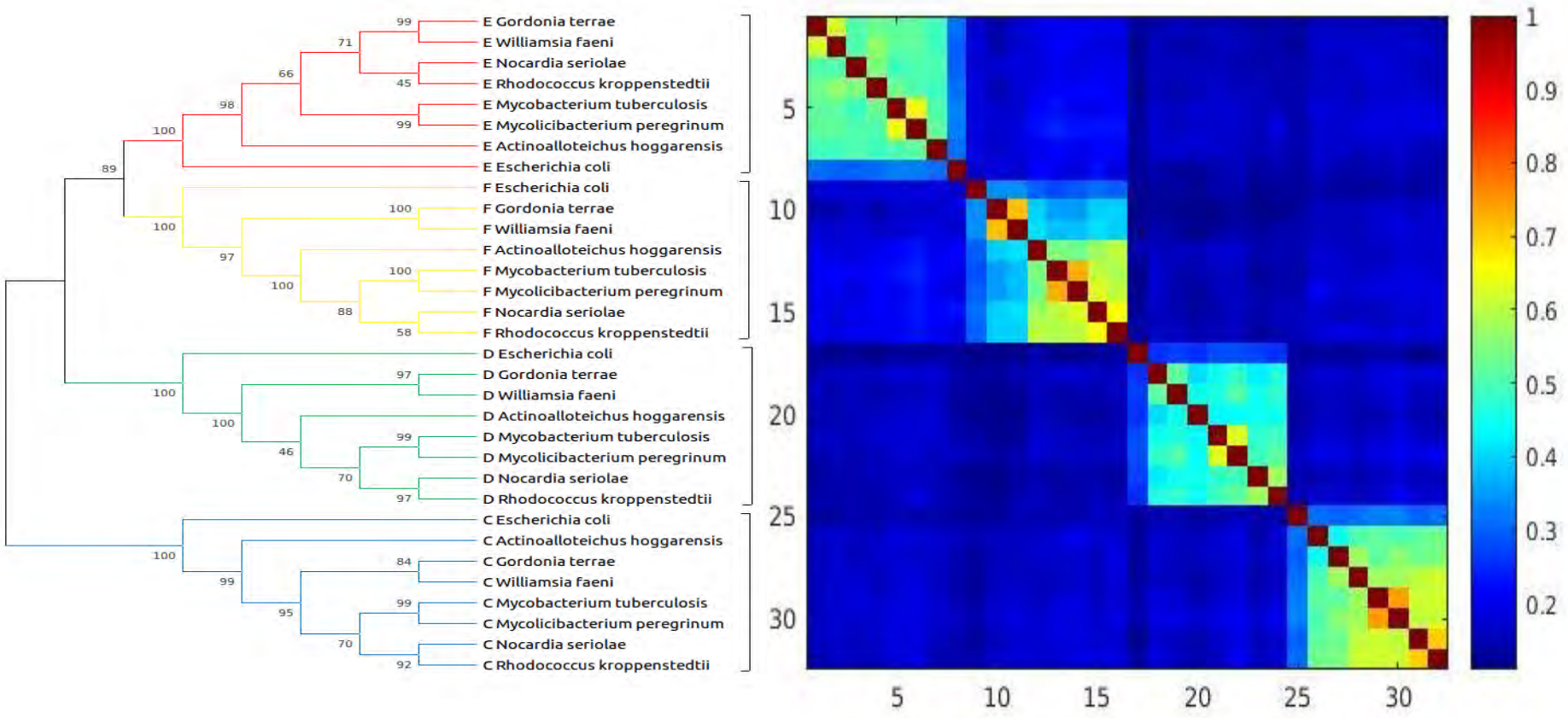


Figure 2-7: Mur amide ligase family phylogenetic tree and all vs all pairwise sequence heatmap. A total of 32 sequences used for the construction of the phylogenetic tree and heatmap. **A)** Phylogenetic tree produced using MEGAX and Maximum Likelihood method used to infer evolutionary history. Best tree shown constructed using the WAG+G+I+F model with 1000 bootstrap replicates. Next to the branches is the percentage of trees in which the associated taxa is clustered together. **B)** Mur amide ligase pairwise sequence heatmap produced using python and MATLAB. Heatmap showing identity scores for every Mur amide ligase sequence versus every Mur amide ligase sequence within the family. Conservation in heatmap increasing from blue to red.

On conducting phylogenetic analysis, groups within the Mur amide ligase clustered together, and statistical support provided by the bootstrap values on the branches was good. Most branches had a bootstrap value over 70, and only four below this number. In the phylogenetic tree shown in figure 2-7A, all sequences belonging to each of the Mur amide ligase enzymes (MurC-F), were grouped together, showing distinct divisions between the family. There were stronger relationships observed within groups rather than as a whole superfamily. This can be related to the motif analysis observed in Appendix A-3, where there is higher conservation within groups, than when motifs were discovered as a whole family as shown in figure 2-2. These observations were also captured by the all versus all sequence identity heatmap in Figure 2-7B, and small conservation boxes were formed within the matrix. There were four conservation boxes observed, representing the four types of Mur amide ligases (MurC, MurD, MurE and MurF). The conservation boxes had conservation higher than 0.5 within the heatmap, indicating a high conservations and high sequence identities within groups. There was lower conservation observed throughout the family of enzymes on the heatmap (conservation lower than 0.5). A strong red diagonal was observed in the heatmap matrix, and this indicates conservation through all the 32 sequences, although this is limited. Similar findings were observed before with MSA and motif discovery.

2.9 Chapter conclusion

At the beginning of this chapter, the main underlying objective was to characterize the Mur amide ligases, pointing out the relationships and spotting any differences within this family of enzymes. After observations of the MSA, motif discovery and phylogenetic analysis results obtained, two main points can be concluded. The first is the presence of some conservation throughout the ligase enzymes, that can be related to the structurally important catalytic portions of the enzymes. Mur amide ligases are classified as a family because of their similar structure and function in terms of catalyzing the formation of peptide bonds between different amino acids and a growing peptidoglycan chain, so it is only natural that they share conserved regions and motifs. The second take home note however, lies in the fact that there are some differences within the family, that mainly map towards the N and C variable terminals. These regions are responsible for substrate

and amino acid binding, and differences would lean towards the fact that each Mur amide ligase enzyme accepts structurally different substrates and amino acids. With this knowledge, finding inhibitors for the family as a whole may be difficult, as each active site differs slightly, but specific inhibitors for each group may be obtainable. In the next chapter, a search for these inhibitors was executed.

Chapter 3: Homology modelling and Docking

3.1 Introduction

After the realization brought about in chapter 2, that each Mur amide ligase possess its own unique features, the enzymes were subjected to experiments individually for the rest of the study. In this chapter, we focus on docking experiments undertaken to potentially find inhibitors for each Mur amide ligase. Preceding docking experiments, homology modelling of MurF was conducted, to model a missing loop region, so docking experiments carried out would be free of bias. Ligands from the SANCDB compound database were then docked blindly onto each Mur amide ligase and results reported. We begin this chapter with a brief introduction of what homology modelling and docking really entails, followed by a detailed explanation of how experiments were conducted, and a results section to report findings.

3.2 Homology modelling

Homology modelling (also known as comparative modelling) is a computational method that is used to predict a protein's 3D structure (Waterhouse et al., 2018). Many proteins have crystal structures that have been resolved experimentally by NMR spectroscopy and X-ray crystallography, but some structures unfortunately have not been resolved, or experiments leading to their discovery have failed (Waterhouse et al., 2018). In order to obtain a functional structure given these situations, homology modelling is then used utilized as a tool to produce required structures (Waterhouse et al., 2018). This involves a target sequence being aligned with a template sequence from a database of resoled 3D protein structures, that it is closely related to (Waterhouse et al., 2018). It has been proven that protein structures remain more conserved through evolution than protein sequences, therefore if the aligned query and template sequence share more than 20% identity, the likelihood of the produced structure being accurate is relatively high (Waterhouse et

al., 2018). There are four steps involved in homology modelling, and these include: template identification, target and template sequence alignment, model building and refinement, and model validation (Waterhouse et al., 2018).

3.2.1 Template identification

Given a protein sequence, the first crucial step in homology modelling is identifying the best template suited for the sequence provided (Venselar et al., 2010). Normally, performing a BLAST search against the Protein Data Bank (PDB) retrieve templates that are closely related to the query sequence by pairwise alignment (Peng and Xu, 2011). This method however is mostly successful with sequences that share homology (Peng and Xu, 2011). For more distantly related homologs, PSI-BLAST is used to perform multiple sequence alignments of templates and query sequences, resulting in more templates being identified and better templates for sequences with a distant relation to structures already resolved (Venselar et al., 2010). Once a list of templates has been produced, the first approach to obtaining a reliable template for homology modelling is to identify hits with a low E-value (Peng and Xu, 2011). These templates are considered to be close in evolution so as to construct a reliable model (Peng and Xu, 2011). Choosing a template with a poor E-value could result in a wrong structure being created by homology modelling, which is why template identification is such a crucial step (Peng and Xu, 2011). Other factors that may need to be taken into consideration when choosing a reliable template is the resolution. Generally, a template with 1.8-Å data can typically create a better structure model than with 2.2-Å data (Venselar et al., 2010). A closer look at whether a template may have a function similar to that of the query sequence, or if it belongs to a homologous operon, could aid in the selection of a reliable template (Venselar et al., 2010). Finally, as stated before, the sequence identity of the target and template is important, and templates chosen must share at least 30% sequence identity with the query sequence. However, this factor is flexible, as accurate models can still be formed when sequence identities are lower than 20%, by combining MSA and structural alignments (Bishop and Kroon, 2011). A variety of programs are then employed to align the query sequence to the template to calculate a new model, which can be checked for aberrant loop regions that alternate rotamer conformations of side chains and don't align well (Bishop and Kroon, 2011). Refinement of this

new model can be done, and several types of scoring methods can be used to measure the accuracy of the model obtained (Bishop and Kroon, 2011). To maximize the quality of model produced in all cases, more than one template can be selected (Peng and Xu, 2011). Once a template/ templates have been chosen, the next step in homology modeling is the alignment of the target sequence to the template sequence.

3.2.2 Target-template sequence alignment

Aligning a query sequence to the template sequence is important in homology modelling for the creation of a reliable 3D structure. This can be carried out by several alignment tools such as CLUSTALW or PROMALS3D (Venselar et al., 2010). Multiple sequence alignment of many templates closely related to the query sequence can aid by providing more information on functionally important residues that need to be accounted for within the structure to be produced (Venselar et al., 2010). The correctness of these alignments contributes greatly to the accuracy of the final structure to be produced, as the template sequences generate a backbone for this model (Venselar et al., 2010). The aligned residues will occupy the same position in the template and model, and coordinates will be copied over from the template to the new structure's backbone (Venselar et al., 2010). When backbone information is in place, a model can then begin being constructed.

3.2.3 Model building and refinement

There are many tools and software designed to predict a proteins structure, such as MODELLER, HH-suite and PRIMO (Sali and Blundell, 1993; Hildebrand et al., 2009; Hatherley et al., 2016). In this study, MODELLER was utilized to construct a MurF ligase, and it is a program that predicts a 3D structure from information provided in the Protein Data Bank (Sali and Blundell, 1993). It was written and developed to be implemented within python environments. There are many methods that are used by the different programs to fulfil modelling of structures, but MODELLER utilizes the modeling by satisfaction of spatial restraints method (Fiser and Sali, 2003). This method uses distance geometry or optimization techniques to satisfy the spatial restraints collected

from the target-template sequence alignment (Fiser and Sali, 2003). MODELLER specifically extracts spatial restraints from two sources, with the first one, that is homology-driven restraints on distance and dihedral angles, being from the alignment of the template and target/query sequence (Fiser and Sali, 2003). The second stereochemical restraints such as bond angles and bond length preferences, are obtained from the molecular mechanics force field of Charmm-22 (Fiser and Sali, 2003). Statistical preferences of dihedral angles and the nonbonded atomic distances are collected from a set of known protein structures (Fiser and Sali, 2003). A model under construction is then calculated using an optimization method that relies on conjugate gradients and molecular dynamics, that minimize spatial restraints violations (Fiser and Sali, 2003). This method can be related to the method used to create protein structures from NMR-derived restraints (Fiser and Sali, 2003). When a model has been built successfully, it will resemble the template structure quite well. Some minor details including accurate backbone conformation, some side-chain rotamers and hydrogen networks are often wrong, so the implementation of various refinement and optimization tools are then employed so as to make the built structure more accurate (Venselar et al., 2010). This can be carried out using MODELLER as well.

3.2.4 Model validation

Finally, once a satisfactory model has been built and refined, validation of this model is required to deem it fit for use in *in silico* experiments. The first point of validation is the use of the Z-Discrete optimized protein energy (DOPE) score (Venselar et al., 2010). A low Z-DOPE score implies a good model, but one has to compare the Z-DOPE score of the template with the Z-DOPE score of the completed structure (Venselar et al., 2010). If they are more or less similar, then the model built can be considered a good model. Other validation software such as PROCHECK, Verify3D and ProSA can then be used to further examine the quality of the model. These methods evaluate the model globally (determines the quality of the whole structure) or locally (determines the quality of each residue within the structure) (Venselar et al., 2010). When a model has been evaluated by the validation programs, it can either “pass” meaning the structure is adequately correct and complying with requirements that define a good structure, or it “fails”, implying the model created is not of satisfactory quality. Problematic regions that could result in a model failing

are variable regions that form loops within the structure, or large inserts that lack structural information (Venselar et al., 2010). If a model is not good enough, the modelling process has to be repeated, using another template or alignment, but if it passes it can well be used for further experiments (Venselar et al., 2010).

3.3 Docking

Molecular docking is a process that aims to predict binding sites for ligands and other molecules on protein surfaces (receptors) (Trott and Olson, 2010). This is a widely used process in drug discovery, for the identification of potential inhibitors for certain proteins, however it can be used in research to study ligand interactions with proteins (Trott and Olson, 2010). Large libraries of drug-like ligands can be docked to proteins, in order to obtain leads that can be carried forward during drug development (Trott and Olson, 2010). Many tools such as Knime, Maestro and Glide can be used to carry out the process of docking, but Autodock Vina was implemented in this project as it is a newer, reliable and accurate software that provides one with high performance and multi-core capabilities. Autodock Vina uses a collection of Python scripts (in Autodock tools), for the setup of docking processes (Seeliger and Groot, 2010). In order to conduct docking of ligands to proteins, it follows two steps, which are the sampling and scoring steps (Rentzsch and Renard 2015). In the sampling step, a large number of ligands are docked to the required parts of the protein, generating different ligand poses (Rentzsch and Renard 2015). Autodock Vina uses many statistical algorithms to exhaust all possible binding conformations of ligands on the receptor surface (Rentzsch and Renard 2015). Once different conformations are collected, they are then scored in the scoring step, for predicted binding affinities. The scoring function basically seeks to approximate the standard chemical potentials of the given system (Rentzsch and Renard 2015). The results normally produced include the different ligand poses and their respective binding energies, that can then be manipulated by a user (Rentzsch and Renard 2015). Docking is a fairly cheaper method as compared to traditional methods for determination of ligand binding sites such as NMR and crystallography, but it is only a prediction method (Rentzsch and Renard 2015).

3.3.1 The rule of five

Analysis of the docking results can then be performed, and determining whether the lead hits are in fact druggable is an important step in drug discovery. Lipinski's rule of five is one tool that can be utilized to determine whether a compound can be orally administered to patients (Lipinski, 2004). Chris Lipinski was one of the first people to point out that drugs typically have different physicochemical and structural properties that could render them inactive (Lipinski, 2004). Such properties that could potentially place compounds outside preferred predictors of "drug-likenesses" include the molecular weight (MW) of the compound or the number of polar atoms and polar surface areas present, just to name a few (Santos et al., 2016). The original rule of five (Ro5) founded by Lipinski and his team for development of oral drugs was actually created because too many compounds were being discovered in medicinal and combinatorial chemistry with substandard physicochemical properties (Lipinski, 2004). It states that, the molecular weight of the compound should not exceed 500 Da, the water partition coefficient (log *P*) should be less than five, the hydrogen bond donors (HBD) cannot be more than five and the hydrogen bond acceptors (HBA) cannot be more than 10 (Lipinski, 2004). Interesting fact, the rule of five was given its name because all boundary values are multiples of five. If compounds observed as inhibitors in lead optimization fail the Ro5, it is highly likely they will fail as oral drugs, or many problems will be encountered with them, but if they do pass, there is no guarantee that those compounds will be druggable (Lipinski, 2004). The Ro5 does not go into detail about specific structural features related to chemistry that are found in the compounds, so there is room for further analysis of any compounds that do pass (Lipinski, 2004).

3.4 Chapter Objectives

1. Model missing loop region in MurF ligase using MODELLER, and validate new model using ProSA, PROCHECK and Verify3D.
2. Dock ligands from the South African Natural Compounds Database (SANCDB) (https://sancdb.rubi.ru.ac.za/) onto all Mur amide ligase enzymes using Autodock Vina.
3. Validate docking of ligands onto Mur amide ligase structures in Autodock Vina.

4. Screen for any lead compounds binding in the active sites of the Mur amide ligases, to be observed with Molecular Dynamics (MD) simulations.

3.5 Methodology

3.5.1 Homology modelling

A BLAST search was conducted using the sequence of MurF obtained from UniProt, against a database of known protein structures (PDB repository). The first hit was chosen, with a low E-value and a high sequence identity (PDB structure 1GG4). The selenomethionine residues in the PDB structure were mutated to methionine using Discovery Studio, and thereafter all non-protein atoms such as water, ions and any ligands bound to the structure were removed. The template sequence and target sequence from UniProt for MurF were then aligned and MODELLER was used to model a new structure. Approximately 100 models were constructed using a standard auto model routine and slow refinement, to increase accuracy of the final model. Z-DOPE scores for the models produced were then obtained and ranked, and the model with the lowest Z-DOPE score was chosen as a possible structure for MurF. This model was validated using PROCHECK, Verify3D and ProSA, and Discovery Studio was used to observe if the missing loop region residues had been modelled correctly.

3.5.2 Docking

To being docking, protein structures for the Mur amide ligases were obtained from the PDB database (MurC: 2F00, MurD: 1UAG, MurE: 1E8C and MurF: modelled 1GG4 structure) and ligands were collected from SANCDB. Waters, ions and all ligands were removed from the protein structures using Discovery Studio, leaving apo proteins. The docking process then followed these steps; 1:Prepare Ligand ; 2:Prepare Receptor (Protein); 3: Set grid parameters; 4: Set docking parameters; 5: Docking in Autodock Vina (in CHPC cluster).

The proteins and ligands were prepared using Autodock tools (prepare_receptor4.py and prepare_ligand4.py respectively), then blind docking was carried out on all the proteins using Autodock Vina contained in supercomputers at the Centre for High Performance Computing (CHPC) in Cape Town, with an exhaustiveness of 192. Validation of docking was conducted with the same parameters, where the ligand co-crystallized with MurE was redocked onto the protein, but a higher exhaustiveness was used (384). Docking results were analyzed in Discovery Studio, and all ligands with a binding energy lower than -8 kcal/mol and binding to the active site of each enzyme were selected for further analysis. The RDKit Descriptor Calculation in Knime was used to calculate hydrogen bond donors, hydrogen bond acceptors, molecular mass and octanol-water partitions (logP) for the selected ligands, so as to be compared to Lipinski's rule of five, to determine whether the ligands are druggable. Two ligands for each Mur amide ligase structure that passed Lipinski's rule of five were then analyzed using MD simulations.

3.6 Results and Discussion

3.6.1 Homology modelling

To begin docking studies on Mur amide ligase enzymes, correct protein structures had to be obtained and utilized, for more accurate results to be produced. Analyzing all the ligase structures to be used revealed that MurF had a missing region from residue 186 to 193. This was a loop region that needed to be modelled, so the MurF ligase structure is complete before docking experiments. The region of missing residues is shown in figure 3-1. Full length sequence alignment displayed in Appendix B-1.

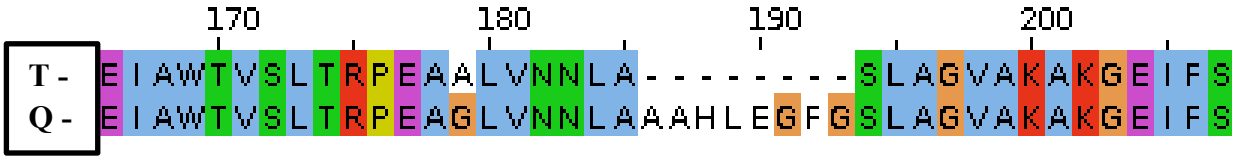


Figure 3-1: Aligned Target and template sequences of the MurF ligase. Structure 1GG4 from PDB database used as a template and the sequence from UniProt used as a target. T- Template sequence. Q- Query/Target sequence Aligned and viewed in Jalview.

Structure 1GG4 was used as a template for homology modelling of MurF and the sequence identified in UniProt for the MurF *E.coli* was used as a target. Modelling was performed with MODELLER and results are shown in figure 3-2.

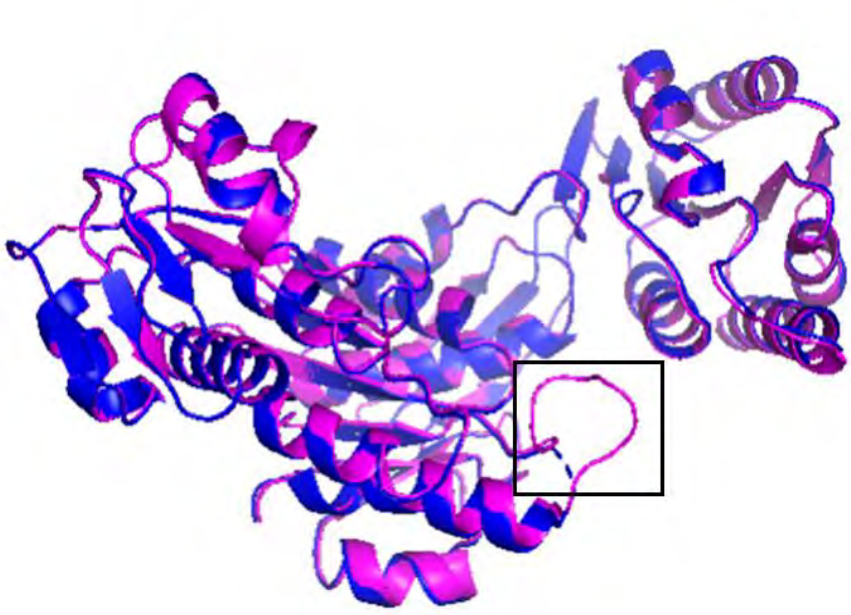


Figure 3-2: Modelled MurF ligase structure. Template 1GG4 (blue) superimposed with modelled MurF structure (purple) as depicted in PyMOL. Missing loop region modelled with MODELLER, using slow refinement.

Modelling of MurF was performed with slow refinement. A 100 models were created and ranked according to their Z-DOPE scores. The template had a normalized Z-Dope score of -1.364, and the modelled structure had a score of -1.083. In figure 3-2, the MurF ligase modelled was superimposed with the template used. It is clear from the figure that the missing loop regions were modelled in place, but further analysis into the quality of the model had to be carried out. Various tools were used to verify this model and results are shown in Figure 3-3, 3-4 and 3-5.

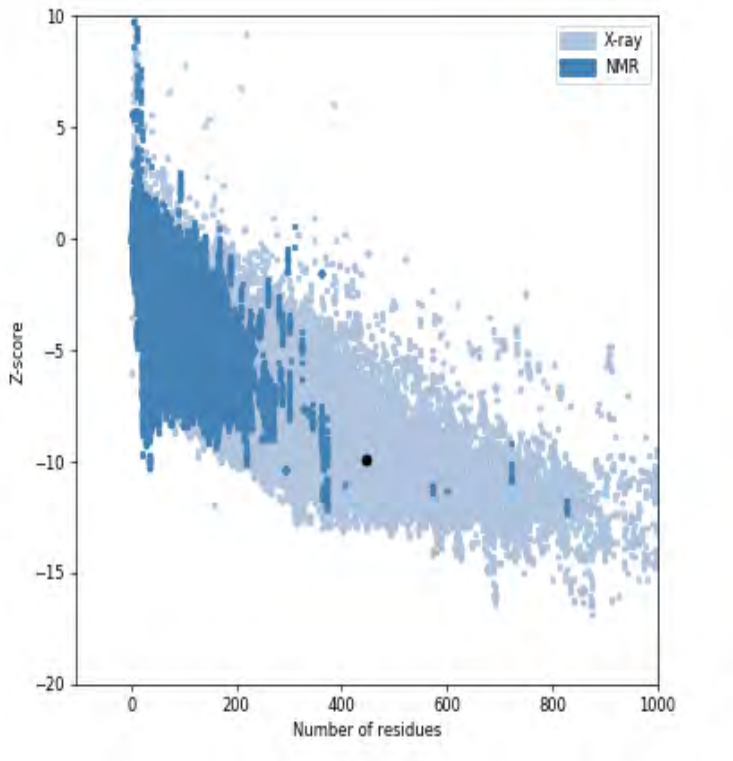


Figure 3-3: Assessment of model using ProSA. A PDB 1GG4 structure was used as a template. High resolution structures of similar sizes created with X-ray crystallography (light blue) and NMR spectroscopy (dark blue). Model falling within known X-ray structure with a Z-score of -9.94.

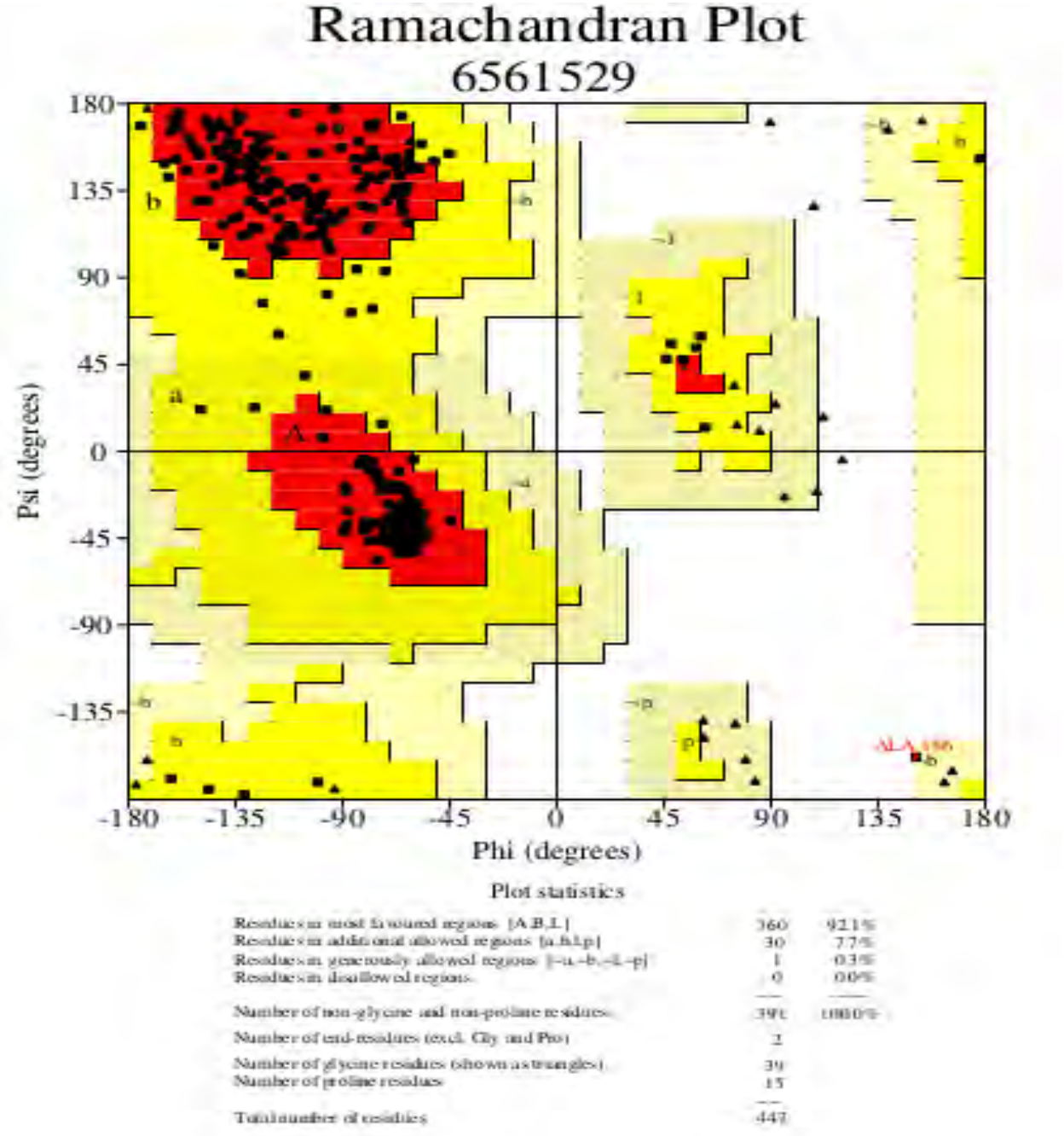


Figure 3-4: PROCHECK local quality validation. Ramachandran plot to observe quality of modelled MurF structure using MODELLER. Plot containing four sections: The most favored region, additional allowed region, generously allowed region and the disallowed region. At least 90% of the residues in the model were in the most favored region, and 0% in the disallowed region.

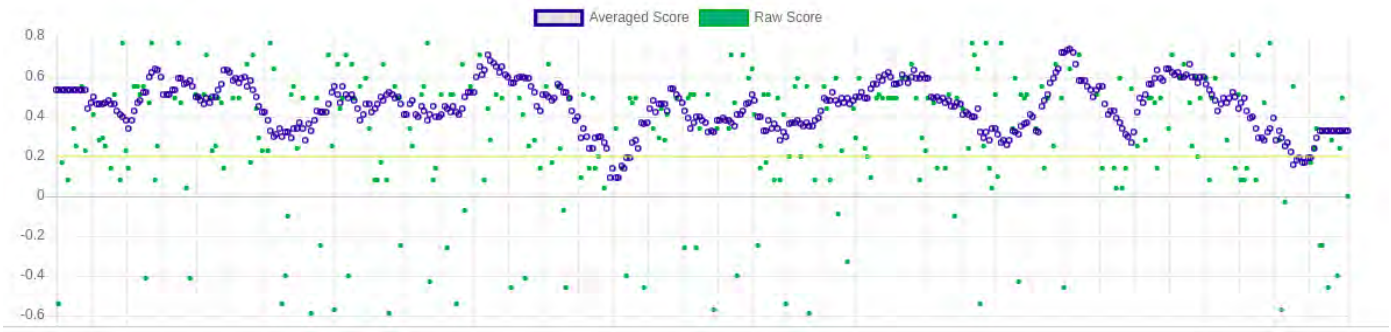


Figure 3-5: Quality of model validated with Verify3D. Averaged (blue) and raw (green) scores for the model shown. At least 80% of the amino acids scored >= 0.2 in the 3D/1D profile in all structures after averages were calculated. Model passed Verify3D.

The model passed in all three verification programs. In PROCHECK (figure 3-4), 92.1% of residues were observed to fall in the favorable regions of the Ramachandran plot, while 0% fell in the disallowed regions. The template was created using X-ray crystallography, and the modelled protein once analyzed in ProSA (figure 3-3), fell within high resolution structures of similar sizes also created with X-ray crystallography. The model had a Z-score of -9.94 which is regarded as a passing score for the quality of a model in ProSA. Finally, Verify3D revealed that 96.64% of the residues had averaged 3D-1D scores >= 0.2. These results suggested that the modelled MurF protein was a good structure for MurF to be utilized in further *in silico* experiments.

3.6.2 Molecular docking

Once the loop region within the MurF structure had been successfully modelled, docking experiments could then be conducted. The ligands and proteins were prepared, before being subjected to Autodock Vina. Blind docking of the 623 compounds obtained from the SANCDB database on each of the Mur amide ligases was performed, and results were visualized in Discovery Studio. It is known that the active site pocket of each of the Mur amide ligases falls within the N-terminal, therefore all ligands binding to the proteins at the N-terminal were selected. The chosen ligands were then narrowed down according to their binding energies (all ligands with binding

energies < -8 were used). Compounds binding to the active site and having a binding energy less than -8 kcal/mol were then subjected to calculations with an RDkit Descriptor calculator, that yielded results which could be compared to Lipinski's rule of five. In table 3-1, the series of steps in the docking process followed in this research is outlined, along with the different numbers of ligands remaining after each step.

Table 3-1: Reduction of SANCDB ligands for MD simulation through molecular docking.

	SANCDB Inhibitors	Active Site Inhibitors	Inhibitors with < -8 kcal/mol binding energy	Inhibitors passed Lipinski's rule of 5	MD simulation inhibitors
MurC	623	126	100	10	2
MurD	623	315	126	7	2
MurE	623	268	67	6	2
MurF	623	337	18	6	2

As seen in table 3-1, after comparison with Lipinski's rule of five, at most 10 inhibiting ligands were identified for each enzyme. The lowest number of inhibitors identified was six, for MurE and MurF. Many compounds did not pass Lipinski's rule of five mainly because their molecular weight was largely deviating from the preferred weight. A high quantity of ligands had molecular weights between 800 and 1,500 Da, which was much larger than the 500 Da proposed for oral drugs. In Appendix B-2, a detailed summary of the values obtained for comparison with the Ro5 is displayed.

At least two ligands from the select few that passed the Ro5 were chosen for analysis with MD simulations. They were chosen based on their binding energy and types of interactions with the active site residues. The binding free energy obtained after docking is a measure of how tightly a ligand is bound to a protein (Stroganov et al., 2008). The more negative the energy, the better the

ligand for that specific protein (Stroganov et al., 2008). Ligands that had relatively low binding energies but had unfavorable bond interactions with the proteins were discarded. All the Mur ligases in complex with their respective ligands chosen, and interactions between each protein and ligand, are represented in Figure 3-6.

In all the Mur amide ligases displayed in Figure 3-6, it is evident that each ligand is interacting with most residues within its vicinity. Each protein-ligand complex was then analyzed further, to reveal interactions with active site residues, if they existed. According to Fiuza et al., 2008, catalytically important residues for Mur amide ligases (taking an *E.coli* MurC enzyme as a model) mainly found in the active site that they identified include Asp50, Lys130, Glu174, His199, Asn293, Asn296, Arg327 and Asp351. There are however, other active site residues unique to each protein that were identified, that aid in the successful functioning of each protein. It is clear from figure 3-6 that all ligands were interacting with at least one active site residue that was important for the catalytic function of the protein. In MurC, both ligands had three interactions with active site residues, as SANC00574 and SANC00575 both interacted with Lys130 and Arg327, but SANC00574 interacted with Glu174 and SANC00575 interacted with Ala30. MurD results showed that SANC00290 only interacted with one active site residue (His183) while SANC00438 interacted with Thr321, Lys115 and Asn138. In MurE proteins, SANC00290 interacted with four active site residues which are Leu26, Ser28, Val40 and Gln190, but SANC00525 had six interactions within the active site: Leu26, Val40, His43, Ala45, Try50 and Gln190. They did however, have interactions with residues Leu26, Val40 and Gln190 in common. Finally, in MurF, both ligands only interacted with one active site residue, Thr31.

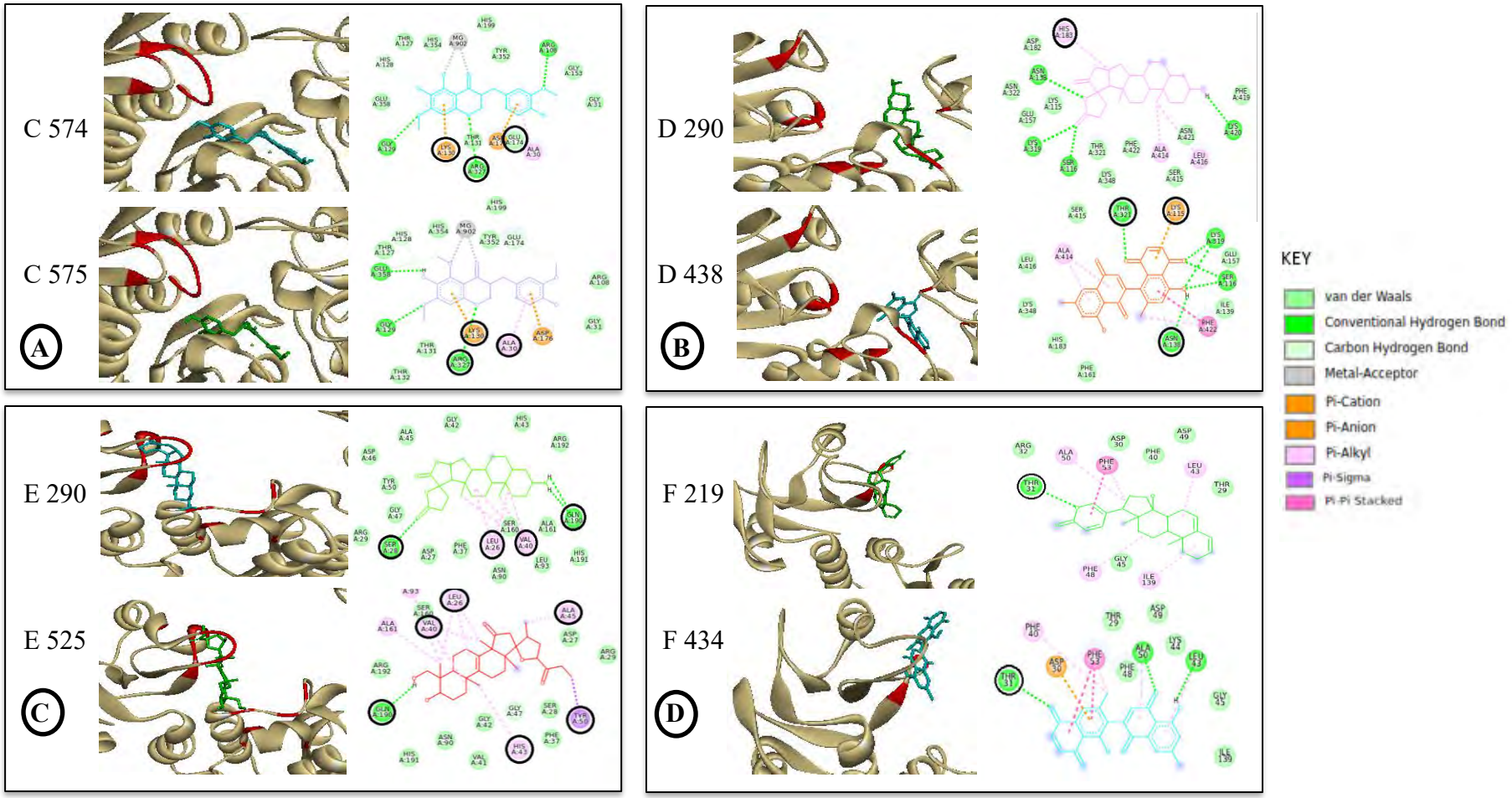


Figure 3-6: Protein-ligand interactions after docking SANCDB compounds onto Mur amide ligases. The best two ligands analyzed for each Mur amide ligase. Catalytically important residues in each active site represented in red. Active site residues interacting with ligands circled in black. A) MurC in complex with SANC00574 and SANC00575. B) MurD in complex with SANC00290 and SANC00438. C) MurE in complex with SANC00290 and SANC00525. D) MurF in complex with SANC00219 and SANC00434. Type of interaction described in a key.

It is understood that if ligands are interacting with catalytically important residues, they are somehow inhibiting the protein sufficiently, as the important residues will not be able to successfully carry out their function. In the case of Lys 130 and His199 identified by Fiuza et al., 2008, this residue is responsible for leading the substrate to the active site, therefore also aiding in the conformational change of the protein from open to closed. Without proper functioning of this residue, the Mur amide ligase would be unable to bind the substrate. Now, analyzing the results obtained within the docking experiments, it can be hypothesized that all ligands identified as hit compounds may be good inhibitors, because of the many interactions observed with catalytically important residues. What may fail these ligands as good inhibitors is the weak protein-ligand interactions observed. Many of the ligands are forming weak hydrogen and van der Waals bonds with their respective proteins, implying that they are not as tightly attached to the protein. To analyze the behavior of the proteins in complex with the ligands, MD simulations needed to be carried out. The simulations would then reveal if the ligands are tightly held within the proteins, therefore successfully inhibiting the protein.

It was however interesting to note that there were some compounds that were identified to be inhibiting all the enzymes in the Mur amide ligase family. This was unexpected as each Mur amide ligase structure contained a slightly different active site, but because of conserved regions within the active site pocket that define Mur amide ligases, this was possible. There were eight ligands universally inhibiting the Mur amide ligases and these were SANC00178, SANC00261, SANC00421, SANC00448, SANC00449, SANC00547, SANC00592 and SANC00595. As exciting as this finding may be, these ligands did not pass the Ro5 therefore could not be studied further in this research.

To validate results obtained with Autodock vina, MurE in complex with its ligand URIDINE-5'-DIPHOSPHATE-N-ACETYLMURAMOYL-L-ALANINE-D-GLUTAMATE (UAG) was utilized. The ligand was removed so as to form an apo protein, then it was re-docked using the

same parameters as those for all docking studies in this research. The results can be viewed in figure 3-7.

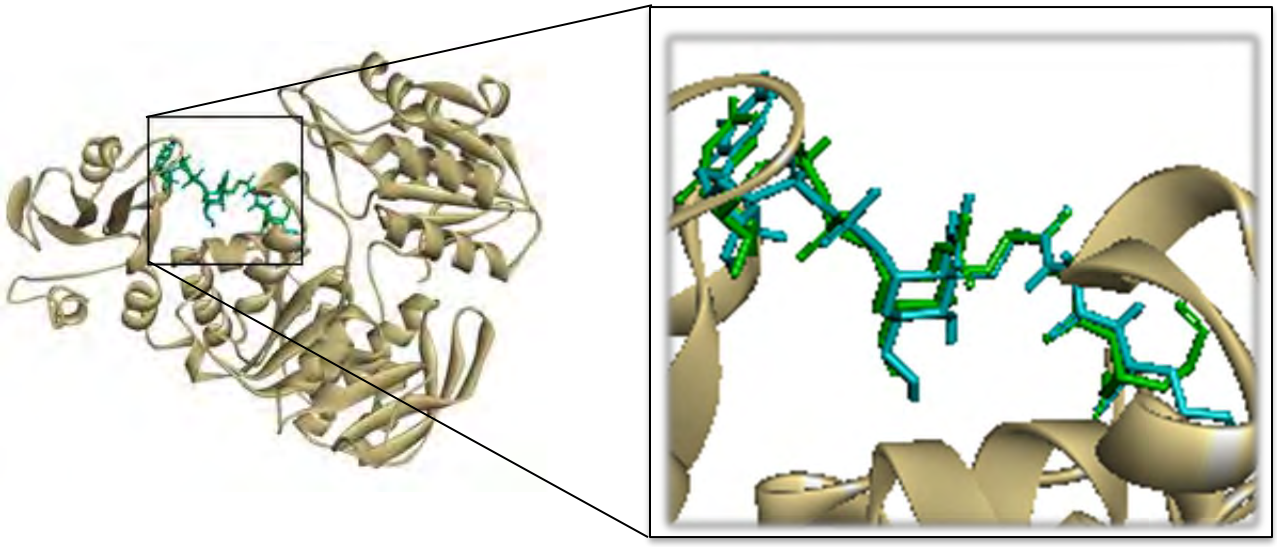


Figure 3-7: Validation for docking studies by re-docking crystallized ligand into MurE. Original ligand co-crystallized in green and re-docked ligand in blue.

The ligand managed to dock in a similar position to that of the co-crystallized ligand, therefore confirming Autodock Vina was fully functional and results obtained in the docking studies are reliable.

3.7 Chapter conclusion

It was realized in experiments undertaken in chapter 2, that each Mur amide ligase structure contained unique features that ultimately defined groups within the family, and therefore further experiments could not be performed with one enzyme representing the family as a whole, but as individual stand-alone enzymes. Within this chapter, we were able to successfully model the

missing loop region within MurF before conducting docking studies. Each Mur amide ligase was subjected to docking with ligands from the SANCDB database, and at least two ligands were obtained for each enzyme, after having undergone a form of virtual screening. The two best ligands identified for each enzyme had docked in the active site of their respective proteins, passed Lipinski's rule of five and had substantially low binding energies. Further analysis on how fixed the ligands were within the proteins needed to be carried out, so they can be considered as good or bad inhibitors of their proteins. Molecular dynamic simulations of the proteins and complexes were then performed and recorded in the next chapter.

Chapter 4: Molecular Dynamic Simulations with Mur amide ligases

4.1 Introduction

To finally observe how the Mur amide ligase enzymes, behave with and without the identified ligands from the docking procedures, molecular dynamic (MD) simulations were conducted. The main aim with the simulations was to confirm or reject compounds identified as inhibitors for each Mur amide ligase. In this chapter, we subjected each Mur amide ligase enzyme and its two docked ligands to MD simulations, then observed the trajectories produced. A total of 12 MD simulations were run and the results were analyzed using RMSD, RMSF and Ratio of Gyration calculations. Molecular dynamic simulations are briefly introduced within this chapter, followed by a detailed methodology and a discussion of results obtained.

4.2 Molecular Dynamic Simulations with GROMACS

In bioinformatics, once lead hit are obtained from docking studies, an observation of how the proteins function with and without hit ligands bound is important, so as to regard the hit ligands as good or bad inhibitors of the proteins. To do this, molecular dynamic (MD) simulations are performed. These are methods employed to simulate molecular motion, mainly for large biomolecules such as nucleic acids or proteins, by an iterative application of Newton's law of motion (Lindahl et al., 2001). There are multiple open-source packages available for MD simulations, such as CHARMM and NAMD, but in this research, the GROMACS (GROningen MAchine for Chemical Simulation) software suit is used.

The GROMACS package is a collection of many libraries and programs that carry out MD simulations (Van Der Spoel et al., 2005). It is normally used on biological molecules with visible and complex bonded interactions, but calculations for non- bonded interactions can be carried out with GROMACS, making it a very attractive software for all kinds of MD simulations on pair potentials (Van Der Spoel et al., 2005). The creation of this package began in the 1980s, at the University of Groningen in the Netherlands and most of the code is written in the C programming language (Van Der Spoel et al., 2005). It consists of about 75 executable programs, of which most of them are tools designed for analysis of trajectories or energy data that has been generated by the MD simulations (Van Der Spoel et al., 2005). Some of the programs useful for setting up simulations include the pdb2gmx, genbox, genion and x2top (Lindahl et al., 2001). The pdb2gmx is a program that generates molecular topology files from PDB coordinate files, and assigns angles, bonds, dihedrals and charges depending on a residue topology. The default settings are set according to pH 7 for protonation states of ionizable groups, but these can be manually changed, when running specific runs (Lindahl et al., 2001). The genbox program is responsible for the solution of molecules in boxes of different shapes. These boxes define periodic boundary conditions for the MD simulation (Lindahl et al., 2001). The programs genion and x2top place ions at favorable electrostatic potential positions, and generate topologies from raw coordinates after detecting the different bonds and angles present, respectively (Lindahl et al., 2001). Additional programs within GROMACS are then employed to carry out MD simulations.

4.2.1 Steps for MD simulations

There are five steps involved in the production of MD simulations. The first step is the setup stage, where a topology file, a GRO structure file and a position restraint file are created in that order. All the information required to describe the protein is stored in the topology file, such as atom masses, bond angles and lengths and charges (Lindahl et al., 2001). Once a topology file is created, a force field and water molecules can be selected in order for GROMACS or any other MD simulation program being used to run topology calculations (Lindahl et al., 2001). When a GRO file is created, the structure of the protein is basically stored within this file. It also stores information regarding the centered structure within a unit cell box that can be chosen, and

examples of boxes available for MD simulations include a triclinic box, a hexagonal prism, dodecahedrons and octahedrons (Lindahl et al., 2001). A position restraint file is the last file to be created in the setup stage of MD simulations, and this file is created so as to be used in the equilibration steps. The second stage in the MD simulation process is protein solvation. Ions and water molecules are added to the structure and topology files, filling the unit cell that was chosen in the first step (Lindahl et al., 2001). Sodium and chloride ions are used to neutralize the system, as sodium ions add positively charged ions to a negatively charged system, and chloride ions add negatively charged ions to a positive system. Energy minimization is the third stage in MD simulation development, that aims to remove any steric clashes or unusual geometry observed to be raising the energy of the system (Lindahl et al., 2001). The default settings can be utilized in this stage, but another option of uploading an MDP file with custom parameters can be chosen.

In the fourth stage, equilibration of the solvent around the protein is a must, and this is performed in two steps. The first step is the equilibration under the NVT ensemble (also known as the isothermal-isochoric ensemble), that maintains the number of particles, volume and temperature within the system (Lindahl et al., 2001). The second type of equilibration is with the NPT ensemble (also known as isothermal-isobaric ensemble) that maintains a constant number of particles, pressure and temperature (Lindahl et al., 2001). During this equilibration stage, the protein must be held in one position so that the solvent can move around it. This then employs the position restraint file created in the setup stage to place the protein in a fixed position while equilibration of the system is performed. When equilibration is complete, the protein is freed from all restraints and the final stage of MD simulations begins. Trajectories are then created, and results can be analyzed with visualization tools such as VMD.

4.3 Chapter objectives

1. Conduct MD simulations on all apo proteins of Mur amide ligase enzymes.
2. Run MD simulations on the Mur amide ligases with docked ligands within the structure to observe the behavior of proteins with and without identified compounds from docking studies.

4.4 Methodology

4.4.1 Molecular dynamic simulation setup and trajectory production

To observe the behavior of the Mur amide ligase enzymes with lead compounds identified in docking studies, MD simulations were set up and produced using GROMACS v5.1.2 in the CHPC cluster. Each Mur amide ligase was subjected to MD simulations as apo proteins and in complex with each of the compounds identified. Ligands SANC00219, SANC00290, SANC00434, SANC00438, SANC00525, SANC00574 and SANC00575 were used in this study. Overall, 12 MD simulations were performed, and the parameters were kept constant in all the simulations. A temperature of 300K was used, with a neutral pH 7 and 1 bar pressure. The GROMOS96 force field was utilized, and topology files were generated for all proteins and ligands using the GROMACS pdb2gmx tool and the PRODRG software respectively. A cubic box type was implemented for periodic boundary conditions with a solute-box distance of 1.0 nm, allowing at least 2 nm distances in the box between any two periodic images of a protein. The systems were solvated with spc216.gro water models, and neutralized by replacing a proportion of water molecules with sodium (NA+) and chloride (Cl−) ions. Using the steepest descent minimization algorithm, minimization was carried out with an energy step size of 0.0001 and a maximum number of steps set at 50,000. Minimization was performed for each enzyme until the maximum force was greater than 1000.0 kJ/mol/nm. Equilibration of each of the systems was performed using the NVT and NPT ensembles, with position restraint on all proteins and ligands. In NVT equilibration, the reference temperature was set to 300 K with a coupling time constant of 0.1 ps,

whereas in the NPT equilibration, 1 bar pressure and a coupling time constant of 2.0 ps were utilized. Finally, the MD simulations were performed for 100 ns, with a time step of 0.002 ps and a total of 50,000,000 steps.

4.4.2 Analysis of Trajectories

After MD production runs were complete, trajectories were produced and viewed in VMD. The root mean square deviation (RMSD), radius of gyration (Rg) and the root mean square fluctuation (RMSF) were then calculated using the GROMACS tools gmx_rms, gmx_gyrate and gmx_msf respectively. These were plotted graphically using xmgrace and analyzed.

4.5 Results and Discussion

Docking studies from chapter 3 yielded at least seven ligands from the SANCDB database for further development as potential inhibitors of the Mur amide ligases. For these hit ligands to be considered good inhibitors and carried further in the drug discovery process, MD simulations had to be performed in order to observe the behavior of the ligases in complex with the identified inhibitors. A total of 12 MD simulations were produced in the CHPC cluster using GROMACS, and each simulation was 100 ns long. The RMSD, RMSF and Rg were calculated for each simulation, and graphs were plotted using xmgrace, to graphically observe what the MD simulations entailed. In Figure 4-1, the RMSDs for all the Mur amide ligase enzymes are shown.

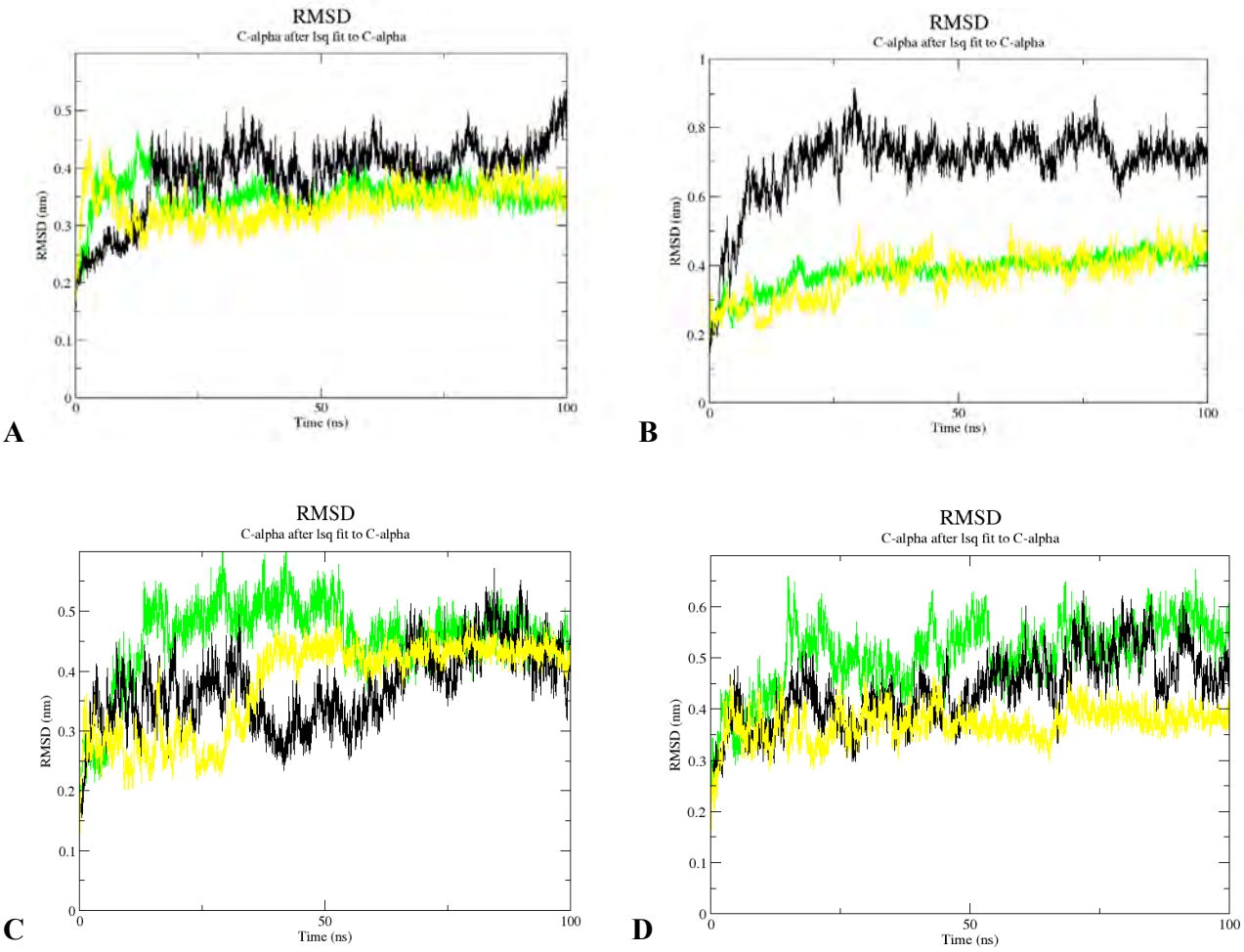


Figure 4-1: The root mean square deviation (RMSD) graphs for all Mur amide ligases. All apo proteins represented in black. Complexes in green: MurC-SANC575, MurD-SANC438, MurE-SANC525 and MurF-SANC434. Complexes in yellow: MurC-SANC574, MurD-SANC290, MurE-SANC290 and MurF-SANC219. A) MurC. B) MurD. C) MurE. D) MurF. RMSD calculated using GROMACS and plotted using xmgrace. Trajectory 100 ns long with a time step of 0.002 ps.

Root mean square deviation is the measurement of an average distance between backbone atoms such as the Cα backbone atoms (Kuzmanic and Zagrovic, 2010). With these distances plotted, a graphical representation of the change in structure throughout an MD simulation aids in

determining whether the protein is stable (Kuzmanic and Zagrovic, 2010). A continuous increase in RMSD indicates sustained changes within the structure and highly fluctuating distances plotted on the graphs indicate low stability of the protein (Kuzmanic and Zagrovic, 2010). Most flexible regions such as loop regions and unwound termini can contribute greatly to the overall RMSD of a structure (Kuzmanic and Zagrovic, 2010). In figure 4-1A, the RMSD of the apo protein for MurC was slightly higher than that observed for the protein in complex with each of its two inhibitors, SANC00574 and SANC00575. The stability of the proteins in complex with ligands was greater than the stability of the apo protein. This could be because the Mur amide ligases are known to exist in open conformations when ligands are not bound (Barreteau et al*.,* 2008). The open conformation is more unstable, and a freedom of movement is allowed, as the protein awaits to bind its substrate (Barreteau et al*.,* 2008). Similar results were observed in MurD and MurE RMSD graphs, where significant differences in the RMSD of the apo proteins and complexes existed. Of the three RMSD plots observed (figure 4-1A, 4-1B and 4-1C), MurD in complex with SANC00438 was the most stable and less changing structure, indicating that SANC00438 may be a good inhibitor for MurD.

There were much less fluctuations in distances observed and the overall RMSD was lower significantly, than the apo protein RMSD. For MurE in complex with SANC00525 and SANC00290, the RMSD first increased, higher than that of the apo protein, indicating unstable states of the complexes. Only later in each simulation, did the RMSD of the complexes decrease and stabilize, indicating that the inhibitor may need a longer duration to sufficiently inhibit MurE. Analyzing the MurF RMSD plots, the complex with SANC00434 was in fact causing the protein to become more unstable as the RMSD increased higher than that of the apo protein. In complex with SANC00219 however, MurF apo protein and the complex were behaving similarly and there was no significant change observed for most of the simulation, but a slight difference was observed towards the end of the simulation. A further look into the RMSF and Rg was needed in order to deduct conclusions, as we could not rely on RMSD alone. Figure 4-2 shows results obtained for RMSF, for each Mur amide ligase.

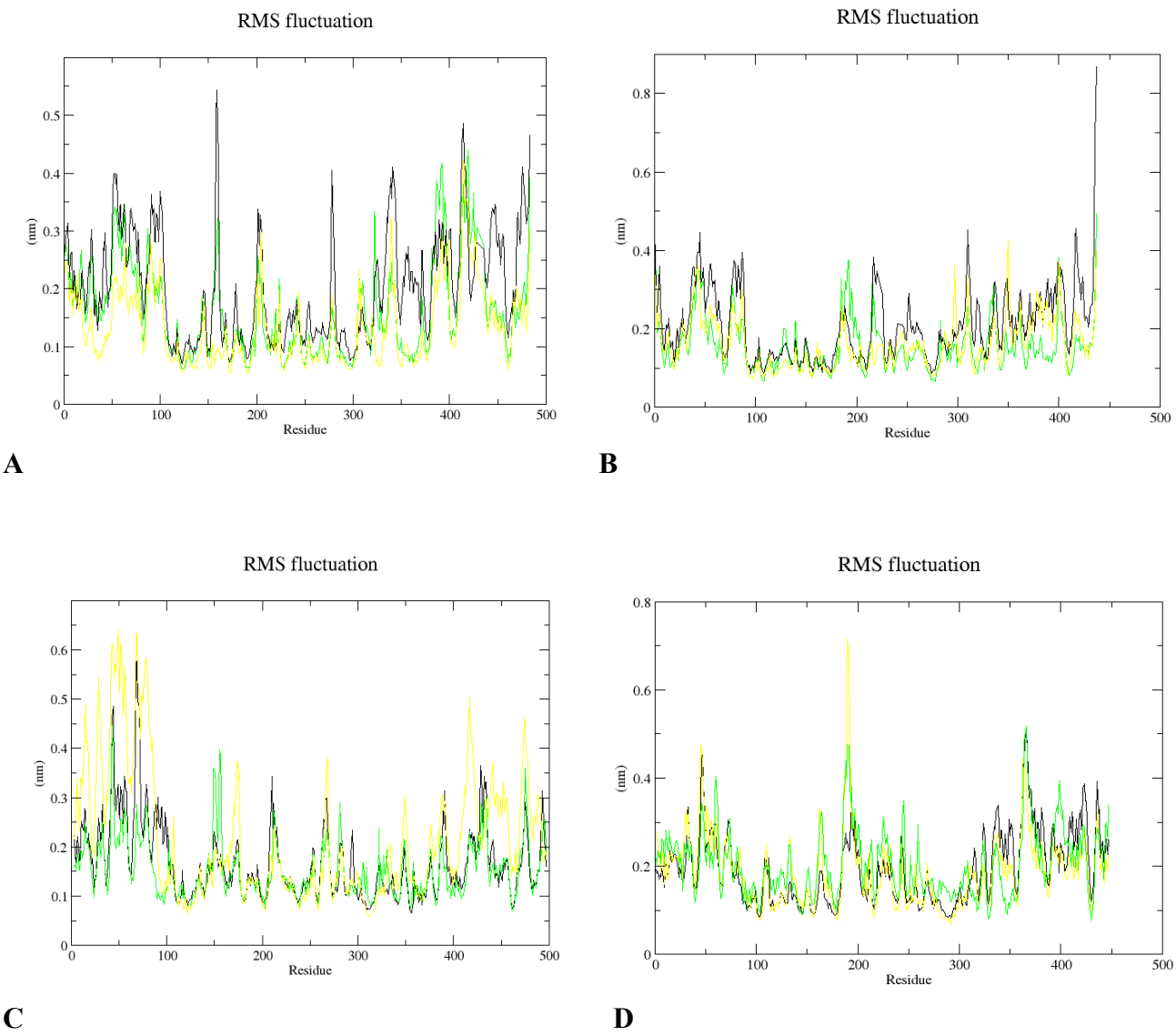


Figure 4-2: The root mean square fluctuation (RMSF) graphs for all Mur amide ligases. All apo proteins represented in black. Complexes in green: MurC-SANC575, MurD-SANC438, MurE-SANC525 and MurF-SANC219. Complexes in yellow: MurC-SANC574, MurD-SANC290, MurE-SANC290 and MurF-SANC434. A) MurC. B) MurD. C) MurE. D) MurF. RMSD calculated using GROMACS and plotted using xmgrace. Trajectory 100 ns long with a time step of 0.002 ps.

The root mean square fluctuation is a measure of the local chain flexibility by measuring each residues flexibility based on Cα atoms (Farmer et al., 2017). The standard deviation of the atoms from an average structure produces data that can be observed on RMSF plots (Farmer et al., 2017). As seen in figure 4-2A and B, RMS fluctuation values for MurC and MurD complexes, are lower than that of the ligand free structures. There are however certain residues with higher fluctuations observed, such as residues 395-398 in MurC complex with SANC574 and residue 299 in the MurD complex with SANC00290. These were part of loop regions, and most loop regions are highly flexible. The other residues were fairly less flexible as compared with the apo protein residues. In MurE, the RMSF results showed high flexibility of residues than the apo protein, for both complexes, in most regions of the protein. The proteins appeared to be highly flexible mainly at the N-terminals, and this is where the active site of the proteins is located. The high flexibility observed in the N-terminal could imply that the docked ligand was not entirely inhibiting the protein, and there were weak interactions between the ligand and active site pocket residues. As displayed in chapter 3, MurE inhibitors identified mainly had weak hydrogen and alkyl interactions with the protein, so this would explain why there is high flexibility observed within the N-terminal. The RMSF of MurF complexes displayed in figure 4-2D, were completely more flexible as compared to the apo protein. This showed that throughout the simulation, the inhibitors did not manage to stabilize the protein, and residues were allowed to freely move.

A final analysis of the ratio of gyration for each system would determine which ligands were successful within the MD simulations and which ligands failed. Ratio of gyration is the measurement of how compact or loose the protein is during the simulation (Lobanov et al., 2008). It is calculated by measuring the average distances of the selected atoms from their center of mass (Cα atoms) (Lobanov et al., 2008). Figure 4-3 clearly displays Rg results of MD simulations performed.

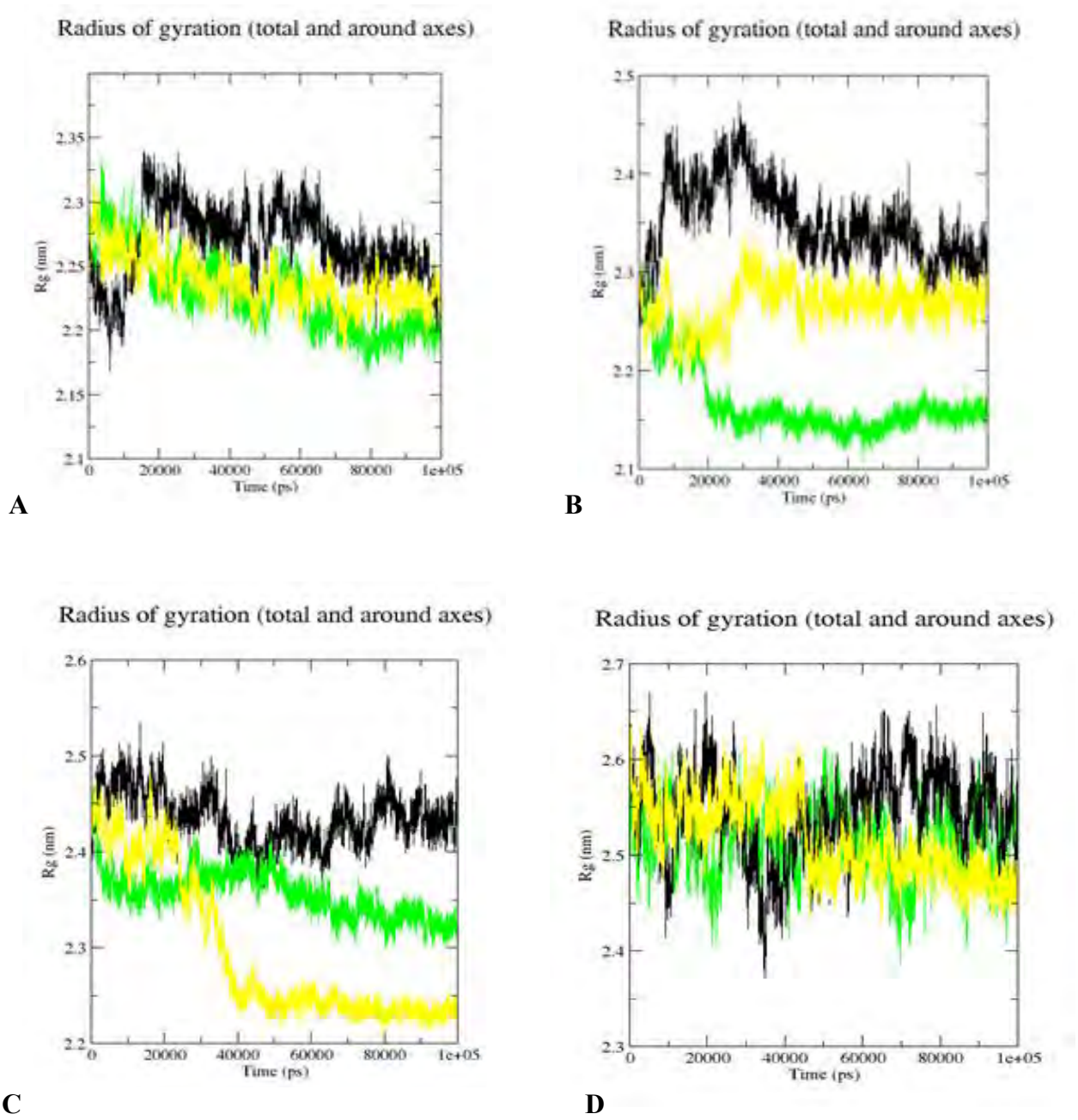


Figure 4-3: Radius of gyration (Rg) graphs for all Mur amide ligases. All apo proteins represented in black. Complexes in green: MurC-SANC575, MurD-SANC438, MurE-SANC525 and MurF-SANC434. Complexes in yellow: MurC-SANC574, MurD-SANC290, MurE-SANC290 and MurF-SANC219. A) MurC. B) MurD. C) MurE. D) MurF. RMSD calculated using GROMACS and plotted using xmgrace. Trajectory 100 ns long with a time step of 0.002 ps.

In figure 4-3, a similar trend can be observed for MurC, MurD and MurE. It is clear that the ligand bound proteins are significantly becoming more compact than the apo proteins. The Rg of MurD in complex with SANC00438 was the most significant, and this result correlated with observations of this complex's RMSD. MurF complexes however, did not show significant compactness as compared with the ligand free structure. The complexes had a Rg very similar to that of the apo protein, implying that the MurF complexes were behaving similarly to the apo protein and ligands were not entirely causing a noticeable change in structure. As mentioned before, the Mur ligases exist in open conformations when unbound to a ligand, but once their substrates bind and reactions begin, the protein enters a closed conformation state (Barreteau et al*.,* 2008). This can be observed with the Rg results shown in figure 4-3. As MurC, MurD and MurE complexes compact, the Rg of the systems decreases. The compacting could be the proteins shifting from an open to a closed conformation. Analysis of this structure change due to inhibitors had to be further analyzed in VMD and results are shown in figure 4-4.

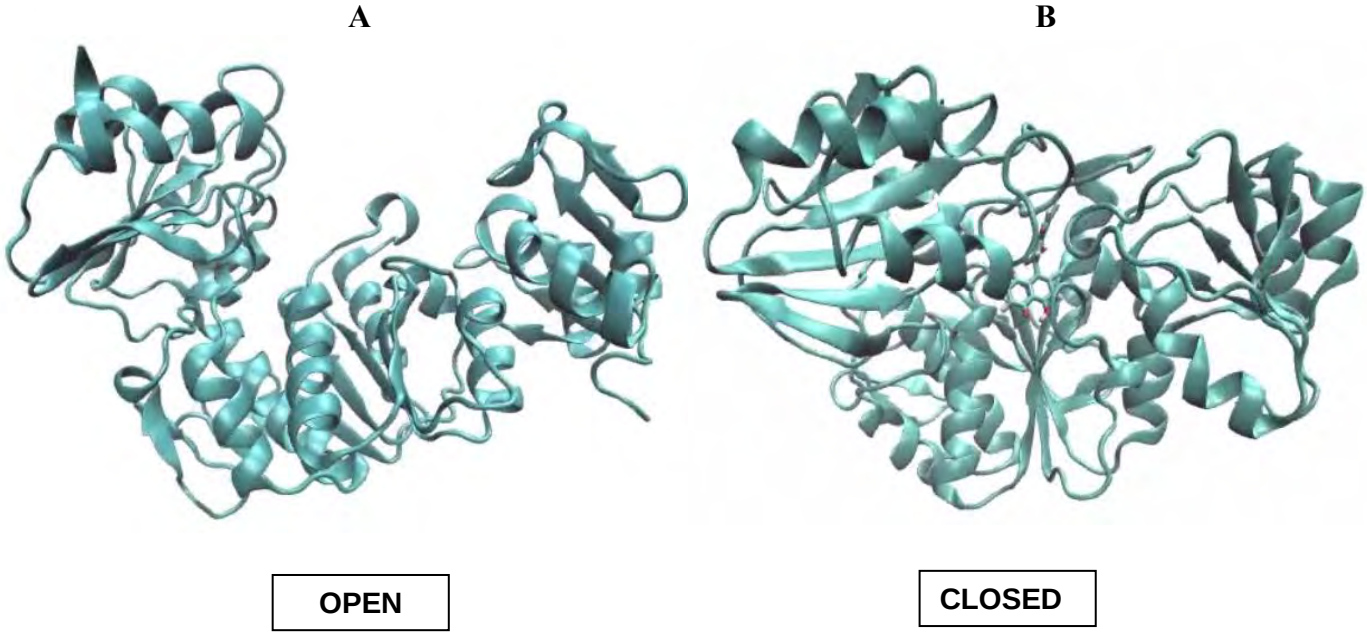


Figure 4-4: Open and closed conformations of MurD. Screenshots taken during visualization of apo and complex trajectories in VMD. A) MurD apo protein in open conformation. B) MurD in complex with SANC00438 in closed conformation.

As seen in figure 4-4A, the protein existed in an open conformation when no ligand was bound. When SANC00438 was docked into MurD, the conformation changed and became more closed. This was observed in all the proteins with ligands docked and analyzed with MD simulations except for MurF ligands, that did not cause a significant conformational change. The visual analysis of the trajectories in VMD correlated with results shown on RMSD, RMSF and Rg plots.

4.6 Chapter conclusion

Ligands identified in docking studies were subjected to MD simulations in complex with their respective proteins. Weak interactions between some of the ligands and proteins identified had created the hypothesis that the ligands would fail MD simulations and therefore would not be used in drug discovery of antibacterial agents. This was true for MurF ligands, that mainly exhibited signs of poor bonding to the MurF protein, as the protein became more flexible, unstable and remained in an open conformation when bound to ligands. MurC, D and E however, were less flexible, more stable and changed to a closed conformation upon ligand binding. These were taken as positive results of the research, with SANC00438 showing the best inhibition for MurD. Although five ligands were identified to be good inhibitors of MurC-MurE, further investigation would need to be conducted on other ligands that may have inhibited the proteins, but were excluded because of lower binding energies after docking in this research, due to time constraints.

Chapter 5: Concluding remarks

5.1 Overall conclusion and future work

Beginning this project, the main aim was to characterize the Mur amide ligases, pointing out the relationships and spotting any differences, then performing lead optimization, where SANCDB compounds would be docked into each Mur amide ligase, and results analyzed with MD simulations. After observations of the MSA, motif discovery and phylogenetic analysis results obtained, it was concluded that there was a presence of conserved regions throughout the ligase enzymes, that can be related to the structurally important catalytic portions of the enzymes. It is only natural that these conserved regions are present because all the Mur amide ligases do belong to a family of enzymes that share similarities. Furthermore, the conserved regions are mainly the ATP binding domains within each Mur amide ligase that allow for the binding of ATP which plays a crucial role in the catalytic cycle of the enzymes. Without this binding site in any of the Mur ligases would result in the loss of function of the enzymes. Another observation noted when characterizing the ligases was the presence of unique patterns within each enzyme. These regions mainly mapped towards the N and C variable terminals, that are responsible for substrate and amino acid binding respectively. These differences are supported by the fact that each Mur amide ligase enzyme accepts structurally different substrates and amino acids. With this knowledge, using one protein as a representative for the whole family to find inhibitors would be difficult, as each enzyme proved to have specificity towards slightly different substrates. Thus, each protein was studied as a unique protein, and experiments were conducted on each of the Mur amide ligases separately.

Docking studies were then executed in the next phase of the research. Each Mur amide ligase structure was subjected to docking with 623 compounds from the SANCDB database in Autodock Vina. Prior to docking however, a missing loop region within MurF was modelled so as to obtain a full structure for experiments. We were able to successfully model the missing loop region within MurF before conducting docking studies. Each Mur amide ligase was subjected to docking with

ligands from the SANCDB database, and at least two ligands were obtained for each enzyme, after having undergone a form of virtual screening. These ligands were SANC00574 and SANC00575 for MurC, SANC00290 and SANC00438 for MurD, SANC00290 and SANC00525 for MurE and SANC00290 and SANC00434 for MurF. The two best ligands identified for each enzyme had docked in the active site of their respective proteins, passed Lipinski's rule of five and had substantially low binding energies. It was interesting to note however, that some compounds from the SANCDB database were inhibiting all the Mur amide ligases. About eight ligands were identified: SANC00178, SANC00261, SANC00421, SANC00448, SANC00449, SANC00547, SANC00592 and SANC00595. Although they did not pass the Ro5 and could not be studied further in this research, future research into them would be advantageous for drug discovery of antibacterial drugs, as these inhibitors can potentially inhibit a whole family of enzymes. One would need to maybe remove toxic regions of the compounds, reducing the size and therefore transforming them to druggable compounds.

The ligands identified in docking experiments were finally subjected to MD simulations in complex with their respective proteins. Analyzing the apo proteins after MD simulations revealed that the proteins exist in an open conformation. Only until a functional ligand/inhibitor docked to the protein, did we observe a closed conformation in most of the complex MD simulations run. The closed conformations were observed to occur in the MurC, MurD and MurE complexes. MurF complexes however, did not obtain the same result. The protein slightly remained in an open conformation and weak interactions between the ligand and protein would suggest a reason for the behavior of the complexes. The MurF ligands were not tightly bound to the protein and they were only interacting with one catalytically important residue within the MurF active site. The quantification of the open and closed conformations would further confirm the validity of the MD simulation trajectories observed and these will be calculated in future work. With these results, MurC - MurE ligands identified would be classified as good inhibitors and the MurF ligands would be discarded. The results obtained here do not however guarantee that the compounds will be druggable, as the Ro5 does not go into detail about specific structural features related to chemistry that are found in the compounds. This leaves room for further analysis of the five compounds identified within this research. Moreover, all compounds that did not pass the Ro5 can still be

investigated, as well as many other compounds with higher binding energies (i.e. >-8) that were not analyzed in this work.

As documented within this thesis, the research presented here has yielded some results that could be taken a step further. *In vivo* and *in vitro* experiments could potentially be carried out, utilizing the Mur amide ligase enzymes and the hypothesized inhibitors identified through docking and MD simulations. Such experiments would further validate the *in silico* results obtained here, and could potentially add towards the search for novel compounds for antibiotic drug discovery.

Thank you for reading

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of molecular biology*, *215*(3), pp.403-410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), pp.3389-3402.

Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S., 2015. The MEME suite. *Nucleic acids research*, *43*(W1), pp.W39-W49.

Barreteau, H., Kovač, A., Boniface, A., Sova, M., Gobec, S. and Blanot, D., 2008. Cytoplasmic steps of peptidoglycan biosynthesis. *FEMS microbiology reviews*, *32*(2), pp.168-207.

Barry 3rd, C.E., Boshoff, H.I., Dartois, V., Dick, T., Ehrt, S., Flynn, J., Schnappinger, D., Wilkinson, R.J. and Young, D., 2009. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nature Reviews Microbiology*, *7*(12), p.845.

Bartlett, J.G., Gilbert, D.N. and Spellberg, B., 2013. Seven ways to preserve the miracle of antibiotics. *Clinical Infectious Diseases*, *56*(10), pp.1445-1450.

Bishop, Ö.T. and Kroon, M., 2011. Study of protein complexes via homology modeling, applied to cysteine proteases and their protein inhibitors. *Journal of molecular modeling*, *17*(12), pp.3163-3172.

Bloom, D.E. and Cadarette, D., 2019. Infectious Disease Threats in the 21st Century: Strengthening the Global Response. *Frontiers in immunology*, *10*, p.549.

Chaisson, R.E. and Martinson, N.A., 2008. Tuberculosis in Africa—combating an HIV-driven crisis. *New England Journal of Medicine*, *358*(11), pp.1089-1092.

Chowdhury, R.R., Vallania, F., Yang, Q., Angel, C.J.L., Darboe, F., Penn-Nicholson, A., Rozot, V., Nemes, E., Malherbe, S.T., Ronacher, K. and Walzl, G., 2018. A multi-cohort study of the immune factors associated with *M. tuberculosis* infection outcomes. *Nature*, *560*(7720), p.644.

Coates, A.R.M. and Hu, Y., 2007. Novel approaches to developing new antibiotics for bacterial infections. *British journal of pharmacology*, *152*(8), pp.1147-1154.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry Iii, C.E. and Tekaia, F., 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, *393*(6685), p.537.

Corpet, F., 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, *16*(22), pp.10881-10890.

El Zoeiby, A., Sanschagrin, F. and Levesque, R.C., 2003. Structure and function of the Mur enzymes: development of novel inhibitors. *Molecular microbiology*, *47*(1), pp.1-12.

Emanuele Jr, J.J., Jin, H., Jacobson, B.L., Chang, C.Y., Einspahr, H.M. and Villafranca, J.J., 1996. Kinetic and crystallographic studies of Escherichia coli UDP-N-acetylMuramate: L-alanine ligase. *Protein science*, *5*(12), pp.2566-2574.

Engel, N.C., 2012. The making of a public health problem: multi-drug resistant tuberculosis in India. *Health policy and planning*, *28*(4), pp.375-385.

Fair, R.J. and Tor, Y., 2014. Antibiotics and bacterial resistance in the 21st century. *Perspectives in medicinal chemistry*, *6*, pp.PMC-S14459.

Farmer, J., Kanwal, F., Nikulsin, N., Tsilimigras, M. and Jacobs, D., 2017. Statistical measures to quantify similarity between molecular dynamics simulation trajectories. *Entropy*, *19*(12), p.646.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, *17*(6), pp.368-376.

Fiser, A. and Šali, A., 2003. Modeller: generation and refinement of homology-based protein structure models. In *Methods in enzymology* (Vol. 374, pp. 461-491). Academic Press.

Fiuza, M., Canova, M.J., Patin, D., Letek, M., Zanella-Cléon, I., Becchi, M., Mateos, L.M., Mengin-Lecreulx, D., Molle, V. and Gil, J.A., 2008. The MurC ligase essential for peptidoglycan biosynthesis is regulated by the serine/threonine protein kinase PknA in Corynebacterium glutamicum. *Journal of Biological Chemistry*, *283*(52), pp.36553-36563.

Flynn, J.L. and Chan, J., 2001. Immunology of tuberculosis. *Annual review of immunology*, *19*(1), pp.93-129.

Ginsberg, A.M., 2008, October. Emerging drugs for active tuberculosis. In *Seminars in respiratory and critical care medicine* (Vol. 29, No. 05, pp. 552-559). © Thieme Medical Publishers.

Golkar, Z., Bagasra, O. and Pace, D.G., 2014. Bacteriophage therapy: a potential solution for the antibiotic resistance crisis. *The Journal of Infection in Developing Countries*, *8*(02), pp.129-136.

Hatherley, R., Brown, D.K., Glenister, M. and Bishop, Ö.T., 2016. PRIMO: An interactive homology modeling pipeline. *PloS one*, *11*(11), p.e0166698.

Hildebrand, A., Remmert, M., Biegert, A. and Söding, J., 2009. Fast and accurate automatic structure prediction with HHpred. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), pp.128-132.

Hinman, A.R., 1998. Global progress in infectious disease control. *Vaccine*, *16*(11-12), pp.1116-1121.

Ijaz, K., Dillaha, J.A., Yang, Z., Cave, M.D. and Bates, J.H., 2002. Unrecognized tuberculosis in a nursing home causing death with spread of tuberculosis to the community. *Journal of The American geriatrics society*, *50*(7), pp.1213-1218.

Jilani, T. and Siddiqui, A. (2019). *Active Tuberculosis*. [online] Ncbi.nlm.nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK513246/ [Accessed 18 Jul. 2019].

Jnawali, H.N. and Ryoo, S., 2013. First–and second–line drugs and drug resistance. In *Tuberculosis-Current Issues in Diagnosis and Management*. IntechOpen.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T.L., 2008. NCBI BLAST: a better web interface. *Nucleic acids research*, *36*(suppl_2), pp.W5-W9.

Katoh, K. and Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), pp.772-780.

Kouidmi, I., Levesque, R.C. and Paradis-Bleau, C., 2014. The biology of Mur ligases as an antibacterial target. *Molecular microbiology*, *94*(2), pp.242-253.

Kumar, S., TaMura, K. and Nei, M., 1994. MEGA: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics*, *10*(2), pp.189-191.

Kuzmanic, A. and Zagrovic, B., 2010. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical journal*, *98*(5), pp.861-871.

Liao, B., Shan, X., Zhu, W. and Li, R., 2006. Phylogenetic tree construction based on 2D graphical representation. *Chemical physics letters*, *422*(1-3), pp.282-288.

Lindahl, E., Hess, B. and Van Der Spoel, D., 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, *7*(8), pp.306-317.

Lipinski, C.A., 2004. Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, *1*(4), pp.337-341.

Liu, Y. and Breukink, E., 2016. The membrane steps of bacterial cell wall synthesis as antibiotic targets. *Antibiotics*, *5*(3), p.28.

Lobanov, M.Y., Bogatyreva, N.S. and Galzitskaya, O.V., 2008. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, *42*(4), pp.623-628.

Mengin-Lecreulx, D., Falla, T., Blanot, D., van Heijenoort, J., Adams, D.J. and Chopra, I., 1999. Expression of the Staphylococcus aureusUDP-N-AcetylMuramoyl-l-Alanyl-d-Glutamate: l-Lysine Ligase in Escherichia coli and Effects on Peptidoglycan Biosynthesis and Cell Growth. *Journal of bacteriology*, *181*(19), pp.5909-5914.

Michael, C.A., Dominey-Howes, D. and Labbate, M., 2014. The antimicrobial resistance crisis: causes, consequences, and management. *Frontiers in public health*, *2*, p.145.

Munro, S.A., Lewin, S.A., Smith, H.J., Engel, M.E., Fretheim, A. and Volmink, J., 2007. Patient adherence to tuberculosis treatment: a systematic review of qualitative research. *PLoS medicine*, *4*(7), p.e238.

Nikaido, H., 2009. Multidrug resistance in bacteria. *Annual review of biochemistry*, *78*, pp.119-146.

Notredame, C., Higgins, D.G. and Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, *302*(1), pp.205-217.

Patin, D., Boniface, A., Kovač, A., Hervé, M., Dementin, S., Barreteau, H., Mengin-Lecreulx, D. and Blanot, D., 2010. Purification and biochemical characterization of Mur ligases from Staphylococcus aureus. *Biochimie*, *92*(12), pp.1793-1800.

Pei, J., Kim, B.H. and Grishin, N.V., 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, *36*(7), pp.2295-2300.

Peng, J. and Xu, J., 2011. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), pp.161-171.

Pruitt, K.D., Tatusova, T., Brown, G.R. and Maglott, D.R., 2011. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research*, *40*(D1), pp.D130-D135.

Rentzsch, R. and Renard, B.Y., 2015. Docking small peptides remains a great challenge: an assessment using AutoDock Vina. *Briefings in bioinformatics*, *16*(6), pp.1045-1056.)

Santos, G.B., Ganesan, A. and Emery, F.S., 2016. Oral Administration of Peptide-Based Drugs: Beyond Lipinski's Rule. *ChemMedChem*, *11*(20), pp.2245-2251.

Schleifer, K.H. and Kandler, O., 1972. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriological reviews*, *36*(4), p.407.

Seeliger, D. and de Groot, B.L., 2010. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal of computer-aided molecular design*, *24*(5), pp.417-422.)

Stroganov, O.V., Novikov, F.N., Stroylov, V.S., Kulkov, V. and Chilov, G.G., 2008. Lead finder: an approach to improve accuracy of protein− ligand docking, binding energy estimation, and virtual screening. *Journal of Chemical Information and Modeling*, *48*(12), pp.2371-2385.

Toma i , T., Šink, R., Zidar, N., Fic, A., Contreras-Martel, C., Dessen, A., Patin, D., Blanot, D., M ller-Premru, M., Gobec, S. and Zega, A., 2012. Dual inhibitor of MurD and MurE ligases from Escherichia coli and Staphylococcus aureus. *ACS medicinal chemistry letters*, *3*(8), pp.626-630.

Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, *31*(2), pp.455-461.

UniProt Consortium, 2014. UniProt: a hub for protein information. *Nucleic acids research*, *43*(D1), pp.D204-D212.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E. and Berendsen, H.J., 2005. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, *26*(16), pp.1701-1718.

Venselaar, H., Joosten, R.P., Vroling, B., Baakman, C.A., Hekkelman, M.L., Krieger, E. and Vriend, G., 2010. Homology modelling and spectroscopy, a never-ending love story. *European Biophysics Journal*, *39*(4), pp.551-563.

Ventola, C.L., 2015. The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and therapeutics*, *40*(4), p.277.

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

The World Health Organization. (2017). Global tuberculosis report 2017. The World Health Organization. https://apps.who.int/iris/handle/10665/259366. License: CC BY-NC-SA 3.0 IGO.

Appendix A

A-1: FASTA sequences used in the research.

>C_Mycobacterium tuberculosis
MSTEQLPPDLRRVHMVGIGGAGMSGIARILLDRGGLVSGSDAKESRGVHALRARGALIR
IGHDASSLDLLPGGATAVVTTHAAIPKTNPELVEARRRGIPVVLRPAVLAKLMAGRTTL
MVTGTHGKTTTTSMLIVALQHCGLDPSFAVGGELGEAGTNAHHGSGDCFVAEADESDG
SLLQYTPHVAVITNIESDHLDFYGSVEAYVAVFDSFVERIVPGGALVVCTDDPGGAALA
QRATELGIRVLRYGSVPGETMAATLVSWQQQGVGAVAHIRLASELATAQGPRVMRLSV
PGRHMALNALGALLAAVQIGAPADEVLDGLAGFEGVRRRFELVGTCGVGKASVRVFD
DYAHHPTEISATLAAARMVLEQGDGGRCMVVFQPHLYSRTKAFAAEFGRALNAADEV
FVLDVYGAREQPLAGVSGASVAEHVTVPMRYVPDFSAVAQQVAAAASPGDVIVTMGA
GDVTLLGPEILTALRVRANRSAPGRPGVLG

>C_Escherichia coli
MNTQQLAKLRSIVPEMRRVRHIHFVGIGGAGMGGIAEVLANEGYQISGSDLAPNPVTQQ
LMNLGATIYFNHRPENVRDACVVVVSSAISADNPEIVAAHEARIPVIRRAEMLAELMRF
RHGIAIAGTHGKTTTTAMVSSIYAEAGLDPTFVNGGLVKAAGVHARLGHGRYLIAEADE
SDASFLHLQPMVAIVTNIEADHMDTYQGDFENLKQTFINFLHNLPFYGRAVMCVDDPVI
RELLPRVGRQTTTYGFSEDADVRVEDYQQIGPQGHFTLLRQDKEPMRVTLNAPGRHNA
LNAAAAVAVATEEGIDDEAILRALESFQGTGRRFDFLGEFPLEPVNGKSGTAMLVDDYG
HHPTEVDATIKAARAGWPDKNLVMLFQPHRFTRTRDLYDDFANVLTQVDTLLMLEVYP
AGEAPIPGADSRSLCRTIRGRGKIDPILVPDPAQVAEMLAPVLTGNDLILVQGAGNIGKIA
RSLAEIKLKPQTPEEEQHD

>C_Williamsia faeni
MNEPIAGAMKQRLPHQLSRVHMVGIGGAGMSGLARILLARGGQVSGSDAKESRGVVD
LRTRGAQIHIGHDSAALDLLDGGPTVVVTTLAAIPKTNPELVAARERGIPILLRPAVLATL
MEGHRNLLIAGTHGKTSTTSMAVVALQHCGLNPSFAVGGALNESGTNAHHGSGELFVA
EADESDGSLLEYTPNVAVVTNVESDHLDFFGTEAAYVQVFDDFADRIVDGGSLVVCLD

DPGSAAFAGRVSDRLSERGISVRGYGEGLHRDLAPGVEPAAELLSWQSRAGGGLARVR
FTDHAGVVTEREVKLPVPGKHMALNATAAVLAGVAAGAEISDLIEGVEGFAGVHRRFQ
LRGTVGGIRVFDDYAHHPTEVRAVLTAARDMVTADGGRVVAIFQPHLYSRTETFATEF
AGSLDLADQVIVLDVYGAREEPIPGISGVTIAEQITKPVAFQPDVSALPRQVIELARPQDVI
LTIGAGDVTMQAGEIVEALRDIGRVNQGDRAVRGEK

>C_Rhodococcus kroppenstedtii
MNQENSLVSTPSGPPQGLPEELSRVHMVGIGGAGMSGIARILLSRGGAVSGSDAKESRG
VKALRARGAQVRIGHDASALDMLDGGPTAVVTTHAAIPKTNPELVAAGKRGIPVLLRP
AVLARLMTGFRTVLVSGTHGKTSTTSMLVVALQHCGADPSFAVGGELNEAGTNSHHGT
GDVFVAEADESDGSLLQYDPDLVVVTNVEADHLDYFGSEDAYTQVFDDFVARLRPGGS
LVVCTDDPGSLALARRVASQAPDGVRLLGYGADDVSVDGVETGARLLGWKPEDTGGV
VHFRLAGEDRDRTVRLGVPGRHMALNAMGAVVAAIELGADPDTVVEGLAGFAGVHRR
FQLRGRERGVRVFDDYAHHPTEVRAVLGAAADLVRTADPDRPGRVLVVFQPHLYSRTV
TFAEEFASALDLADEVMVLDVYGAREEPVPGVSGATVARSVTKPVHYQPDLSAVPRQV
AHLVGAGDVVVTMGAGDVTMLGSQILDALRSTLPRR

>C_Nocardia seriolae
MERPDSDAPTATEPVAGEHTDLPEHLRRVHMIGIGGAGMSGIARILLSRGGMVSGSDAK
ESRTVLALRARGAQVRIGHDGSALDLLPGGPTAVVTTYAAIPKTNPELVEAHRRGLLVL
LRPAVLASLMQGHKTLLVSGTHGKTSTTSMLVVALQHCGFDPSFAVGGELNEAGTNAH
HGTGGIFVAEADESDGSLLQYAPDVAVVTNIESDHLDYFGTDEAYVQVFDDFVDRLEPG
GRLVVCLDDPGSLGLARRSVAKLKDNPDGIQVIGYGSGGALDDSGAETGVPVGARLLS
WEPRDVGGVAVFQLSDEPAPRTLRLSVPGRHMALNALGALLAARATGADMGEIIQGLE
GFGGVHRRFQLTGRENGVRVFDDYAHHPTEVRAVLSAAAELVAQEARDGARSRPGRVI
VVFQPHLYSRTATFATEFGAALSLADEVVVLDVYGAREEPLPGVNGALVANAVTKPVH
YQPDMSRVGKQVGALARPGDVVITMGAGDVTMLGGQILDGLRARPGHGR

>C_Actinoalloteichus hoggarensis

MAGMTEAERIAPERQDRGDTGGTMELRNTDGALPPVLRRVHLVGIGGAGMSGIARILL
ARGAEVSGSDAKDSRTVLALRAQGAQVALGHEAAHLDQLDEPPTAVVVSTAIRATNPE
LAAARERGITVLRRAEALAELTTGSRVACVAGTHGKTSTTSMLTVALQRCRLDPSFAIG
GDLNESGSNAHHGTGDVFVAEADESDGSFLFFTPDVAVVTNVEPDHLDHHGTAEAYTE
VFESFVSRVRPDGVMIIGVDDPGSAALAGYAEAQGIRVRRYGRDVEAPGDARIVGYAPA
ELLGHVTVDLLGEPGGAPERLEIKVAVPGEHMAGNAVAALLAGLELGANRDELLAGLA
GFRGVRRRFELKGEAGGVRVYDDYAHHPTEVEAQLRAARPVAGEHRLIVVFQPHLYSR
TKTFAAEFGTALALADQVVVLDVYGAREEPEPGVTGSLVVSQVPLPAEQVRFEPSFEQL
PEVVARLADPGDVVITMGAGDVTMLGPEILGRLASVRER

>C_Mycolicibacterium peregrinum

MVGIGGAGMSGVARILLDRGGLVSGSDAKESRGVVALRARGAEVRIGHDASSLDLLPG
GPTAVVTTHAAIPKTNPELVEARRRGIPVILRPVVLAKLMAGYTTLMVTGTHGKTTTTS
MLIVALQHSGFDPSFAVGGELGEAGTNAHHGSGKCFVAEADESDGSLLEYTPNVAVVT
NIEADHLDFFGSEQAYTAVFSAFVDRIAPGGALVVCTDDPGAAALAEHTDLLGIRVLRY
GSTPAGDLAGTLLSWEQQGTGAVAHIQLAGEPHPRAIRLAVPGRHMALNALAALLAAIE
VGAPAEAVLDGLAGFEGVRRRFELVGSLGGVRVFDDYAHHPTEVRATLEAARTVVDQT
GGRVIVAFQPHLYSRTATFATQFGSALSAADEVFVLDVYGAREQPLPGISGATVAEHVS
APVTYVPDFSAVAAAVAAAAQAGDVVLTMGAGDVTMLGREIVTELGIKANRTAPGRS
STESP

>C_Gordonia terrae

MVGIGGAGMSGLARILLARGGQVSGSDAKNSRGILELRTRGARVQVGHDPSALDQIPG
GPSVVVTTHAAIPKTNPELVAARSRGIPVLLRPRVLAQLMAGYRSLLIAGTHGKTSTTSM
AVVALQHAGTDPSFAIGGELNESGTNAHHGGDPIFVAEADESDGSLLEYTPDVVVVTNI
DADHLDFFGSIEAYVEVFDRFAARITTGGSLVVCLDDPGSAALAGRVAEPLTARGVEVL
GYGRGTHAALAPTVPNVATMLSWEPRGVGGAATVRFTAPLCPDVEDRQLILPLPGEHM
ALNAIAAVIGAARTGTPAGAVATDISDRFDGIVAGIGSFGGVHRRFEFRGQRGGVEVIDD
YAHHPTEVRAVLSAAQDMVSARGGPGAERGRVIAVFQPHLYSRTVEFADDFAAALDLA

DTVVVADVYGAREAPIPGVSGRTISDKLTVPGIFAPDLSRLAAQVARLARSGDVVLTLG
AGDITMQGPEILAALASASTTGDAANGADPAGSPADAPGESGDRTGRVS

>D_Mycobacterium tuberculosis
MSGLPRSVPDVLDPLGPGAPVLVAGGRVTGQAVAAVLTRFGATPTVCDDDPVMLRPH
AERGLPTVSSSDAVQQITGYALVVASPGFSPATPLLAAAAAAGVPIWGDVELAWRLDA
AGCYGPPRSWLVVTGTNGKTTTTSMLHAMLIAGGRRAVLCGNIGSAVLDVLDEPAELL
AVELSSFQLHWAPSLRPEAGAVLNIAEDHLDWHATMAEYTAAKARVLTGGVAVAGLD
DSRAAALLDGSPAQVRVGFRLGEPAARELGVRDAHLVDRAFSDDLTLLPVASIPVPGPV
GVLDALAAAALARSVGVPAGAIADAVTSFRVGRHRAEVVAVADGITYVDDSKATNPH
AARASVLAYPRVVWIAGGLLKGASLHAEVAAMASRLVGAVLIGRDRAAVAEALSRHA
PDVPVVQVVAGEDTGMPATVEVPVACVLDVAKDDKAGETVGAAVMTAAVAAARRM
AQPGDTVLLAPAGASFDQFTGYADRGEAFATAVRAVIR

>D_Escherichia coli
ADYQGKNVVIIGLGLTGLSCVDFFLARGVTPRVMDTRMTPPGLDKLPEAVERHTGSLN
DEWLMAADLIVASPGIALAHPSLSAAADAGIEIVGDIELFCREAQAPIVAITGSNGKSTVT
TLVGEMAKAAGVNVGVGGNIGLPALMLLDDECELYVLELSSFQLETTSSLQAVAATILN
VTEDHMDRYPFGLQQYRAAKLRIYENAKVCVVNADDALTMPIRGADERCVSFGVNMG
DYHLNHQQGETWLRVKGEKVLNVKEMKLSGQHNYTNALAALALADAAGLPRASSLK
ALTTFTGLPHRFEVVLEHNGVRWINDSKATNVGSTEAALNGLHVDGTLHLLLGGDGKS
ADFSPLARYLNGDNVRLYCFGRDGAQLAALRPEVAEQTETMEQAMRLLAPRVQPGDM
VLLSPACASLDQFKNFEQRGNEFARLAKELG

>D_Williamsia faeni
MSASPERPWGERVLVAGAGVAGASAARFLLNQGVSVTVCDNRAAALEPLVSAGAAAV
SPDTLMATPGWIADFDLVVASPGFRPDNPLVVAAADAGLPVWGEVELAWQVDRSGVT
GAPRQWLVVTGTNGKTTTTGMIESIMAAAGVSVLACGNIGLPVLDALVAEPRIDVLAVE
LSSFQLFWAPSVRPTAGVVLNIAEDHLDWHGSMAAYVEAKAIALSGPIAVVGLDDPVA
AGLRGSDPDGRTIGFRLGAPGPGELGVVDGRLVDRVDWEASEDLSSVVPIDPVSLFDAD
RVRPPGPSGVLDALAAAALCRAVGVSAPAVAEGLDTFTVGGHRSAPVAIVGTTTFINDS

KATNPHAAQASILGHRRVVWVAGGLLKGASVDELVIAIRDRLAGVVIIGADRDIIKDAM
TRHAPLVPTVTRVTGEDGAVNIAPASAVPASADAVMDSAVRAAWDLCTADPDTPDAV
VLAPAAASLDMFSGYGARGDSFAASARALPGAQVLGAS

>D_Rhodococcus kroppenstedtii
MAGATDHRAGENSGAVTDPRDLRDRRVLVAGGRVSGRALLPTLGTLGARVTVADSDE
SALAACADLGADTVLMDRIAPDRAAVGEFALVITSPGFRPDAPLLTTALTVGVPVWGDI
EFSWRVDRAEVYGPARTWLVVTGTNGKTTTTSMLASILDASGRPSLACGNIGVPVLDAL
RSESPRAEVLAVELSSFQLFWAPSVRPAAGVVLNIAADHLDWHGGMDAYVEAKLQAL
RGDVAVVGLDDAVASSLRSRATAPRTVGFTAARPAAGDLGVEDGMLVDRAFGDGVPL
VSAAEIEPPGPAGLQDALAASALARAAGVDADAVARGLSAFRVGPHRAATVRELGGVQ
FVDDSKATNPHAARSSILAHERVVWLAGGLLKGAEVDDLVTEVRSRLVAAVVFGRDG
DAIAAALGRHAPDVPVVHVPPGDDASVADETDVTNGDQVMTLAVRAAAHRAEPGTTV
LLAPSTASFDQFRDYGHRGDSFTAAVADLTEADLGTGARPGTRG

>D_Gordonia terrae
MSSEPSGPDPRELAGRIVVVGGAGTAGLSATRFLVTLGADVLLADDRFAGRDASPADRE
TAAGAVPGVAELASRGVEPIAVADLLADPARMATAAVVIVSPGFAPTHHLVRAATDSG
VPVWGEVELAWRVDAAGLLGEPRTWLVVTGTNGKTTTTSMLAGIVEAAGRAGAACG
NIGLPVLDAMQADPRVDVLCAELSSFQLHWAPSVRPDAGVVLNVADDHLDWHGTFDA
YALAKAGALRGAVGVVGLDDPVASGLPAAGRRIGFTLSEPQPGQLGVENGRLTDRAFD
AGRLFDADAVRPAGPSGVADALAASALALAAGIGPDAIRAGLAAFRPAGHRGEVVASR
DGVAFVDDSKATNPHAAQAAVAAFERVVLIAGGLLKGASMDEMLTAVRARLVGVVAI
GRDRDLVIEAIARHAPEVPTVTVFTGDDDTVNVHRPGQAGPQASASDRPGSGSSASVAV
MDRAVEEAWAFAVASSPAPDAVLLAPAAASLDMFAGYGARGDAFAAAARRIAGSPRA
GR

>D_Actinoalloteichus hoggarensis
MDGSPFAGLSVLVAGAGVSGVSAARALLAQGATVSVVDGSPARIESMNLPAVTGLVGV
DVVPPGIDLVVTSPGWRPTAPLLVAAADAGVEVIGEVELAWRLDQARPDGPADWLVVT
GTNGKTTTVSMVESILLAAGLDAVACGNVGLPVVDAVLSGHRVLAVELSSFQLHWSSS

VRPAAGLVLNVAEDHLDWHGSAAAYGAAKARALLGPVAVAGVDDEPAAALLAASPA
DRRIGVTAAEPAEGQFGVSAGVLVDRVFGPDGDAVSLIDAELIRPVGPTGIVDALAAAA
LTRSYGVAPESVEAGLRDFRPGAHRAVVVAEHDGIVYVDDSKATNPHAAAASLRGHAR
VVWLAGGLLKGASVDDLVAETGDRLAAAVLFGADRALIAAALARHAPDVPVYEVATG
DDKAMREAVDIASQVARPGETVLLAPAAASMDMFTDYAHRGDSFAAAVLDRQAAEG
VAEESAEGDGGAPERDTRPVARPAGTERADSDRIGAERLGGDDTGAGRPGSGADTGSD
DVVERGGSV

>D_Nocardia seriolae
MSEHSPRVLRSPGPMLEFLRGRDVLVAGWGVSGRSLIEPLQDIGARPVVTDGGGKALAE
AAELGLGTATTQDLLEPDALQRFALVITSPGWRPDSPVLVSAVTEGIPVWGDVEFAWW
VDQARIYGPVRKWLVVTGTNGKTTTTQMTHSILRAAGIPSVACGNIGLPILDALRRNPGP
QILAVELSSFQLHWAPSVRPEAGVVLNVAEDHLDWHGGLDAYAAAKARALTGRVGVV
GLDDAVAAALARKSKARRTVGFRIGVPADGELGVVDGKLLDRAFTKAAILAEVGDISP
QGPAGVADALAAAALTRSIDVAPQFVKEGLQEHKVGPHRAAFVRELSGVEFVDDSKAT
NPHAARSSILAHPNVIWVAGGQLKGATVEDLIEECREHLVAAVLFGQDAAVIAAALARH
APDVPVVELNSGDDARMGDPFTEIEPEAVMARAVRVAAGFAHRGDTVLLAPAAASLD
MFADYTHRGRAFAAAVGALDEKDIGRTE

>D_Mycolicibacterium peregrinum
MSLLTPGAPVLVTGAGVTGRAVLAALAPLNVAATLCDDNAEALRVLAEQGTAVIDPAG
AVASIGDYALVVTSPGFPPTAPVLAAAAGAGVPVWGDVELAWRLDDAGHYGPPRRWL
VVTGTNGKTTTTSMLQEMLLADGRRSLLCGNIGDPVLAVLDQPAELLAVELSSFQLHW
APSLRPDAGVVLNVAEDHLDWHGSMAGYAADKARALAGRVAVVGLDDPVAAGLLPA
AGAPVRVGFRMGEPAAGELGVRDGKLIDRAFGDDVELADAATIRVAGPVGVLDALAA
AALARAVGVTPASIATALASFEVGRHRAELVGEADGVRYVDDSKATNPHAAQASITAF
DRVVWIAGGLLKGASVDDLVRHVANRLVAVVLIGRDRRMVADALSRHAPDVPVVEVV
AGEDSGVLETNESIGDHVTRVIDVGDRPVSDAVMAAVVDAARGLAGAGDTVLLAPAG
ASFDQFSGYGQRGDAFAGAVRAAIG

>E_Mycobacterium tuberculosis

MSSLARGISRRRTEVATQVEAAPTGLRPNAVVGVRLAALADQVGAALAEGPAQRAVTE
DRTVTGVTLRAQDVSPGDLFAALTGSTTHGARHVGDAIARGAVAVLTDPAGVAEIAGR
AAVPVLVHPAPRGVLGGLAATVYGHPSERLTVIGITGTSGKTTTTYLVEAGLRAAGRVA
GLIGTIGIRVGGADLPSALTTPEAPTLQAMLAAMVERGVDTVVMEVSSHALALGRVDGT
RFAVGAFTNLSRDHLDFHPSMADYFEAKASLFDPDSALRARTAVVCIDDDAGRAMAAR
AADAITVSAADRPAHWRATDVAPTDAGGQQFTAIDPAGVGHHIGIRLPGRYNVANCLV
ALAILDTVGVSPEQAVPGLREIRVPGRLEQIDRGQGFLALVDYAHKPEALRSVLTTLAHP
DRRLAVVFGAGGDRDPGKRAPMGRIAAQLADLVVVTDDNPRDEDPTAIRREILAGAAE
VGGDAQVVEIADRRDAIRHAVAWARPGDVVLIAGKGHETGQRGGGRVRPFDDRVELA
AALEALERRA

>E_Mycolicibacterium peregrinum

MAMKLRPSRPAGQYLVPLAEQVQAASATGNPLPELRVTGVTLRGQDARPGDLFAALPG
SAVHGARYAPDAVAAGAVAVLTDAAGAAELDALEVPILVHPDPRSVLGEVAAEVYGR
PSERLTVIGVTGTSGKTTTTYLAEAGLRAAGRVAGLIGTVGVRIAGRDLPSALTTPEAPD
LQALLALMVESGVDTVVMEVSSHALTLGRVDGIVFALGGFTNLSRDHLDFHPTMQDYF
EAKAGLFDPGSRNHARAAVVCVDDDAGVAMARRAEKVTTVSAAGAAADWQVRDVS
AAGVGSQVFTLVDPAGAAHPMRIGLTGAYNIANAALAVALLDRAGVTPEQAAPGLRAA
TVPGRLQPIDRGQEFLALVDYAHKPGALQAVLETLRASGPGRIGVVFGAGGNRDTGKR
APMGRVAAELADLVVVTDDNPRDEDPALIRAAIMAGAAEAQTYAEVVEVADRRAAIER
AVAWAGPGDTVLIAGKGHESGQTGGGHTRPFDDRDELAAALEALKEGESGT

>E_Rhodococcus kroppenstedtii

MNSRSATTTRSDHTVPSSSLRPQRPPVTPIGRLADAIGATVHGGASAAEISVTGIELRAQD
VLPGDLFAALPGAHSHGARFAADAVARGAVAVLTDPEGADLVANSAGGSAGAHSDVV
VLVHPHPRSVLGSVSATVYGDPSASMIVVGITGTSGKTTTSYLVEAALRAGGRTPGLIGT
VETRIGDTPVPSALTTPEAPHLHAMFAAMLERGCDSVVMEVSSHALSLGRVDGTRFAV
GAFTNLSQDHLDFHHTFEDYFAAKARLFAAESDVHAAESVVCIDDEWGRRMADIAAAS
GPVTTVSTDGEARGAAAVARNWQVRADGAQTFEAVLGDRVVSATLRLPGRYNVANAL
VGAAVAARVGVDPQVAVDAMATVDVPGRMQRVDRGQAFLAVVDYAHKPAALEVAL

ATLRRQTTGRVAVVVGAGGDRDTEKRPLMGAVSAAAADLVVITDDNPRSEDPAAIRAA
VRTGAMGVDDAARGEVVEISPRADAVAYAVRWARPGDVVLVAGKGHEVGQKVGDV
VHPFDDREALARAIDELGEDGTS

>E_Gordonia terrae
MASRRRSPDRSGGPIRPRQVTPTPVSVLSTATGARLDLVGGADAETDVTGVTLRAQEVR
PGDLFAALPGSSTHGARFADQAVAAGASAILTDPAGRAELARVLPPETAVAVLIHPTPRT
VLGNVSARVYGNPSQRLKLIGITGTSGKTTTSYLIEAALLAAGHSVGLIGTVETRVNGIA
QPSSLTTPEAPTLQALLAAMLEDGVDIVVMEVSSHALALGRVDGAHFAIGAFTNLSQDH
LDFHQTMRAYFDAKAQLFVPSAPTHAVRAVVCVDDEWGRRMAELARRPGEAAHLQP
VLVSTGPTPAHWHAGESSLEADGSSRTPFAEPDGDHELRIPLPGRYNVANGLLAVAVAH
AAGVPVPTALEAVASVSVPGRLQRIDRGQPFLALVDYAHKPGAVDAVLTTLRGQSTGR
VAIVIGAGGDRDTGKRPLMGEAAARGAELVVVTDDNPRTEDPGAIRAEVLAGARAVPE
TQRPRGAEPIREIGDRAAAIEAAIGWAKPGDIVLVAGKGHEAGQEIDGVKHPFDDREVV
AHALESRVQESRAQESTPGTPDSATGDRS

>E_Actinoalloteichus hoggarensis
MKAPRPAHVTPSSVAELASVVAAGLQVEKTSLTPDARVVTGATLRAQDVRPGDLFAAL
PGARAHGADFADQAAAAGAVALLTDEEGSRRAVAAGVALPTLVHPDPRAAIGPAAAR
VYGDPSKLLRVLGITGTSGKTTTCYLIDGGLRAAGVSTGLVGTVETRVAGERLNSAFTTP
EAPDLQALFALMVERGVGHVSMEVSSHALSLGRVGGTEFAVGGFTNLSQDHLDFHRDL
EDYFAAKARLFDGRARHEVVCVDGGWGARLVTPDTITVATGADAEASWSVRDARALP
TGEQTFVALGPDGVRLDLRIRLPGPFNIANALLAAACLHADGIPADAIAAGLSAVDVPGR
MERVVEGQPFVAMVDYSHKPGAVAAVLDAARAQTEGRVLVVLGCGGERDTGKRPAM
GEEAARRADLLVVTDDNPRGEDPTSIRSAMLTGALAVPRAQRGEIVEIGDRRAAITEAV
ARARPGDILVVAGKGHETGQHVGEIVLPFSDRDVLAEAVRARETEENR

>E_Nocardia seriolae
MPAQSSPQVLRPAAPAATSLRKVEELTGARSAGSPEGSENVQVTGIEQRSNAVRAGDLF
AGLAGSHAHGARFAHDAVERGAVAILTDPAGAELIGEIAVPVLVHDAPRAILGELSSAL
YGDPSRKLRIVGITGTSGKTTTSYLVEAGLAAAGLTTALIGTIETRIGGVRVPSALTTPEA

PQLHAMFALMVEQGVDAVVMEVSSHALALGRVDGVHFTVGAFTNLSQDHLDFHSDFE
DYYAAKRRLFIPDGPDSAASARPSVAARASVICVDDEWGQRLAAEIGDDLGPAETAGA
RLRTVATRAGVRADWLADGITAGPGGTQEFIAMGPDVNVDVRLRLPGTYNIANGLLAV
AVCVAAGVEAEAAAVALGDVDVPGRMQRVDEGQEFLAVVDYAHKPAAIESVVATLRT
WLAEQGEGRLAVVVGAGGDRDAGKRPLMGAAAARGSDLAIITDDNPRSEDPAAIRAA
VLAGAEQVPDAERGEVREVGDRAAAIKAAVAWAQPGDVLLIAGKGHETGQEISGVKHP
FDDREVLAAALSEKTRDLTHS

>E_Williamsia faeni
MTLRRYVEHVSAPDSSVLRPALVTPTSVEALAATGNARIDLIGSATAQTEVGGVTLRAQ
DVRPGDLFAALPGSRFHGAQFAADALAAGASAVLTNQAGRDVIAGLDPTPNCVVLVHP
EPRAVLGDLSSQVYGHPSRSLTLIGITGTSGKTTTSYLTEAALMAAGHSVGLIGTVETRIA
GIRQPSALTTPEAPQLQGLLAAMLEQGVDSVVMEVSSHALYLGRVDGTRFALGAFTNLS
QDHLDFHPTMRDYFNTKALLFQPDSAVHAGASVICVDDDWGREMFRIAGRGAVSVSTT
GSADGVDWHAGTSEVARDGAQRVTVTDPTGQSYEMTVPMPGRFNVANALTAVALAS
GAGVPVETAIASVGDVAVPGRVERVVRGQPFLAVVDYAHKPAALEAVIATLRGQTQGR
IAVVVGAGGDRDTGKRPMMGEAAARGAEFVVVTDDNPRTEVPADIRAQVIAGAAAVP
LAERPSGAEPVQEIGDRAGAIAAAVRWARAGDVVLVAGKGHEAGQEIHGIKHPFDDRV
VLGQAIDELGTQQNPTDTESLQ

>E_Escherichia coli
MADRNLRDLLAPWVPGAPERALREMTLDSRMAASGDLFVAVLGHQADGRRYIPQAIA
QGVAAIIAEAKDEATDGEIREMHGVPVIYLSQLNERLSALAGRFYHEPSEKMRLVGVTG
TNGKTTTTQLLAQWSQLLGETSAVMGTVGNGLLGKVIPTENTTGSAVDVQHVLAGLVA
QGATVGAMEVSSHGLVQHRVAALKFAASVFTNLSRDHLDYHGDMEHYEAAKWLLYS
THHCGQAIVNADDEVGRRWLAKLPDAVAVSMEDHINPNCHGRWLKAIDVNYHDSGAT
VRFTSSWGDGEIESRLMGAFNVSNLLLALATLLALGYPLADLLKTATQLQPVCGRMEVF
SAPGKTTVVVDYAHTPDALEKALQAARLHCTGKLWCVFGCGGDRDKGKRPLMGAIAE
QFADVVVVTDDNPRTEEPRAIINDILAGMLDAGQAKVMEGRAEAVTNAIMQAKENDV
VLVAGKGHEDYQIVGSQRLDYSDRVTAARLLGVIA

>F_Mycobacterium tuberculosis

MIELTVAQIAEIVGGAVADISPQDAAHRRVTGTVEFDSRAIGPGGLFLALPGARADGHD
HAASAVAAGAAVVLAARPVGVPAIVVPPVAAPNVLAGVLEHDNDGSGAAVLAALAKL
ATAVAAQLVAGGLTIIGITGSSGKTSTKDLMAAVLAPLGEVVAPPGSFNNELGHPWTVL
RATRRTDYLILEMAARHHGNIAALAEIAPPSIGVVLNVGTAHLGEFGSREVIAQTKAELP
QAVPHSGAVVLNADDPAVAAMAKLTAARVVRVSRDNTGDVWAGPVSLDELARPRFTL
HAHDAQAEVRLGVCGDHQVTNALCAAAVALECGASVEQVAAALTAAPPVSRHRMQV
TTRGDGVTVIDDAYNANPDSMRAGLQALAWIAHQPEATRRSWAVLGEMAELGEDAIA
EHDRIGRLAVRLDVSRLVVVGTGRSISAMHHGAVLEGAWGSGEATADHGADRTAVNV
ADGDAALALLRAELRPGDVVLVKASNAAGLGAVADALVADDTCGSVRP

>F_Escherichia coli

MISVTLSQLTDILNGELQGADITLDAVTTDTRKLTPGCLFVALKGERFDAHDFADQAKA
GGAGALLVSRPLDIDLPQLIVKDTRLAFGELAAWVRQQVPARVVALTGSSGKTSVKEM
TAAILSQCGNTLYTAGNLNNDIGVPMTLLRLTPEYDYAVIELGANHQGEIAWTVSLTRP
EAGLVNNLAAAHLEGFGSLAGVAKAKGEIFSGLPENGIAIMNADNNDWLNWQSVIGSR
KVWRFSPNAANSDFTATNIHVTSHGTEFTLQTPTGSVDVLLPLPGRHNIANALAAAALS
MSVGATLDAIKAGLANLKAVPGRLFPIQLAENQLLLDDSYNANVGSMTAAVQVLAEMP
GYRVLVVGDMAELGAESEACHVQVGEAAKAAGIDRVLSVGKQSHAISTASGVGEHFA
DKTALITRLKLLIVEQQVITILVKGSRSAAMEEVVRALQENGTC

>F_Mycolicibacterium peregrinum

MIEMTVARIAEIVGGELADITAEEAAATTVTGTVEFDSRAVGPGGLFLALPGARSDGHDF
AAGAVAAGAVAVLAARPVGVPAIVVEPEPGAADAASGALEFDTDGSGAAVLAALAKL
AAAVAAELVAGGLTIIGVTGSSGKTSTKDLLVAVLDPLGQVIAPPGSFNNELGHPWTVL
RATPETDYLILEMSARHPGNIAALAAIAPPQIAVVLNVGTAHLGEFGSREAIANTKAELP
QAVPASGVVILNVDDTAVAAMAGKTAARVVRVSREPGSEVDVWAGPVALDGLARPRF
TLHAGGGQVEVALAVHGDHQVGNALCAAAVALQCGAGLDQVATALAGAGPVSRNR
MQVGNRADGVTVINDAYNANPDSMRAGLKALAWMAREGTGESAGEAGGKRRSWAV
LGEMAELGDDAITEHDAIGRFAVRLDVSRLIVVGTGRTMNAMHHGAVMEGSWGSEST
MVDDADAALALLQAELQPGDVVLVKASNSVGLGALADALVAAEDADDRNADR

>F_Williamsia faeni

MAAATGGKLVGPDQDIDGASIDSRGIRPGQLFVPIVAERDGHDFIDGALAGGATAYLTA
QAGDQGKPDATAIQVADTALALADLGRAARARLSGPVVGVTGSVGKTSTKDLLAAVL
ATTYVTAANERSFNNELGLPLTLLGAPDNTEAVVVEMGARGIGHIALLCDIARPTVGIVT
RVEGVHLELFGDIDAVAQAKGELIEALPADGLAVLNADHPVVAAMAQRSAAPVLTFGL
TSGADVYAENIVMDSELRASFTLHTPWGTAQVALEARGEHQVANALAAAAAAGGIGV
SVEQIAAGLGGAAISGLRMELSRTDAGVQILNDAYNANPTSMSAALDSFATLAAGRRV
AVLGVMAELGADAAAQHVAITARARSLGIEVIAVATPDYGPDATHVDDVAAAVAALG
DLESGDAVLVKGSRVAALERVADALSKS

>F_Rhodococcus kroppenstedtii

MRALSIGEIADIVGGTLHDVPDPDVRVTGTVEFDSRKVTAGGVFLALPGERVDGHDHAA
DAVAAGAVVVLATRPVGVPAVVVEPTPRPDLPAYDPTGAGPAVLAALAALASASVTGL
IENGLTVVGITGSSGKTSTKDLIRAVLERAGTVVAPPGSFNNELGHPWTVLRADETTRFL
VLELSARGRGHIAELARIAPPRIGVELNVGTAHLGEFGSRENIAEAKSELVQALPAAADG
GVAVLNADDRLVAAMAEKTTARVVTYGHGGSADVRATDVVLDAEARASFTLHTASGS
APVALQVHGEHQVSNALSAAAVGLECGMALDDIAAALGDARAASERRMDVRTRPDGV
TVINDSYNANPDSVRAALKALVTMARADRSSARRSWAVLGEMAELGEDSVIEHDRIGR
LAVRLAVDKLVVVGTGRPSHGMHQGAVMEGSWGDESVLVPDADAAVELLREQVAPG
DLVLVKASQSVRLWTVAEAVLAEQGENR

>F_Gordonia terrae

MGGELLGPDVQIDGASIDSRSIRPGQLFVPIVAERDGHDFIPAAIDGGAGAYLTSRTDPVG
GTAIRVADTALALADLGRAVRTRIPDRVVGITGSVGKTSTKDLLAGVLGTTYRTAASEK
SFNNELGLPITLAAAADDVEAVVLEMGARGIGHIDLLCSIAHPTVGVITRVEGVHLELFG
DIESVARAKGELVESLPADGIAVLNADHRYVAAMADRTSARVLRYGLSPDADVYASDI
VMDDELRASFRLHTPWGSTDVRLGARGEHQVPNALAAAAAGGGLGVAVEDIATGLLG
AQLSGLRMELTTSSRGVVVLNDAYNANPTSMSAALRSLAALPATRRVAVLGTMAELG
ADSPAQHHAIALEARDLGIEVIAVAEAGYGDSARHVSDVDGVAAELGTLEEGAAVLVK
GSRVAALERVAQALLD

>F_Actinoalloteichus hoggarensis

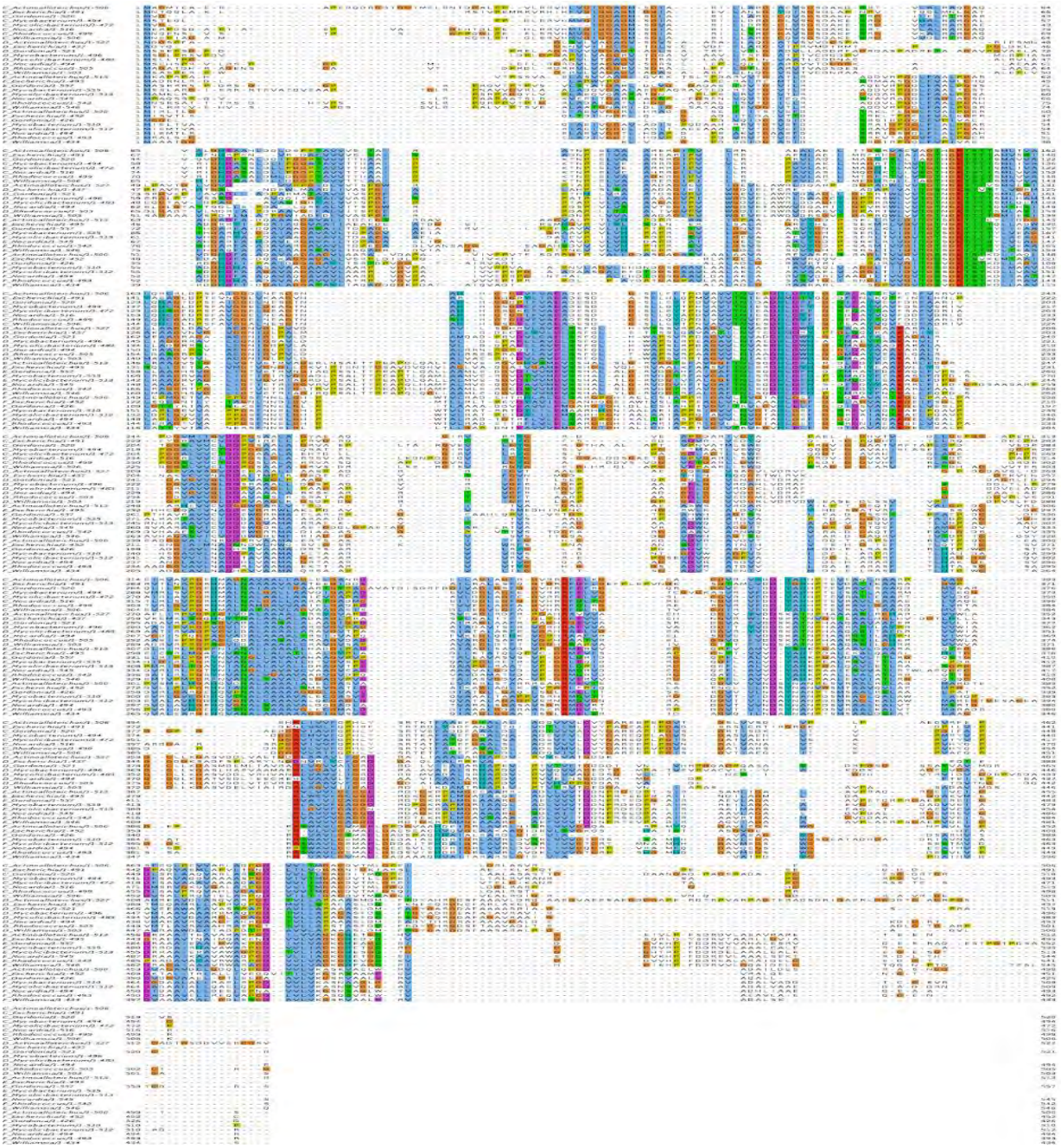
MIRLTLGEIAEVVGGRLHRADGSETITSTVEFDTRKLTEGGLFVALPGERVDGHDFAAQ
AVAAGAAGVLAARAVDAPAVIVPPVTESDRPGTPVALLGDASGAGASVLAALAAVAR
EVASRLSRAGLTVVGVTGSSGKTSTKDLIAQLLEPMGETVAPPGSFNNELGHPWTVLRA
DEHTRHLALELSARGPGHIAELCSIAPPRIGAVLNVGTAHLGEFGSQEAVARTKGELPAA
LPAEADGGVAILNRDDPRVAAMAERTSAKIIFFGHGKDADVRAVDVTLDQTARPSFRLV
TPQGEAPVSMELHGAHQVDNALAAAAAIALELGASVAEVAERLSTARRRSARRMEVTER
SDGVVVVNDSYNANPESVRAALRTLASMARDTEGEPRRSWAVLGVMGELGPDGVAA
HDEIGRLAVRLNISRLVVVGEQARPMYQGAVLEGSWGNEAVLVPDVAGAVDLLGDQL
QPGDVVLVKASKVAELWRVADAILDESRASNGGTS

>F_Nocardia seriolae

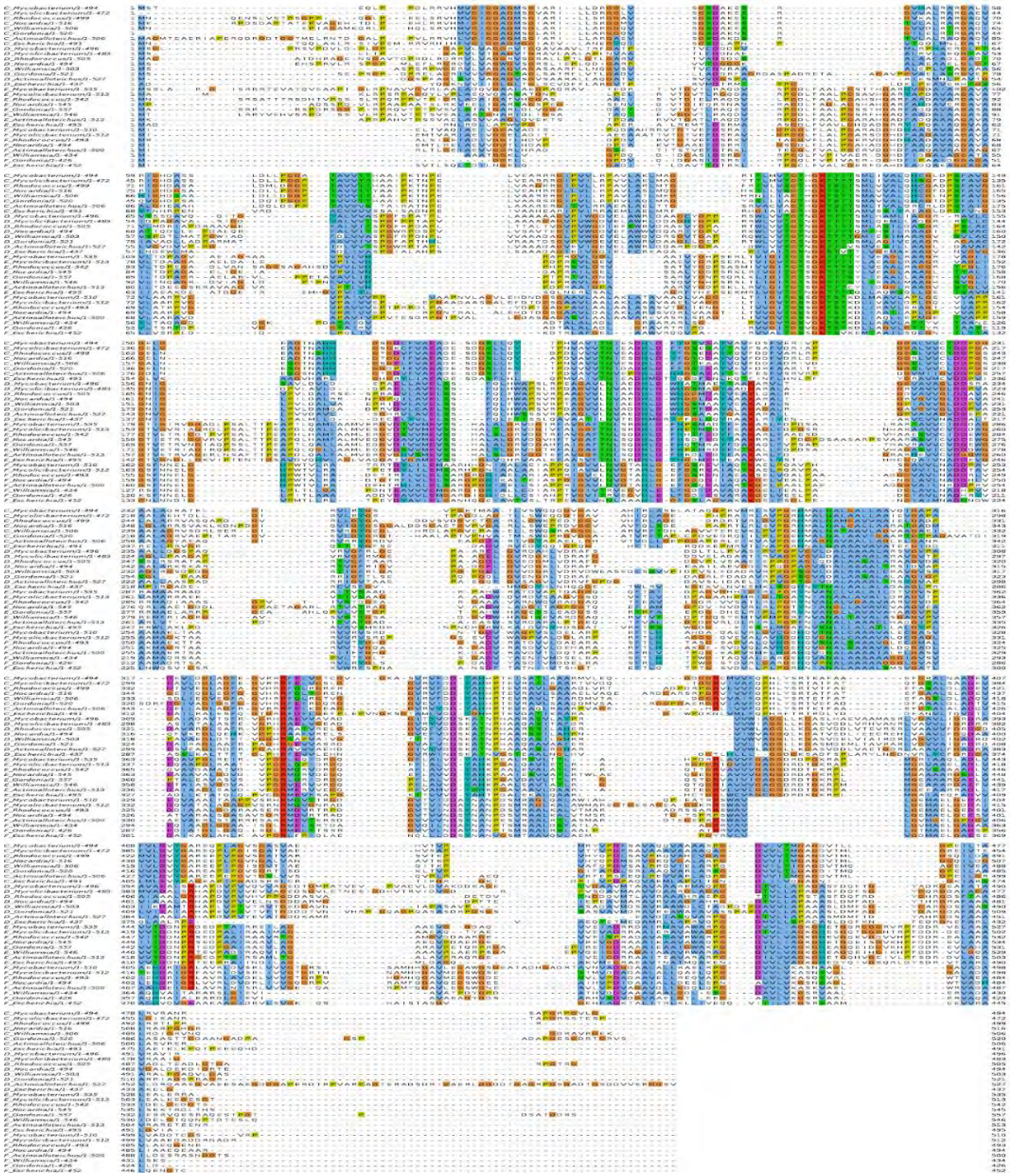
MIEMTLRELAEVVGGTLHDAPDPEAKVTGAAEFDSRRIGSGDLFLALPGAHADGHDFA
AQAVAQGAVAVLAARPVGVPAIVVEPRPGNTRALALEHDTDGSGAAVLDALAKLARS
VDRLVAAGTLTVVGVTGSSGKTSTKDLIAATLAPLGPVVAPPGSFNNELGHPWTALRAD
TGTRFLVLELSARGPGHIKALTEIASPGIGVVLNVGTAHLGEFGGQDIIAQAKGELVEALP
ANGLAVLNADDHLVAKMAGRTAARVVTVGQAAGADVRASDVVLDEHARARFTLHAK
GESVDVQLAVHGEHQVGNALSAAAVALDQGADLATVARALSGASAVSGRRMDVRTR
ADGVTVVNDSYNANPDSVRAALKALVTMSRTGTSPRPSWAVLGEMAELGEESVLEHD
RIGRLVVRLDVDRLIVVGQGRPVRAMFQGAVQEGSWGEEAIHVPDIDAAIALLDEELEP
NAVVLVKGSNSAGLWTVADHLIAAEQEAAR

A-2: PROMALS3D and MAFT MSA results

PROMALS3D:

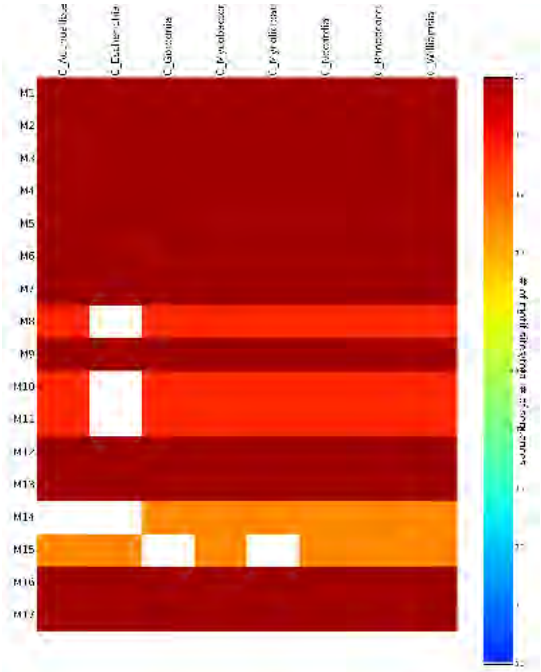


MAFT:

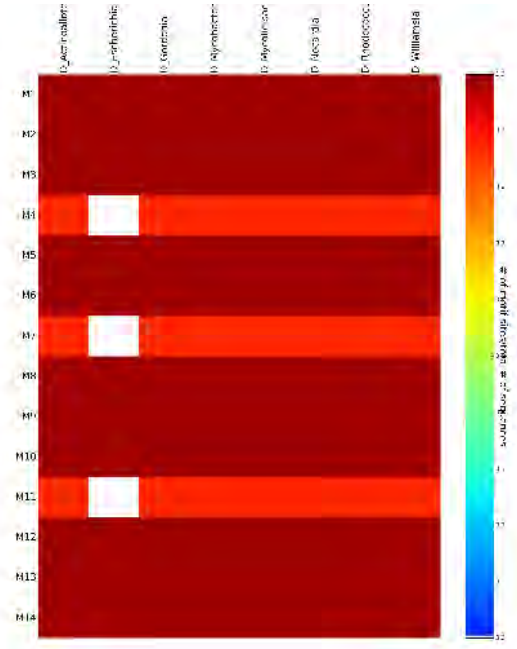


A-3: Motif heat maps created for each group of the Mur amide ligase family.

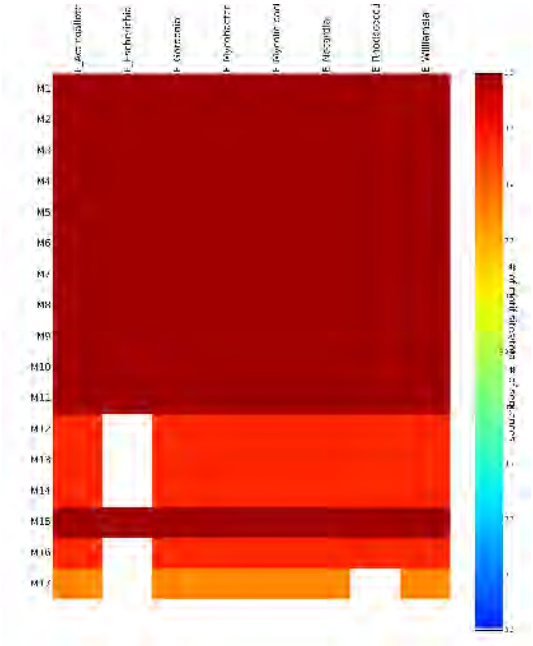
MurC



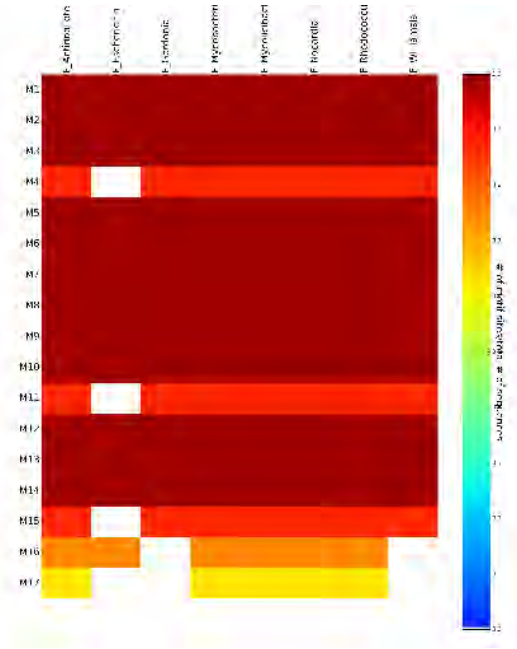
MurD



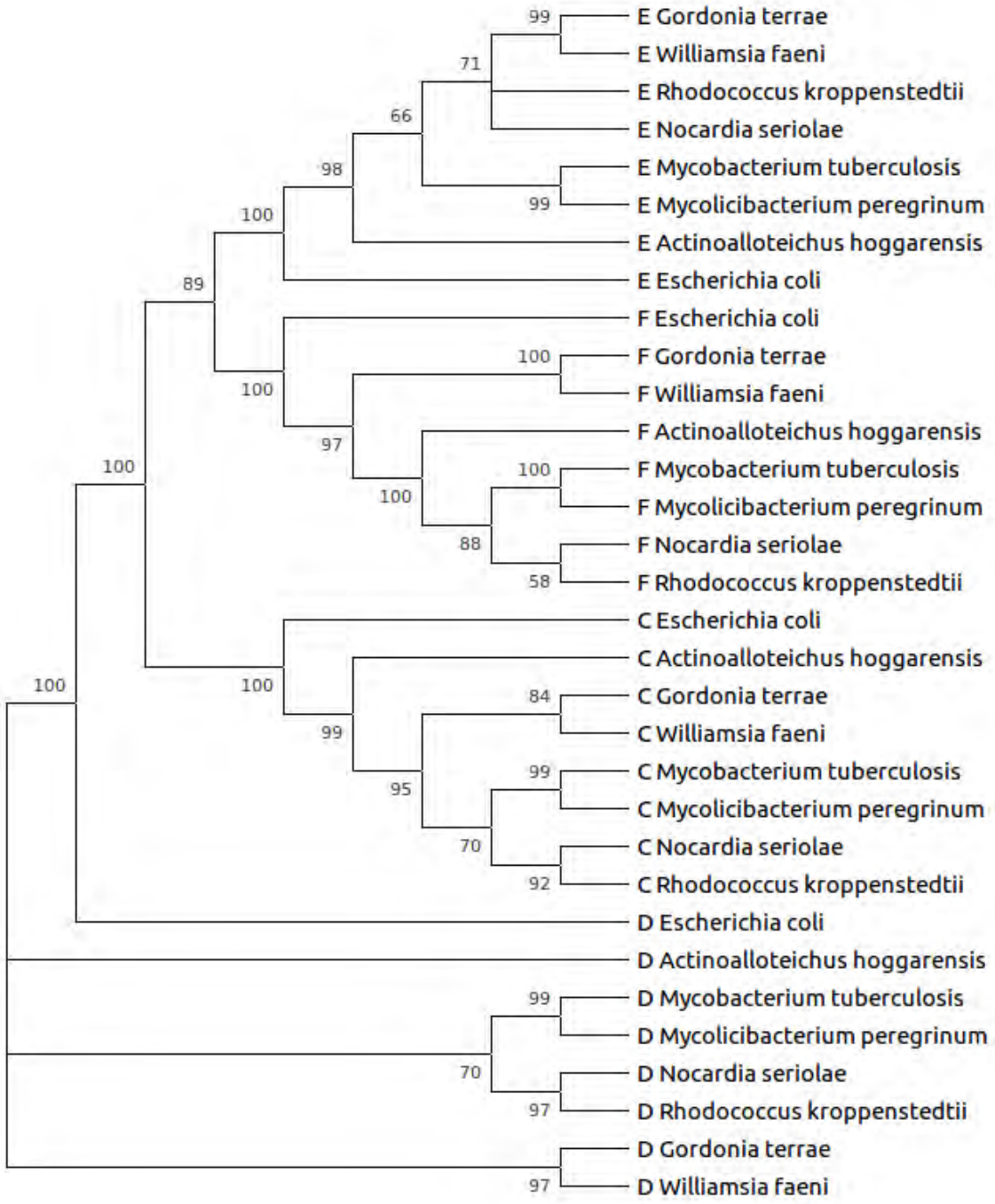
MurE



MurF

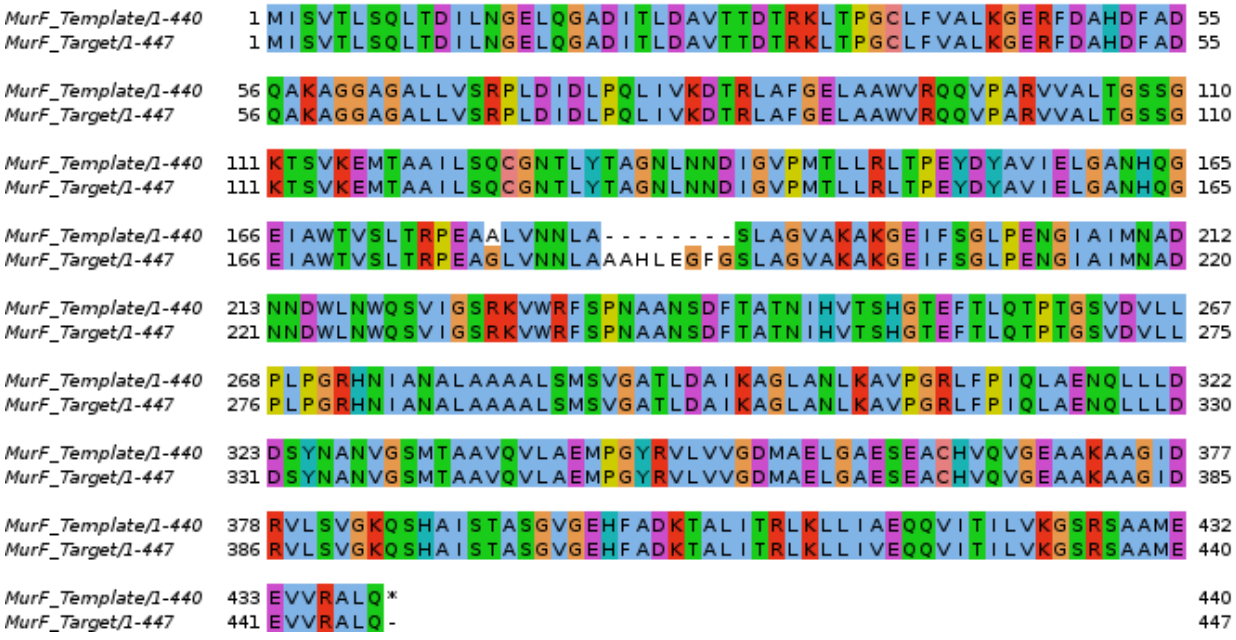


A-4: Bootstrap tree used to compare to phylogenetic tree used in this research.



Appendix B

B-1: Full length sequence alignment used when modelling loop region of MurF.



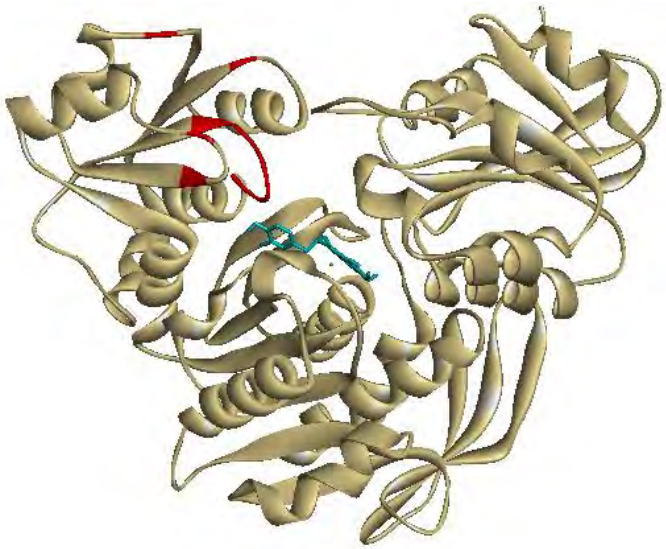
B-2: Summary of values for physicochemical and structural properties of identified ligands for comparison with Lipinski's rule of five ranges.

	SANCDB Ligand	Binding Energy (kcal/mol)	LogP < 5	MW < 500 Da	HBA <10	HBD < 5
MurC	132	-8.4	0.1	410.1	8	4
	141	-8.2	1.4	337.1	5	2
	225	-8.1	2.2	316.1	6	3
	242	-8.1	1.6	316.1	6	3
	281	-8.2	2.8	462.3	7	2
	282	-8.1	3.2	476.3	7	2
	307	-8.0	3.7	360.0	4	2
	368	-8.2	2.5	288.1	5	2
	574	-8.7	2.2	346.1	7	3
	575	-8.5	2.5	360.1	7	2
MurD	237	-9.0	4.9	392.3	5	0
	290	-9.6	3.6	401.3	5	2
	434	-9.0	3.1	374.1	6	2
	438	-9.4	3.1	374.1	6	2
	465	-9.2	1.8	474.2	8	4
	446	-9.5	2.0	472.2	8	3
	466	-9.1	2.6	448.2	7	1
MurE	290	-10.1	3.6	401.3	5	2

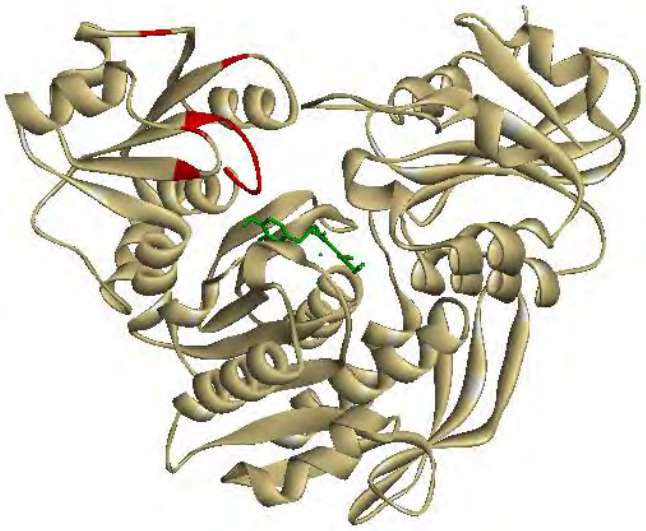
	131	-9.0	1.9	335.1	5	2
	133	-9.0	0.4	352.1	6	4
	416	-9.1	4.7	486.0	5	4
	465	-9.0	1.8	474.2	8	4
	525	-9.4	4.8	472.3	5	2
MurF	136	-9	2.7	303.1	5	3
	219	-9.6	4.9	366.2	3	1
	347	-9.4	3.9	434.1	7	2
	363	-9	4.3	304.2	2	1
	434	-9.2	3.1	374.1	6	2
	438	-9.2	3.1	374.1	6	2

B3: Full structures for docking results of MurC - MurF.

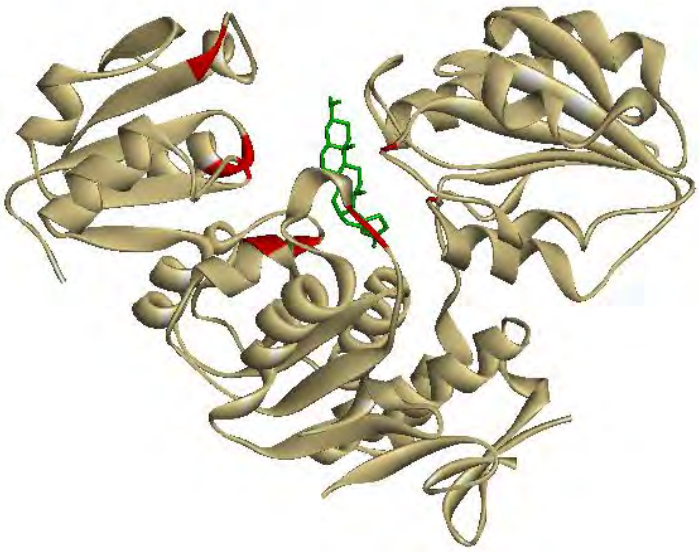
MurC-574



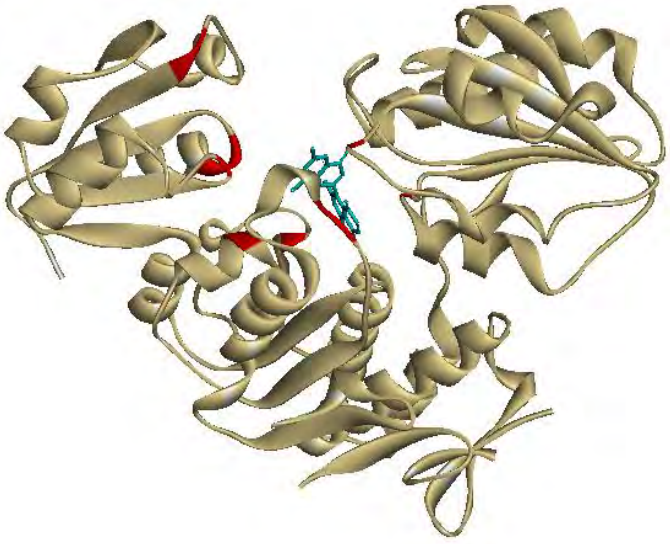
MurC-575



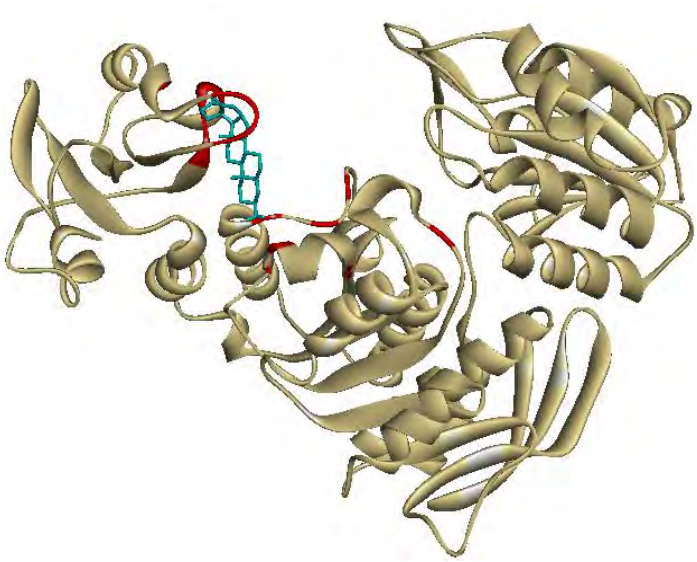
MurD- 290



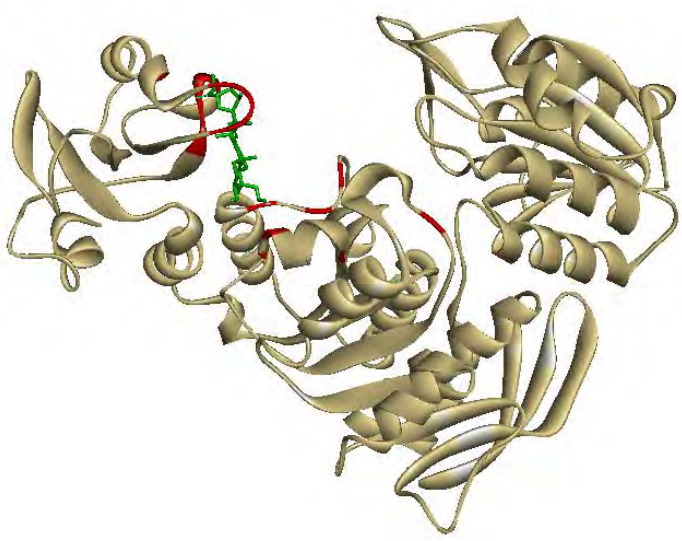
MurD- 438



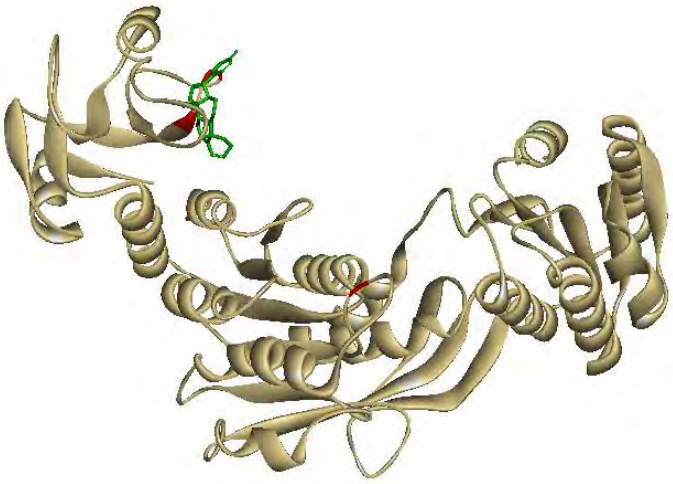
MurE- 290



MurE- 525



MurF- 219



MurF-434