Homology model of 30S ribosomal subunit from *Mycobacteria tuberculosis*

Mathenge Peterson Gitonga

Reg No: I56/63878/2010

Center for Biotechnology and Bioinformatics

University of Nairobi

A thesis submitted to the Board of Postgraduate Studies,

University of Nairobi, in partial fulfillment

For

The Award of Master of Science in Bioinformatics

C 2013

DECLARATION AND APPROVAL

This is my original work and has not been presented for a degree Program in any academic institution.

Name: Mathenge Peterson Gitonga

Registration: I56/63878/2010

Signature: _____Date_____

This proposal has been submitted with our approval as university supervisors:

Prof, Francis Mulaa, PhD

Department of biochemistry &

Centre for Biotechnology and Bioinformatics

University of Nairobi

Signature: _____Date_____

Fredrick Sijenyi, PhD

Senior Scientist,

DNA Software, Inc

Ann Arbor, Michigan

U.S.A

Signature: _____Date_____

Prof Peter W. Wagacha

School of Computing and Informatics

University of Nairobi

Signature: _____Date_____

P.O Box 30197- 00100

Nairobi, Kenya

DEDICATION

I dedicate this thesis to my family, to my mum whose encouragement has been my inspiration to move on and never give up, to my brothers for being a shoulder to learn on when things get tough. Also to Sister Mary and Baldo children Home for molding me to be the person I am today.

ACKNOWLEDGEMENT

First and foremost, my utmost gratitude to God my father for the good health and sound mind he bestowed upon me throughout this research. I acknowledge my research supervisors; Prof Fred Sijenyi for his dedication toward my success, always being ready to assist in spite of his busy schedule and the distance between us. I appreciate your dedicated commitment.

Prof Francis Mulaa, for a constructive critique during the research process. Also for his help in bridging my knowledge gap in biology and making the entire process interesting. I also appreciate his guidance throughout the whole process. Prof Peter Waiganjo Wangacha for the vital role he played in bringing in the computation aspect in the research and for enabling my curiosity in computational sciences to see the light of the day. I acknowledge Prof. James Ochanda through whom I got the opportunity of studying in the best Bioinformatics center in the country. To my colleagues; Harrison, Eric and Maureiq for your support during my studies. To Dr Gachara and Julius for your encouragement and guidance in my academic life.

I am highly indebted to my family for being a shoulder to cry on. For their encouragement, love, care and support throughout my academic life, and most importantly during my research work. Finally I would like to appreciate the Baldo children's home for instilling good discipline in me and for molding me to be the man I am today.

ABSTRACT

Mycobacterium tuberculosis, the causative agent of the tuberculosis, has infected more than a third of the world population to date. It has been known to be a very aggressive bacterium that is highly resistant to current drugs that target Tuberculosis. Antibiotics such as viomycin and capreomycin have been shown to bind to functionally important regions of the bacterial ribosome inhibiting protein synthesis process thereby affecting the bacterial cells viability. It is hypothesized that, a three dimensional structure of the 30S ribosomal subunit of the bacterium, will bring about a novel approach on drug target, and will be important in the development of a new class of anti-bacterial compounds. It will provide a structural scaffold on which structure based drug design studies can be performed. *In silico* screening of ligands can be carried out to identify compounds that show binding potential on ribosome, ribosomal RNA (16S rRNA) or the ribosomal proteins. Since current methods for obtaining three dimensional structures of the macromolecules are slow and tedious, we demonstrate a faster and inexpensive way of generating structural models *in silico* by employing both *de novo* and homology modeling methods

In this thesis, we report modeling of the three dimensional structure of the 30S ribosomal subunit from *Mycobacteria tuberculosis* through the structure prediction methods mentioned above. We report a high resolution ribosomal structure comparable in quality to experimentally determined crystal structures. It is hypothesized that, the structure will bring about a novel approach on drug target, and will be important in the development of a new class of anti-bacterial compounds. It will provide a structural scaffold on which structure based drug design studies can be performed. *In silico* screening of ligands can be carried out to identify compounds that show binding potential on ribosome, ribosomal RNA (16S rRNA) or the ribosomal proteins. Compounds identified this way can be further studied for antibacterial activity. We hypothesize that the generation of the 30S ribosomal subunit from *Mycobacteria Tuberculosis* will provide a structural scaffold that will allow *In-silico* structure based drug design

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATION

RNA –ribonucleic acid

DNA-deoxyribonucleic acid

RMSD-Root Mean Square Deviation

RCSM- Research Collaborator for Structural Bioinformatics

rRNA- Ribosomal RNA

tRNA- Transfer RNA

TB-Tuberculosis

NMR-Nuclear magnetic resonance

BLAST-Basic local alignment and search tool

MTB- multi drug resistant Tuberculosis

FMET- *N-Formylmethionine*

PDB- Protein data bank

1 CHAPTER ONE

1.1 INTRODUCTION

The cell, is the basic unit of life with a human body having approximately 100 trillion cells (Yonath, 2010). These cells are made up of protein and they differ in numbers from one organism to another. There are many different types of protein in different organelles of the body in different organism. There are over 7000, different types of protein in typical eukaryote(Bashan and Yonath, 2008). In addition, the number and the type of the protein depends on the type of the cell and the class of the cells. For instance, in the liver there are many different types of the protein as compared to other organelles due to its function(Yonath, 2010).These proteins are constantly degraded in the body hence there is a required constant production to replace them.

Protein synthesis is performed by complex of apparatus comprising of Ribosomes, messenger RNA (mRNA) and the transfer RNA (tRNA)(Yonath, 2010). Ribosomes are made up of two segments that come together during protein synthesis and dissociate after the protein is synthesized(Bashan and Yonath, 2008) (Figure1.1) and (figure 1.2). Eukaryotic ribosomes differ from the prokaryotic ribosome mostly in the translation processes, and also in that, the eukaryotic ribosome is 40% larger than the prokaryotic ribosome due to presence of additional rRNA element called the expansion segment and extra protein moieties (Ben-Shem et al., 2011, Liu et al., 2011)(Ben-Shem et al., 2011, Liu et al., 2011)(Ben-Shem et al., 2011, Liu et al., 2011)

In the prokaryotic ribosome, the small subunit (30S) contains a single copy of a ribosomal RNA (16S rRNA) and 21 proteins(Ramakrishnan, 2002) (Figure1.3). The larger subunit (50S) contains two Ribosomal RNA (5S and 23S) and 31 proteins.

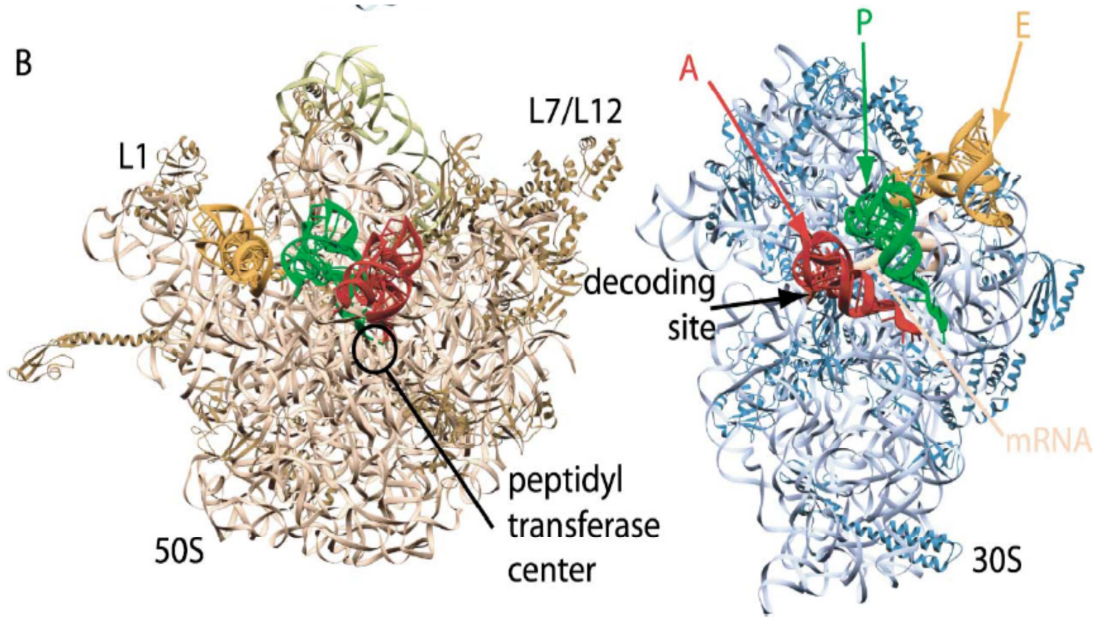


Figure 1:1, The larger and the smaller subunit and the location of the active (A), peptidly(P) and the exit(E) sites Prokaryotic ribosome (Ramakrishnan, 2002)

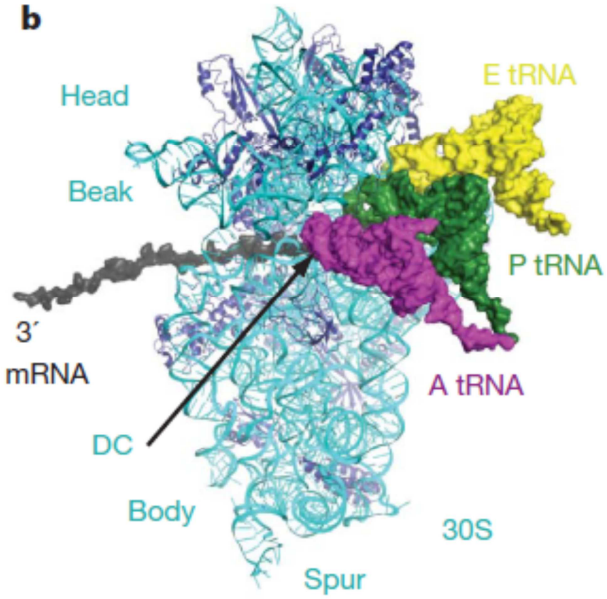


Figure 1:2 The head, beak body and the spur of the 30S ribosome structure (Schmeing and Ramakrishnan, 2009)

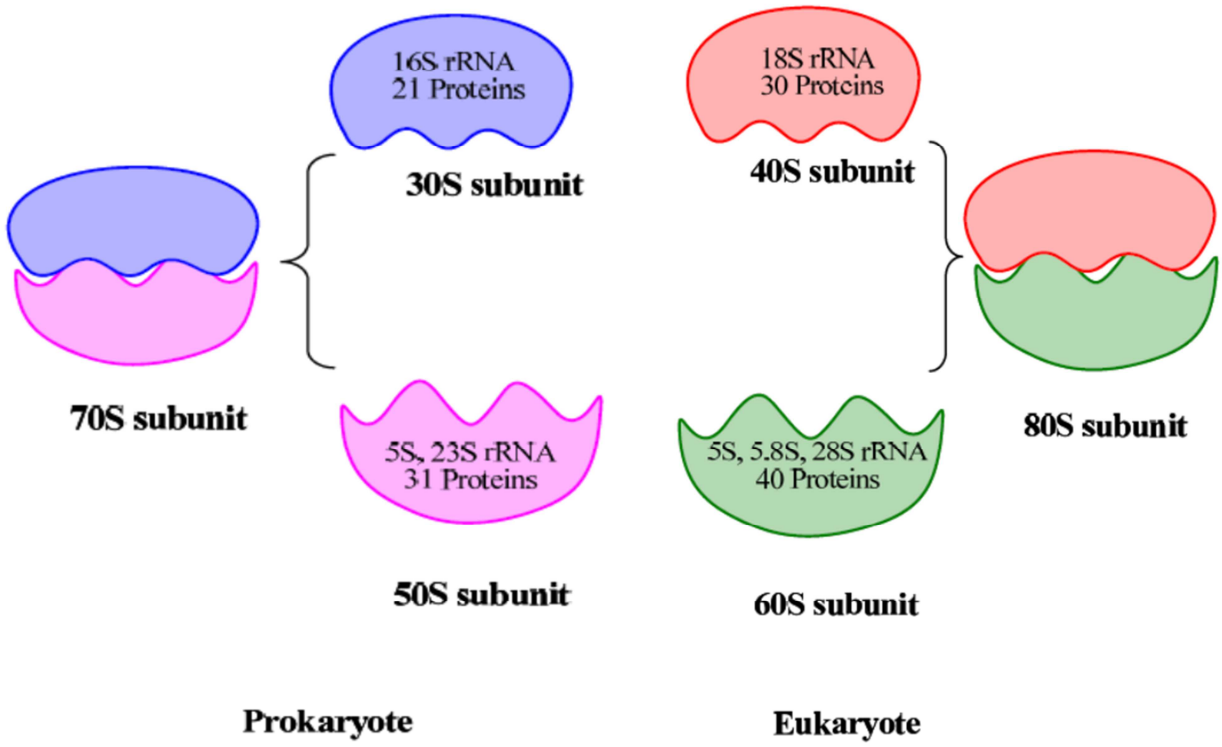


Figure 1:3 Prokaryotic70S ribosome on the left is made up of the smaller 30S subunit contains 21 proteins and one 16S RNA and the lager 50S subunit which contains two ribosomal RNA and 31 proteins. On the right is the Eukaryote ribosome which has got the small subunit 40s containing one 18S RNA and 30 proteins; the lager subunit the 60S contains three ribosomal RNAs and 40 proteins.(Sijenyi et al., 2012a)

In eukaryotic ribosomes, the lager subunit, (60S) consist of three ribosomal RNA (rRNA) molecules (25S, 5.8S and 5S) and 46 proteins. The small subunit (40S) is comprised of one rRNA chain (18S) and 33 proteins (Moore, 1998)(figure1.3).

Translation is the process by which messenger RNA (mRNA) binds to the ribosome and directs the production of protein from amino acids (Yonath, 2005). Although the process was understood some years back, some of the detailed information was not available until recently. E.g. how the antibiotics bind to the ribosome. This was due to lack of 3 dimensional ribosome structures (Ramakrishnan, 2002), owing to its large size and complexity. Three dimensional structures can be modeled through several methods; experimental methods which includes x-ray crystallography, nuclear magnetic resonance and Cryo-electromagnetic spectroscopy. X-ray crystallography stands as a gold standard in the protein modeling, but it is limited only to crystallized structure, this makes it difficult to get three dimensional structure of the protein that cannot be easily crystallized or those that cannot be crystallized at all (Yonath, 2005). Nuclear magnetic resonance (NMR) on the other hand is limited to the size of the structure. Cryo-electromagnetic spectroscopy often ends up yielding low resolution structure and also requires a lot of computational and calculation effort (Pieper et al.). Experimental methods mentioned above are however limited by the time it takes to generate a model and also the size of the molecule that can be modeled at any given time (Pieper et al.). Other methods of protein and the RNA three dimensional structure prediction include the comparative method and the *de novo* modeling (Liu et al., 2011). These methods are often faster than the experimental methods and often lead to structures of comparable quality. In addition, they allow structures to be generated for molecules that are not easy to crystalize or create NMR samples for. Hence their suitability for the work reported in this thesis.

Ribosome as Drug Target

Understanding the ribosome is crucial in the understanding of how protein are synthesized and also in the understanding the different states of the ribosome (Moore, 2012). Ribosomal proteins have been used as the drug target for a long period of time. Some antibiotics e.g. capreomycin have shown target to the rRNA which is highly conserved (Moore, 2012, Akbergenov et al., 2011). Since most of the antibiotics target the protein synthesis process, getting three dimensional structure of the ribosome and understanding how it functions, is crucial in the

process of understanding how those antibiotics work (Liu et al., 2011). It has also been noted that some antibiotics target the 16S rRNA of the ribosome, which is one of the highly conserved regions of the gene(Moore, 2012) hence by getting the three dimensional structure of 16S RNA would help in the understanding of interactions between RNA and antibiotics. It will also help in understanding how mutations are induced e.g. In the case of resistance to streptomycin(Meier et al., 1994) . Due to time complexity of the 3 dimesional structure modeling process in the primary method of modeling the X-ray crystallography and the others, we demonstrate a quicker and a better way of modeling a large molecule such as the bacterial ribosome using homology and *de novo* modeling. This study focuses on the *in silico* modeling of the 30S ribosome from *Mycobacteria tuberculosis*, one of the leading opportunistic killer pathogens in the world.

Tuberculosis

Tuberculosis (TB), caused by the *Mycobacterium tuberculosis* bacteria, is one of the leading infectious disease that is a worldwide public health threat. The control and management of TB has been complicated by the emergence of a drug resistant strain, and latent infection (Stephan *at el.*, 2010). This has prompted scientists to dig deeper and come up with new methods to address its spread and to overcome the problem of drug resistance. One way of addressing drug resistance would be to develop new drugs that target new areas of the pathogenic bacteria. The bacterial ribosome is a validated drug target (Shasmal and Sengupta, 2012), and thus a good candidate for the study of new anti-bacterial anti-infective. Structural studies on *M. tuberculosis* ribosome may help in understanding mechanistic models that interpret collected biochemical data(Shasmal and Sengupta, 2012).

1.2 RESEARCH QUESTION

Can *Mycobacteria tuberculosis* 30S ribosome be modeled into a three dimensional structure, and what are the potential sites that can be exploited for structure based drug design?

1.3 OBJECTIVES

Overall objective

1. Generate the structure for the 30S subunit of *M. tuberculosis.*

Specific objectives

1. To determine the three dimensional structure of the *M. tuberculosis* 16S rRNA.
2. To generate protein homology models for the 21 proteins from the 30S subunit of *M. tuberculosis*.

1.4 RATIONALE

The 30S ribosomal subunit's three dimensional structure is vital in the understanding proteins synthesis in bacteria, the location of the active site in the ribosome and also in the comprehension of how certain antibiotics e.g. viomycin, are able to inhibit protein synthesis. The knowledge acquired can be used in the future in coming up with a novel drug that can be used to target the ribosome in the Prokaryote of interest. However, this can only be achieved by having a quick and easier way of modeling; unlike the Nuclear Magnetic Resonance and the X-ray Crystallography that may take more time.

2 CHAPTER TWO

2.1 LITERATURE REVIEW

2.1.1 Homology Modeling

Homology modeling is the process by which the three dimensional structure of a bio macromolecule such as RNA or a protein is generated through a series of computational steps(Krieger et al., 2003). This requires the use of an existing homologous structure as a template and then threading in the sequence to be modeled. Usually it is accomplished by using a series of computational algorithms. In comparative (Homology) modeling, two conditions have to be met; first, the query sequence must have a detectable similarity to the template. Secondly, it must be possible to compute an accurate alignment between the target sequence and the template structure. This is because homology modeling relies on detectable similarities covering most of the modeled sequence and at least one known structure (Walker, 1996, Walker, 2005). *De novo* methods on the other hand allow the prediction of the tertiary structure from the sequence alone, without relying on similarity of the fold level between the template and any of the known structure (Jacobson, 2004). These two methods usually complement each other during structural modeling as regions that are not similar to both query and template have to be built in from sequence information.

Homology modeling involves multistep processes which can be summarized in several steps (Krieger et al., 2003). The first step involves template recognition and initial alignment; this involves the use of tools like BLAST (Altschul et al., 1990, Altschul and Gish, 1996) to get the best alignment between the query and a possible template. Alignment correction is the second step in homology modeling, it involves the use of more sophisticated methods to align, one such tool that is commonly used in multiple sequence analysis is clustalW, (Li, 2003). However, sequence alignment alone is not sufficient as it does not capture the secondary structure information in both template and query structure which is very important in ensuring that the correct fold in modeled. Tools such as RNA123 have incorporated structure based sequence alignment algorithms that allow the secondary structure information to be incorporated during

sequence alignments (Sijenyi et al., 2012b). After sequence alignment, modeling is achieved by a series of modifications on the template structure guided by the chosen alignment. These include operations such as insertions, deletions and *de novo* building of new segments followed by energy minimization to address steric clashes that might be introduced during this process. The fourth step involves structure validation which includes checking the bond length, bond angles and any presence of backbone gaps or any structural inconsistencies in the model.

Homology modeling is mostly preferred because it can be accurate if a highly homologous template is identified, as compared to other physics based approaches in the comparative modeling (Kolinski and Bujnicki, 2005). It is also faster than other modeling methods like *de novo* structure prediction. The only shortcoming of homology modeling compared to the others is that it has to use a three dimensional structures that have already been modeled using other methods as a template when modeling a new structure.

2.1.2 Ribosome

Ribosome is a large nucleoprotein particle that synthesizes protein in all living cells, using messenger RNA as the template and aminoacyl-transferace RNA as the substrate (Schmeing and Ramakrishnan, 2009). The Ribosome is made up of RNA and protein. The ratio of ribosomal RNA residue to protein residue in the ribosome has been maintained at two to one (2:1) with the exception of mitochondria ribosome (Schmeing and Ramakrishnan, 2009).

RNA

Ribonucleic acid is made up of ribose sugar, a phosphate group and a base. Unlike DNA, RNA is usually single stranded. The bases present are Uracil, Adenine, Guanine and Cytosine. Cytosine is replaced by thiamine in DNA (Figure 2.1). Apart from the four nucleotides, it is believed that there are other many modified nucleotides in the RNA especially in the tRNA, mRNA and ribosomal RNAs (Aduri et al., 2007) . In a normal paring, cytosine pairs with the guanine and adenine pairs with the uracil, this is referred to as the canonical pairing (Gardner et al., 2011). But mismatches are also observed in RNA whereby guanine may pair with another monomer apart from cytosine. This is referred to as non-canonical pairing.

RNA differs from the DNA in several ways which includes: The DNA is normally double stranded while RNA is single stranded. The base Uracil is usually present in the RNA but in

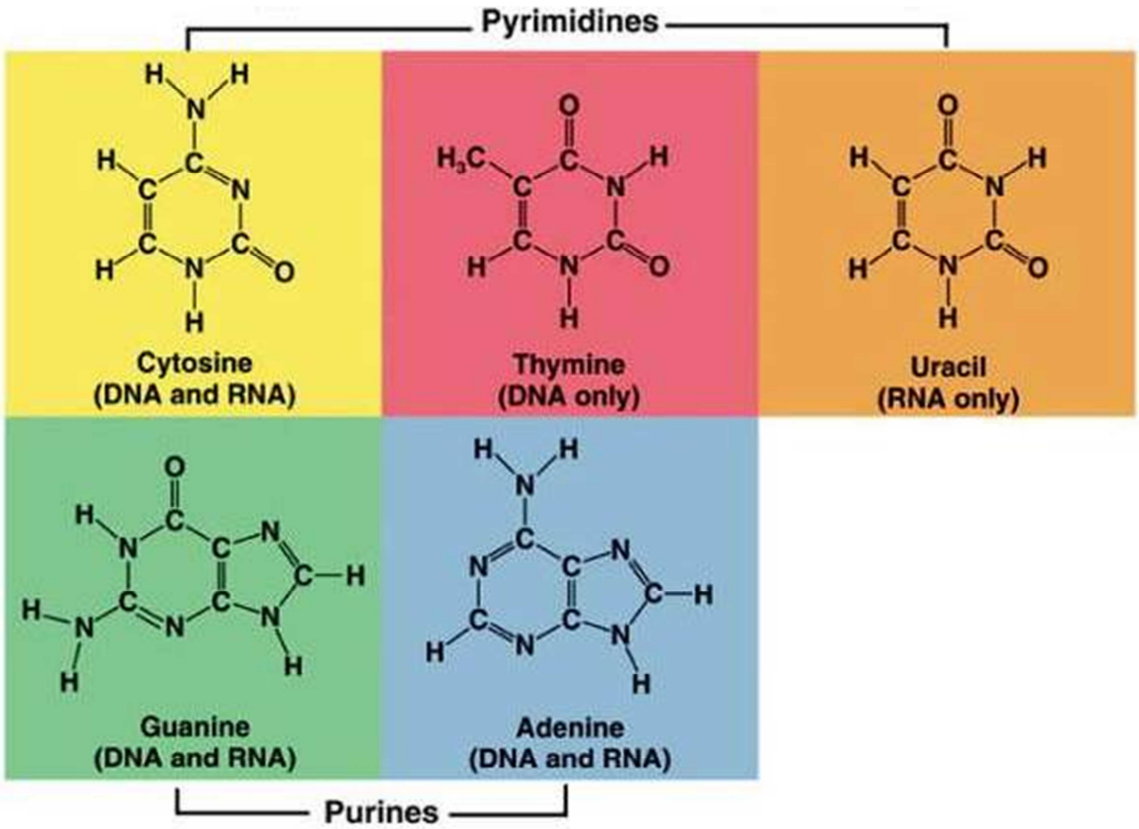


Figure 2:1 Nitrogenous bases that makes up the nucleic acid : Thymine is found in DNA in place of Uracil in RNA (Mader and Baldwin, 2007)

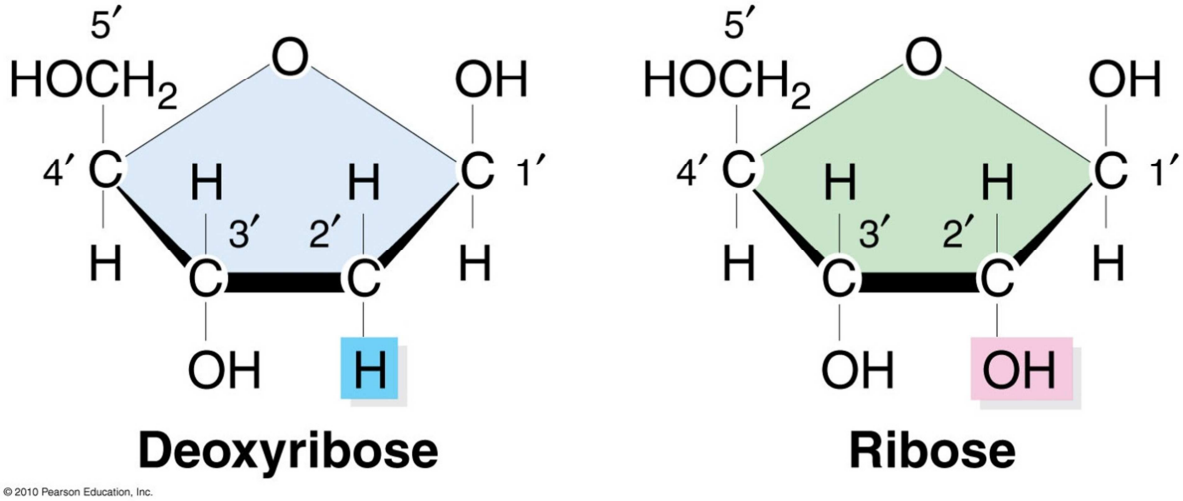


Figure 2:2 Differences between DNA and RNA nucleoside is the presence of an hydroxyl (OH) in the sugar ring in RNA which is missing in DNA nucleoside (Enger et al., 2003)

DNA it's replaced by thiamine. The sugar present in the RNA is ribose but in DNA it is a deoxyribose sugar (Figure 2.1).

There are different types of RNAs which differ in size and in function. This includes; messenger RNA (mRNA) which contains the transcribed part of a gene from the DNA, and it's carried to the ribosome for translation purpose. Transfer RNA (tRNA), which mediates the process of translation by providing the required nucleotides (Zuo et al., 2013). Other RNAs include the short interfering RNA, which is mostly used in gene silencing by cleaving and destroying the massager RNA(Hackenberg et al., 2013). MicroRNAs induce translation repression (Kim and Nam, 2006) by binding in a specific way to the massager RNA. There many others RNA which have different functions in the cell such as snoRNAs (Ge et al., 2010), piwi RNAs (Leger et al., 2013)etc. Unlike protein, RNA has several internal structures that are vital in its functioning. The internal structures include the internal loop, hairpin, motif, pseudo knot and bulges. (Figure2.3). each of the internal structure has a specific function in the RNA. The pseudoknot which are observed in almost all the RNAs, are involved in the catalytic function of the RNA whereby they act as the initiator of the gene translation that regulate plasmid copy number this help in stabilizing different organelles (Gupta et al., 2012) .

16SRNA

It is found in the smaller subunit of the prokaryotic ribosome, the 30S, and it is the only RNA in the subunit. It is one of the highly conserved RNA across all species. It varies in size from organism to organism, for instance, in *Escherichia coli* it is 1531 nucleotides long while in *Mycobacteria tuberculosis* it is 1538 nucleotides long. Due to its large size, it is easily sub divided into four domains; the 5' domain, central domain, 3' major and the 3' minor domain (Figure 2.4).

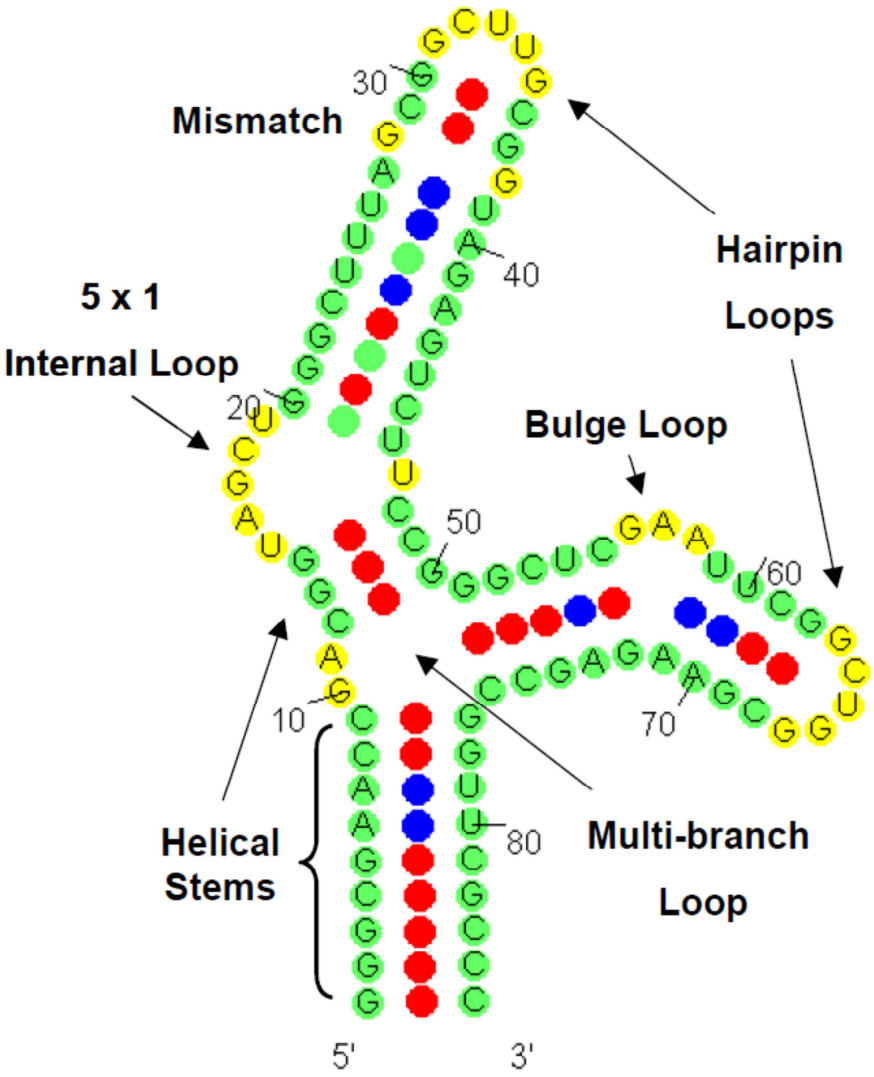


Figure 2:3 Secondary structure of the RNA containing different components. In the cell, these parts are vital in the normal functioning of a given RNA structure (Sijenyi et al., 2012b)

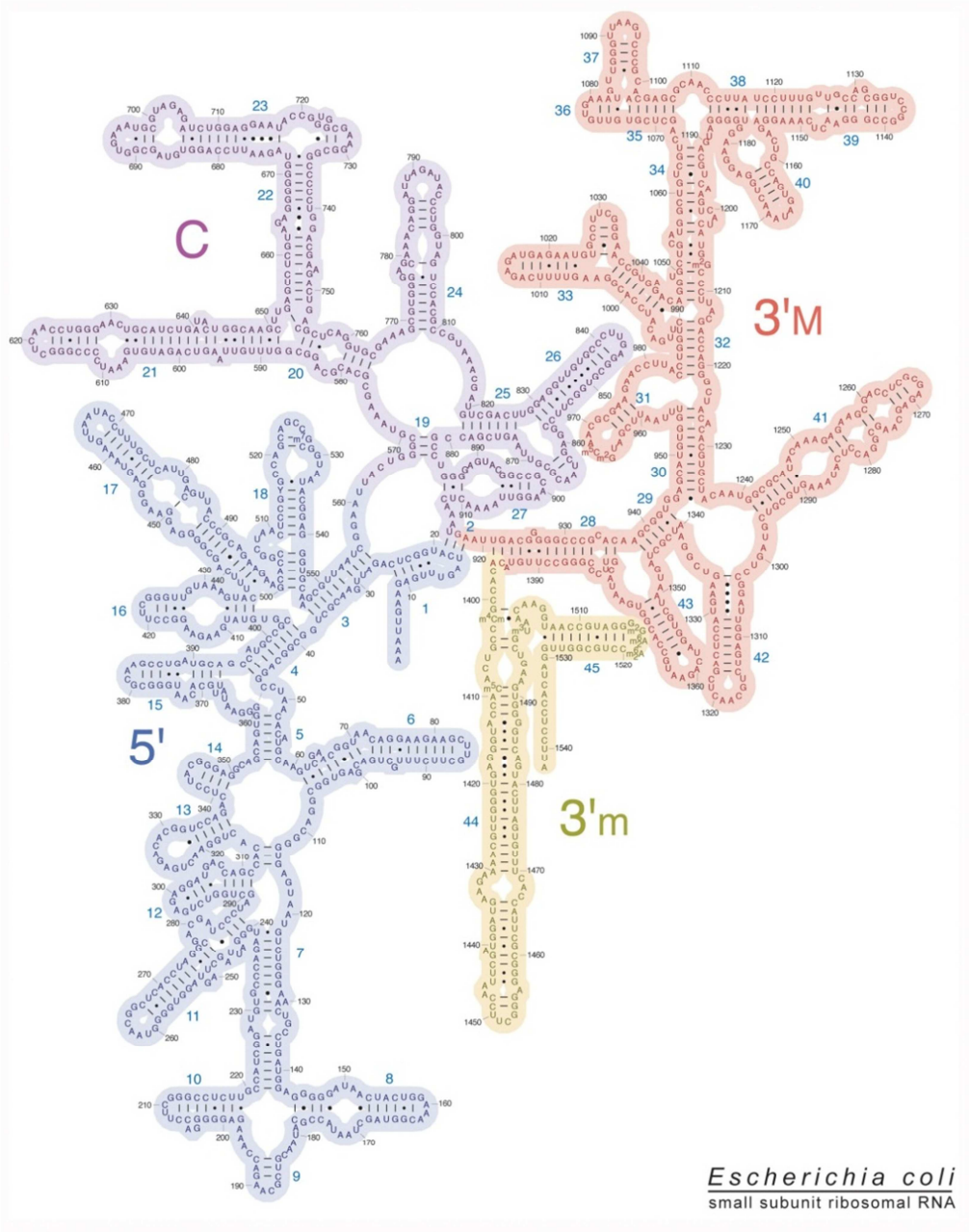


Figure 2:4 Secondary structure of the 16S rRNA having different domains colored differently ; five prime (5'), central domain(c), three minor (3'M), and the three major (3'm)(Cannone et al., 2002)

RNA as a drug target

Antibiotics have had different targets in bacteria, some have targeted the cell wall synthesis e.g. penicillin which targets the final cross linking of the peptide side chain during peptidoglycan synthesis (Tipper and Strominger, 1965). Others have targeted the DNA and the RNA e.g. Rifampin (Walsh, 2003). Others like the aminoglycosides and tetracycline have targeted the ribosome especially small subunit (Walsh, 2003). But in spite of different antibiotics targeting different parts of the bacteria, bacteria have always found a way of becoming resistant to the drugs. Bacteria can develop antibiotic resistance through a variety of pathways, the main one being through mutation. For instance the bacteria *M. tuberculosis* become resistant to antibiotic streptomycin through a mutation (Meier et al., 1994). Streptomycin targets the S2 protein in the small ribosomal subunit, but mutation in the 16S ribosomal RNA at position 513(A->C/T) and 516(C->T) lead to conformation difference of the binding site of streptomycin, leading to resistance of the bacteria to the antibiotic (Meier et al., 1994)

Due to the drug resistance challenges it is believed that RNA can be a good lead for drugs target unlike the protein (Hermann and Westhof, 1998) . This is because; Unlike the DNA, RNA has no proofreading mechanism, hence if targeted; it has a low chance of recovery to normal function. Secondly, RNA plays a vital role in the body, every protein that is synthesized in the body first has to be coded into RNA before being taken to the ribosome for translation purpose(Lind et al., 2002), hence targeting the RNA e.g. massager RNA will ensure that no correct information is translated in the ribosome. Also when the ribosomal RNAs are the target e.g. when 16S rRNA, ribosome is targeted it cannot function normally and hence no protein synthesis, leading to cell death.

16S RNA is one of the highly conserved RNA across species, it codes for the ribosomal protein in the small subunit. A mutation in the 16S ribosomal subunit leads to a different structural conformation of the protein synthesized in the ribosome. The aminoglycoside antibiotics that target the 16S rRNA target the active site of the complex (mostly between positions 1409 to 1491) and thus preventing any mRNA from getting in (Hermann and Westhof, 1998).

Drugs that target rRNA do so by either interacting directly with the ligand–biding surface or by allosteric effect on the structure (Pearson and Prescott, 1997). Hence the availability of a three dimensional structure of the rRNA would allow RNA pharmaceutical companies to develop

drugs that are directed to the active site of the ribosome through *in silico* bidding affinity tests; this would expedite the process of drug discovery.

Insilco studies have shown that RNA can take different conformation shapes at any given time, and that different drugs may dock at the different structure. Hence availability of RNA three dimensional structures can be exploited in the new drug discovery analysis (Lind et al., 2002)

PROTEIN

Proteins are made up of peptides or several polypeptides joined together. The peptides are made up of amino acids which are the building blocks of protein. Proteins are classified according to their domains; the distinct components of three-dimensional structure which is able to carry out certain molecular function e.g. binding, catalysis and others (Jones et al., 1992, Cymerman et al., 2008). Small proteins tend to have one domain but a bigger protein may have several domains (Murzin et al., 1995).

Protein function is mostly associated with its structure folding more than its sequence, in that, protein with low sequence similarity may have similar folding and also similar function. Thus, two proteins may have similar sequence but have different folding hence having different function(Cymerman et al., 2008).

Protein structure can be classified into levels, which include primary level; refers to amino acid sequence making up the protein. Secondary level refers to a classification of protein into beta sheet or alpha helix. Tertiary structures refers to the three dimensional structure of the protein and their folding. Finally quaternary structure is the classification where several tertiary structure join up together to form one big complex (Cymerman et al., 2008)

Protein classifications fall under several categories; family level, this involves clustering of proteins with the same evolutionary origin whereby proteins with a residue similarity of more than 30% and with same function and folding are grouped together .(Lo Conte et al., 2000)

The superfamily involves proteins that have a low sequence similarity but their function and structural similarity indicate that they have the same evolutionary origin (Lo Conte et al., 2000)

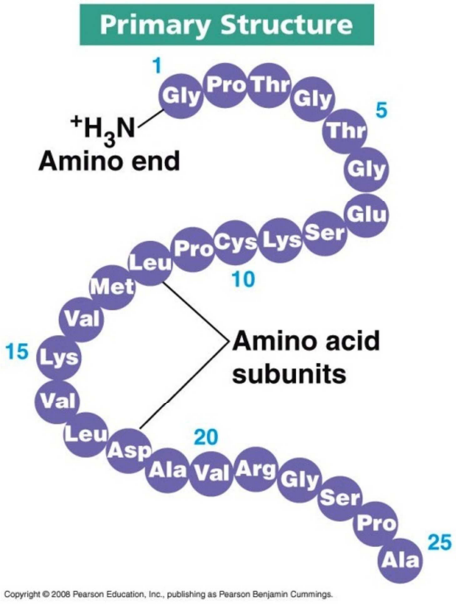


Figure 2:5 Protein primary structures (Mader and Baldwin, 2007)

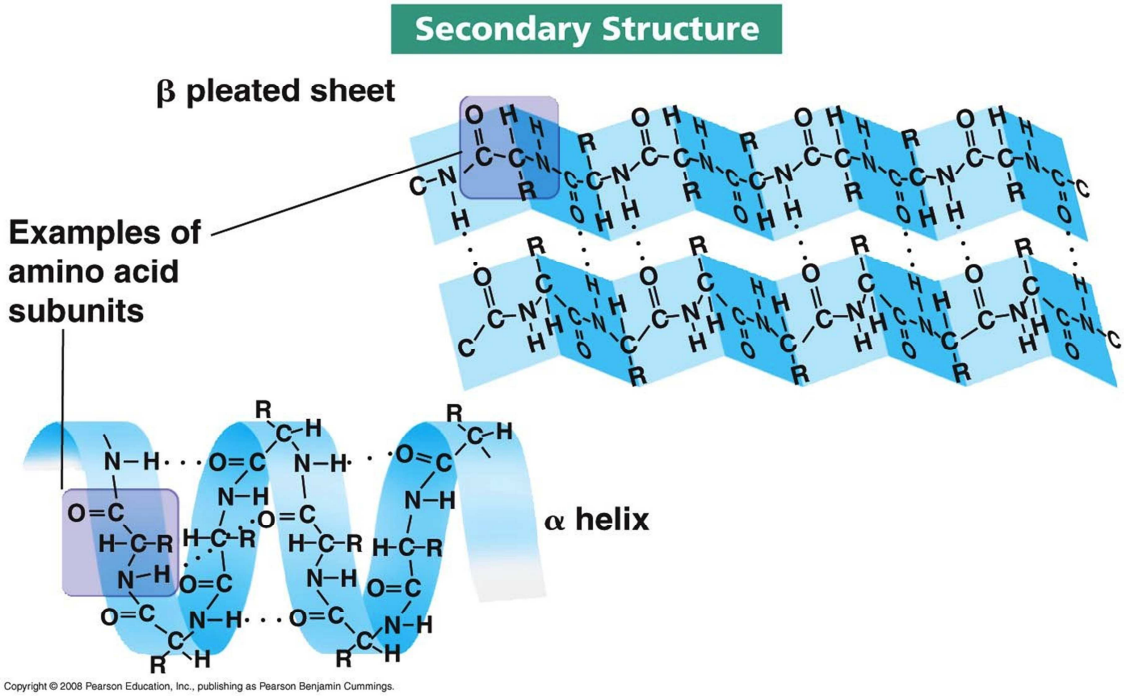


Figure 2:6 Protein secondary structure with beta and alpha(Mader and Baldwin, 2007)

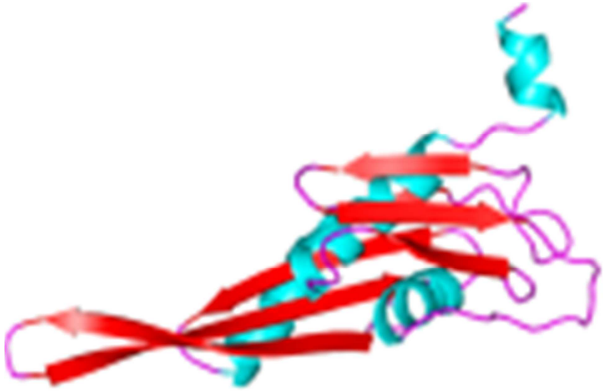


Figure 2:7 Tertiary structure of the protein showing both the beta (in red) and the alpha helix (in cyan) of protein. 2AVY_s5(Schuwirth et al., 2005)

Proteins can also be clustered into common fold. This happens when the family and the superfamily secondary structure has the same conformation and the same topology (Murzin et al., 1995). Finally, proteins can be classified according to their folding patterns; ,all alpha, all beta, alpha and beta, alpha plus beta and multi domain(Murzin et al., 1995).

The large subunit of the ribosome houses the peptidyl transferase center that catalyzes peptide bond formation (Shasmal and Sengupta, 2012). These centers includes; the P site, the site occupied by the peptidly tRNA just before peptide transfer, it holds the tRNA with the nascent peptide chain. The A (aminoacyl) site which accepts the incoming aminoacylated tRNA, and the E (exit) site which holds the declylated tRNA in place as it leaves the ribosome. In addition there is a T site through which the amino acyl tRNA pass on its way to the A site. This happens during the translation (Moore, 2012). The smaller subunit determines the sequence of the product to be made by mediating the base paring interaction between messenger RNA (mRNA) and the transfer RNA(tRNA)(Moore, 2012).

Translation is the process of protein synthesis (Figure 2.8). During which many ribosomes comes together and act simultaneously along the mRNA forming structure known as polysome. They act fast and efficient hence producing protein continuously, and in a high speed of approximately 20 peptide bond per seconds (Yonath, 2010). This is a result of the high level of protein degradation in the body which requires protein to be produced to replace them.

Translation is carried out in three stages, initiation elongation and termination (Schmeing and Ramakrishnan, 2009).

In the initiation process, the ribosome places the initiator codon fMet over the start codon of the mRNA in the P-site in the bacteria hence triggering the whole process of protein synthesis to start (Krupkin et al., 2011). The second stage is the elongation, which involves the sequential addition of the amino acid to the growing polypeptide chain. Elongation process continues until the stop codon in the sequence is moved to the A site which signals the end of the coding sequence(Schmeing and Ramakrishnan, 2009). This is followed by the cleavage of the nascent polypeptide chain from the P-site tRNA hence releasing the newly synthesized protein from the ribosome.

Many of the antibiotics have been targeting this process of translation, to impair protein synthesis in the bacteria. But it has been found out that bacteria are becoming resistance to most of the drug (Amabile-Cuevas, 2012). This is as a result of reckless abuse of antibiotics not only

by the clinicians, but also by the lay public, who demands or self-prescribe antibiotic for a number of ailments without first having a test to ascertain the condition. This has led to the pathogenic microorganism developing resistance to the antibiotic (Kim et al., 2009). Bacteria resistance to drugs manifests in several ways; The bacteria can inhibit the entry of antibiotic into the cell, it can produce substances that degrade the antibiotic, it can have a mechanism of removing the antibiotic from the cell before it acts on it and it can change the conformation of the target area by the antibiotic through mutation -usually in the ribosomal RNA. For instance, the resistance of *M. tuberculosis* to the antibiotic streptomycin. As mentioned in section 2.1.2. was initiated by a single nucleotide mutation (snip) at position 513(A->C/T) and 516(C->T)(Meier et al., 1994). Since streptomycin acts by inhibiting protein synthesis during translation, mutation in the rRNA lead to change of the sequence of the S2 protein which in turn lead to a different conformation structure making it difficult for the antibiotic to infer its action to the bacteria.

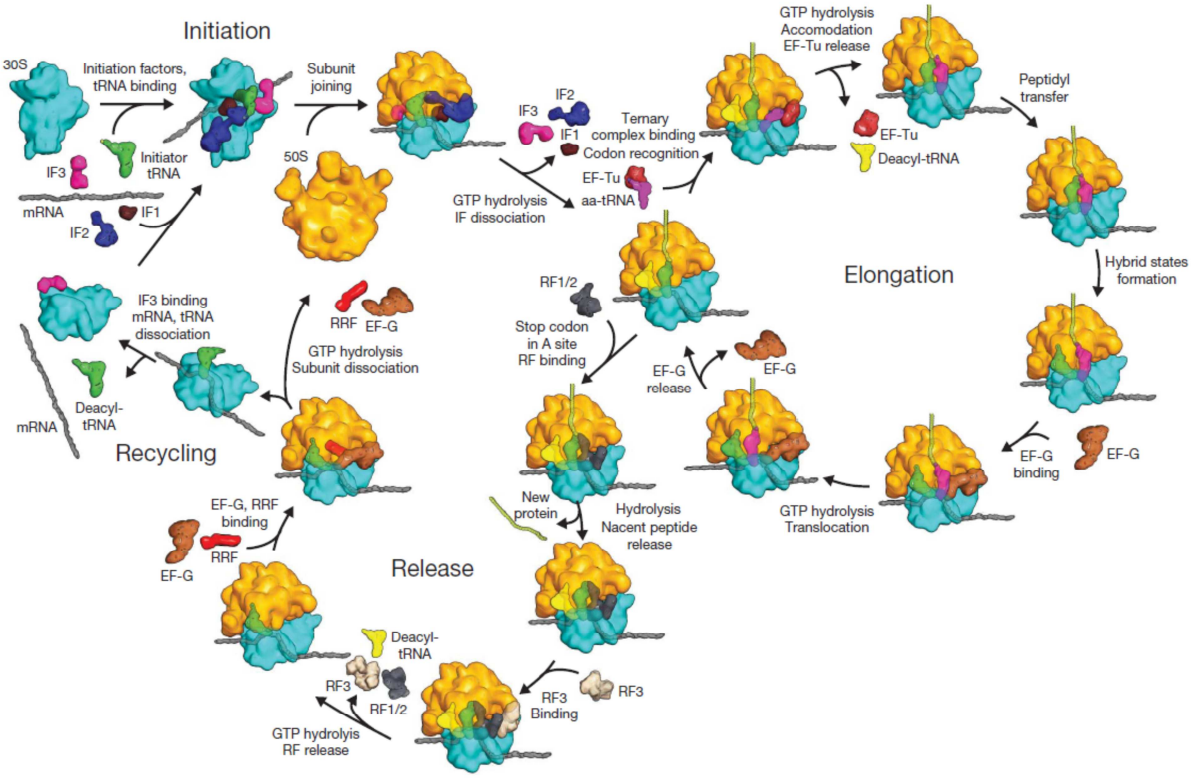


Figure 2:8 Protein synthesis in the ribosome starting at the initiation, Elongation and release (Schmeing and Ramakrishnan, 2009)

2.1.3 Tuberculosis

Mycobacteria tuberculosis the causative agent of tuberculosis, is a rod-shaped, gram positive, G-C reach bacterium discovered in 1882 by Dr. Robert Koch (Cole et al., 1998). It is identified by Acid fast staining test. This follows its addition layer made up of lipids glycoproteins and polysaccharides above the peptidoglycan layer which makes it difficult for other staining test to be used to identify it (Glickman and Jacobs, 2001).

Mycobacteria is transmitted mostly though inhalation of aerosolized droplets containing infectious *M. tuberculosis* from an infected person through a cough to a healthy person. the inhaled bacilli accumulate and get access to the alveolar microphages where they multiply this leads to primary pulmonary (Glickman and Jacobs, 2001). The resultant primary parenchymal Ghon focus usually drains via lymphatic nodes and the combination of the Ghon focus, local lymphangitis and the regional lymph node involvement is known as the Ranke complex(De Backer et al., 2006).

Tuberculosis is an opportunistic disease, in that it can remain in the body in the latent stage until the body immune system goes down hence it is mostly associated with the HIV and AIDS condition. The latent stage is mostly found in the lungs in the alveolus in form of hard globules ball known as Ghon and can only be diagnosed through X-ray. Since tuberculosis has a latent stage which can be prolonged for a period of time, and it has very fast means of communication from one person to another, studies have shown that about 8-10 million cases of tuberculosis are reported every year and 2- 3 million succumb to the disease yearly (Harries, 1990). Due to the new strain of the bacteria(Mtb) the number of the infected people have risen with an increasing rate, and according to world health organization 1 out of every 3 people is infected with tuberculosis worldwide (Dye, 2006). Due to the alarming rate of the infection and the death, the global strategic plan for controlling Tb in the coming decade is reducing death rate, illness and its incidences.(Dye, 2000)

3 CHAPTER THREE

3.1 METHODOLOGY

3.1.1 Proof of concept

To validate the use of *RNA123* software (Sijenyi et al., 2012b) in the homology and *de novo* modeling of the 30S ribosomal subunit, we initially set out to model the structure of a bacteria whose structure had already been solved using X-ray crystallography. RNA structure of *Escherichia coli* (PDB ID 2AVY) was modeled using crystal structure of *Thermus thermophilus* as the template (PDB ID 2JOO). To automate the modeling process, the 16S rRNA template structure was subdivided into four domains the 5', central, 3' major and 3' minor domain. This was done based on the secondary structure of the query and of the template. As for the protein modeling, the process followed is as described in detail in the section 3.1.4.

Result for proof of concept

RNA results

The 16S rRNA from *E.coli* was modeled using the *T.themuphilus* as the template and then the structure obtained was compared to the original structure of the *E.coli* modeled by X-ray crystallography method (PDB 2AVY) through the calculation of the root mean square deviation (RMSD) between the two models. See Figure's 3.1 and 3.2 for a superposition of the models obtained for the 5' and the 3 'major domains with their corresponding crystal structure domains.

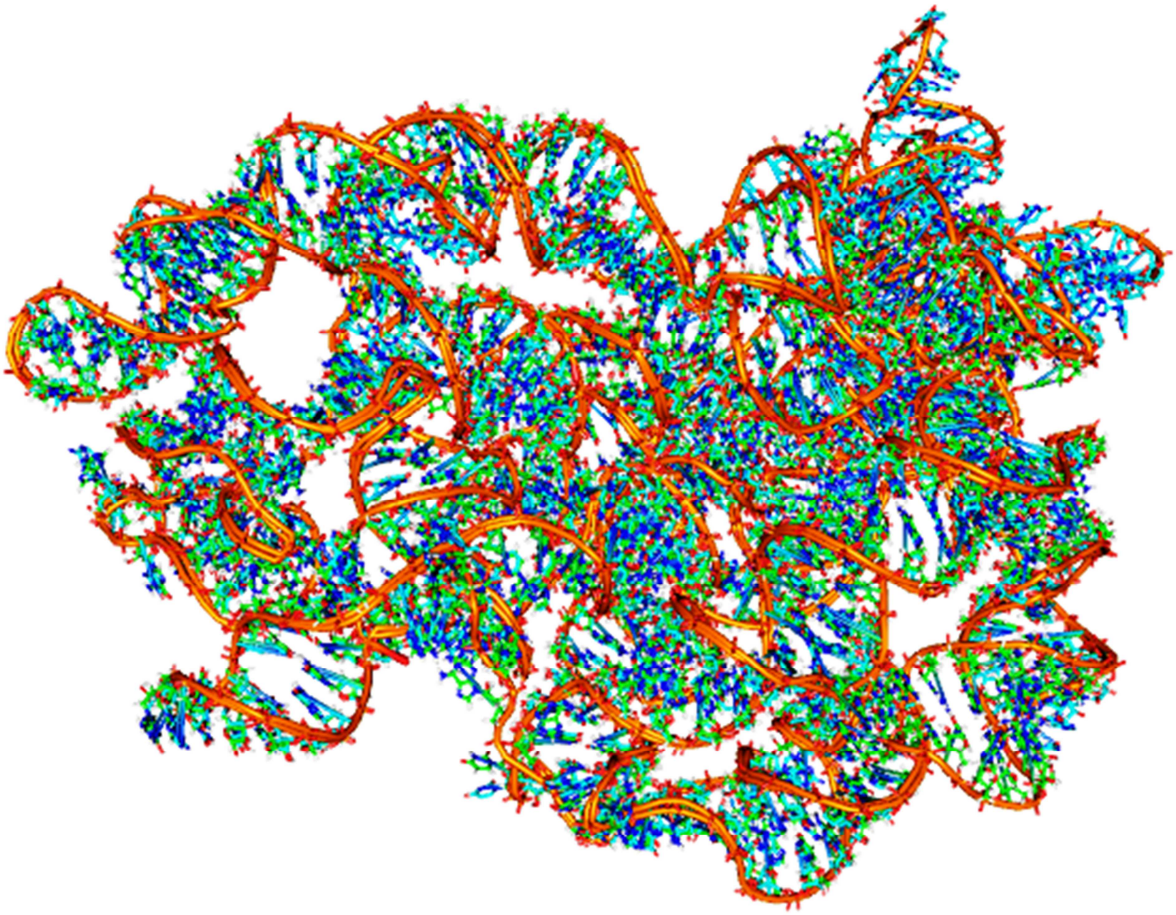


Figure 3:1 Homology modeled structure of RNA 5’ Domain superimposed to the structure solved by X-ray crystallography method (RMSD 4.5 Å). the green colour representing the modelled structure and the blue colour representing the X-ray crastallograpy moddelled template.

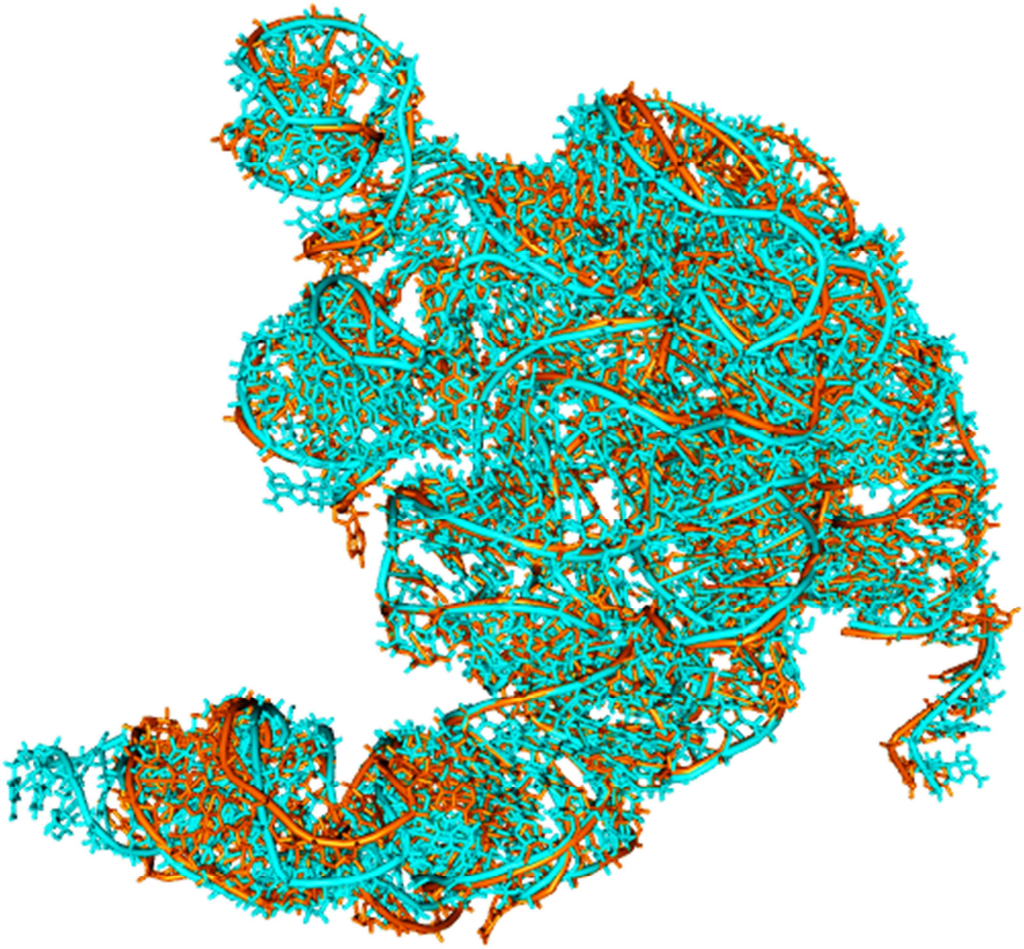
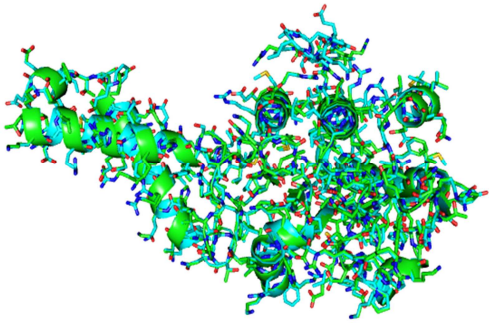
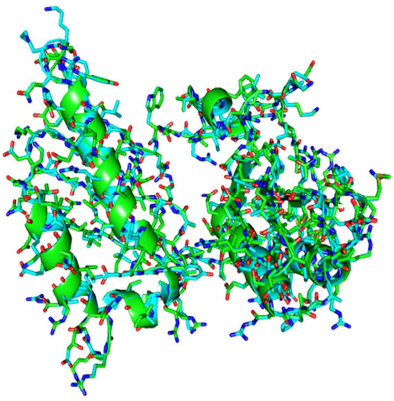


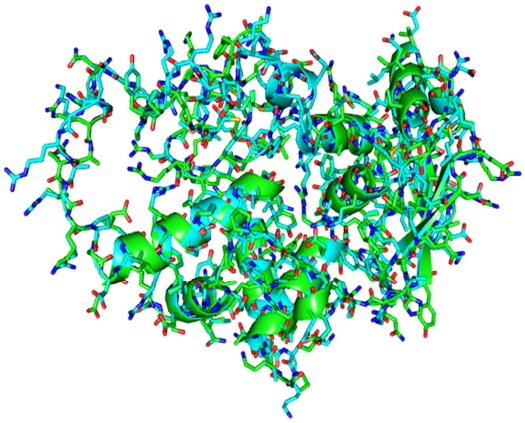
Figure 3:2 Homology modelled structure of 3 mojor domain of E.coli superimposed on the original structure, with RMSD of 2.5 Å. The cyan colour representing the moddelled structure and the other colour representing the template.



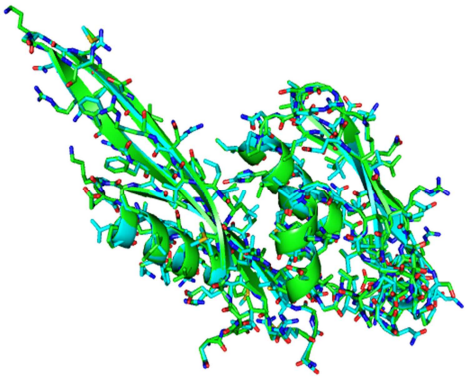
S2 protein



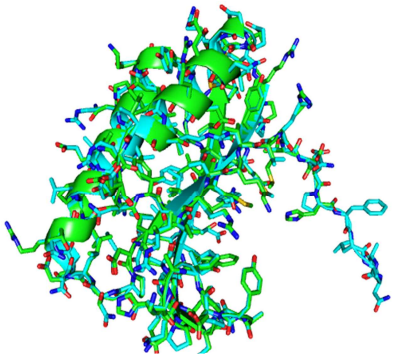
S3_protein



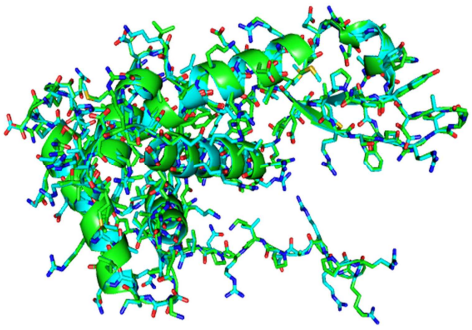
S4_protein



S5_protein



S6_protein



S7_protein

Figure 3:3 S2 to S7 protein of the modeled structure through homology modeling, in green superimposed on the original structure by X-ray crystallography

Table 1 containing the RMSD of the protein modeled, the number of or residue and the amino acid number in each protein

Name of the protein	Root mean square deviation (RMSD)	Residue number of the query and template	Number of the amino acids in the template
S2	1.51	1833	240
S3	1.92	1625	232
S4	2.74	1643	205
S5	1.86	1093	166
S6	2.43	818	135
S7	2.27	1229	178
S8	2.26	968	129
S9	2.57	1022	129
S10	2.25	779	103
S11	2.57	870	128
S12	1.93	945	123
S13	2.36	910	117
S14	3.13	475	100
S15	1.78	699	89
S16	2.68	644	82
S17	2.34	648	83
S18	2.91	554	74
S19	2.83	627	91
S20	2.24	652	86

The table above shows different protein and their root mean square deviation, the number of atoms involved in calculating the RMSD and the number of amino acids in each protein.

3.1.2 Generating protein template and the query

The sequence of the *Tuberculosis* strain H37Rv was used as the query in the modeling process (Glickman and Jacobs, 2001). This sequence was obtained from gene bank which can be accessed at **http://www.ncbi.nlm.nih.gov/genbank/**.

In this project, the 16S rRNA sequences from the *Mycobacterium tuberculosis* bacteria was used as the query with the published *E. coli* structure (PDB ID 2AVY) being used as the template. *E. coli* was used as the ideal template because its secondary structure and that of *M. tuberculosis* are similar(Schuwirth et al., 2005). Also, the 16S rRNA in the two has almost equal length. *E. coli* and the *T. thermuphilus* were chosen for this work they have higher resolution crystal structures available(Schuwirth et al., 2005).

As for protein modeling, using the protein sequence of the *M. tuberculosis*, homologous proteins were searched for using BLAST (Basic local alignment tool) (Altschul et al., 1990). This process identified the *E. coli* as good candidates for use as templates as all the 21 proteins had homologous counterparts in both organisms. The query and the template were submitted for modeling in the Swiss model server for protein homology modeling (Kiefer et al., 2009, Arnold et al., 2006, Schwede et al., 2003).

3.1.3 Generating individual protein templates coordinates

Protein co-ordinates for the template structures were obtained from the Research Collaborator for Structural Bioinformatics (RCSB) protein data bank, which is a biological macromolecular resource found online at www.rscb.org.

After getting the coordinates, sequence alignment was carried out. A good template should have a sequence similarity of above 75% to the query sequence which most of the protein query had with the template. Sequence identity of 50% between the template and the query can also be used although it requires structure correction. But sequence similarity of below 25% cannot be used for homology model as this indicates that the organisms are virtually different (Venselaar et al., (2003)).

3.1.3.1 Ideal template for homology modeling

The template and the query should have relatively same length. This is because if the query is longer than the template, when modeling the non-homologous regions will require the use of other modeling methods such a *de novo* modeling which are more complex and are also limited to the size of structure to be modeled at any given time.

For the RNA, the template and the query should have the same number of multiloops. This is important because of the complexity involved in modeling RNA structures that have different multiloops. (Sijenyi et al., 2012b).

In modeling both the RNA and the protein it is also recommended that a high resolution template be used. Using a low resolution template often leads to a poor quality models

3.1.4 Protein modeling using Swiss Model

Online Swiss model sever provides users with three approaches based on the task and the experience of the user (Arnold et al., 2006). The three modes are; the Alignment mode, project mode and the automated mode.

The automated mode is used when the template and the query has a high similarity of more than 50% and it require Uniprot accession code or amino acid sequence of the query, after which the sever does a BLAST search and selects the template depending on the E-value of the result. Apart from the e- value the saver also checks for the template with the best resolution, and best stereochemistry among the hits and it selects it (Arnold et al., 2006).

In the Alignment mode, the user supplies the query and the template to be used in the modeling in a faster format. The sever models the query with the use of the template given. In this mode there is a place for inputting the query and the another place for inputting the template (Schwede et al., 2003).

Project mode involves preparing a file containing the template and the query, the file is made in a way that the two are superimposed on each by the use of deep view viewer software (Arnold et al., 2006). This mode is used when one is working on many proteins at a go and unlike the others, in project mode the use has got a full control on template selection (Arnold et al., 2006).

During the process of protein modeling, Swiss model uses template structure database to search for the template, this database is usually derived from protein data bank (Westbrook et al., 2003). To make it easy for the Swiss model to search for the template in its database, the protein entering the database are usually stored as individual chains, and they are tagged with the protein information e.g. the protein resolution, method used in modeling their energy and others (Westbrook et al., 2003, Arnold et al., 2006). In the process of template search the sever may encounter several problems and may not find a match (Kiefer et al., 2009). In this case, the sever uses BLAST(Mount, 2007) to search for similar protein. BLAST may employ several methods for searching of the protein in question; interactive method which is used to search for

homologous protein or the use of hidden markov model method (HMM) (Altschul et al., 1990, Wheeler et al., 2005, Arnold et al., 2006).

3.1.5 RNA Modeling

After cutting the RNA structure into domains of approximately 500 nucleotides, each domain was modeled separately by the use of the RNA123 software.

The process of modeling using the RNA123 was done by first ensuring that the template and the query were of the same length.

The first step in the modeling was to generate the secondary structure of the template and the query using the RNA123 software. In cases where the template and the query secondary structure were not consistent or the query secondary structure was different from the secondary structure provided. Manual realignment was done for the correct pairing.

This was then followed by the Optimization process after which one had to ensure that there are no mismatches on the sequence between the two sequences i.e. the template and the query.

This was then followed by the modeling process. Different domains modeled were then joined together into one big structure the 16S ribosomal RNA.

3.1.6 Validation of the structure models

Molprobity checks for the flipped atoms in the structure, it checks for presence of the hydrogen if not available it adds them. Molprobity is also able to generate a Ramachadrian plot which shows the protein that fall out of allowed conformational ranges during modeling.(Davis et al., 2007, Chen et al., 2010). Procheck is mostly used to access the general structure of the protein by checking the stereochemistry of the protein structure and root mean square deviation (RMSD)(Laskowski et al., 1996a)

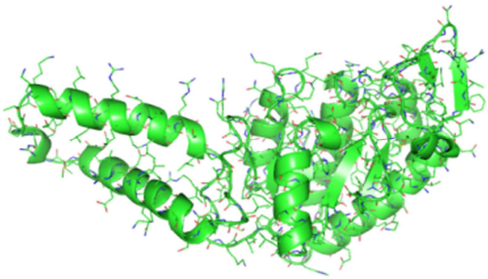
After modeling, the structures acquired had to be checked for structure consistency, correct fold formations and general structure viability. This process was accomplished by running the models through analysis tools such as PROCHECK (Laskowski et al., 1996b) and MOLPROBITY (Davis et al., 2007).

4 CHAPTER FOUR

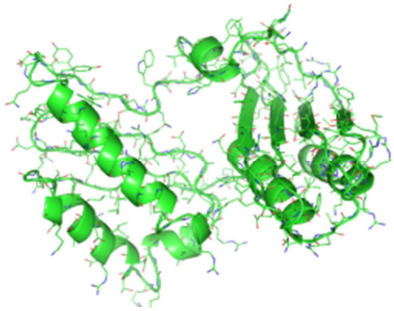
4.1 RESULTS

4.1.1 Protein Results

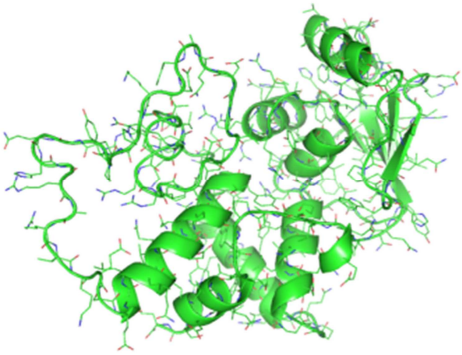
Protein modeling was carried out and 20 proteins were modeled using Swiss model server, pictures of the structures were then taken and labeled accordingly starting with the first protein S1 to S20. The picture shows alpha and the beta sheets folding of the protein. The structure modeled were then used in making of the 30S ribosomal subunit of the *M. tuberculosis*



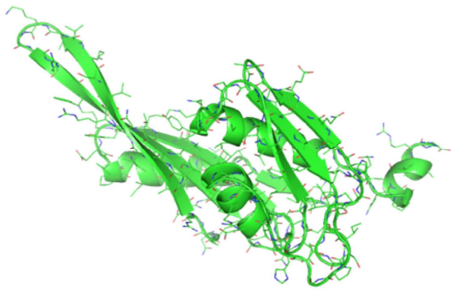
H37Rv_s2_protein



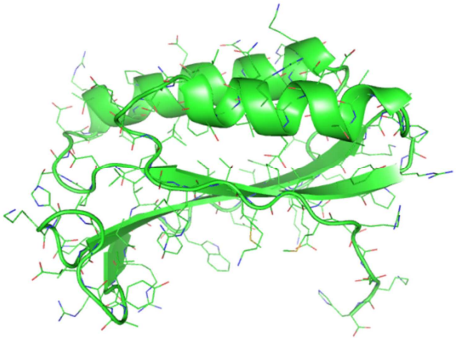
H37Rv_s3_protein



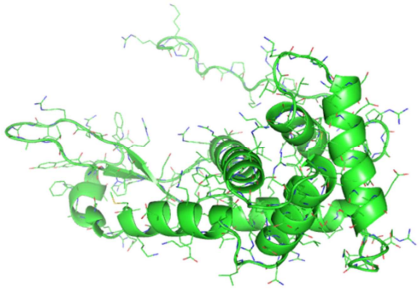
H37Rv_s4_protein



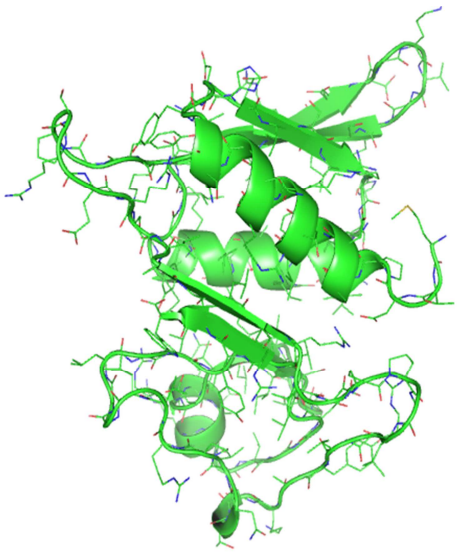
H37Rv_s5_protein



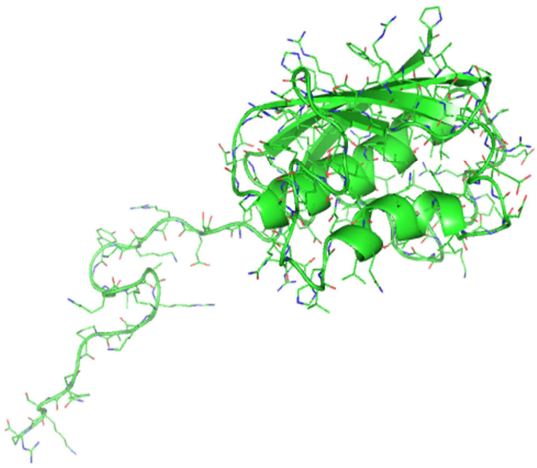
H37Rv-s6_protein



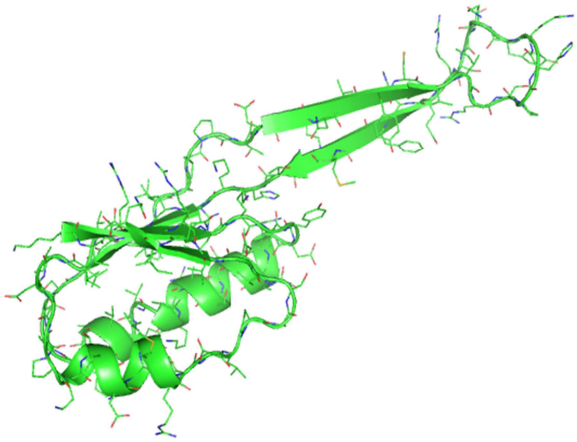
H37Rv_s7_protein



H37Rv_s8_protein

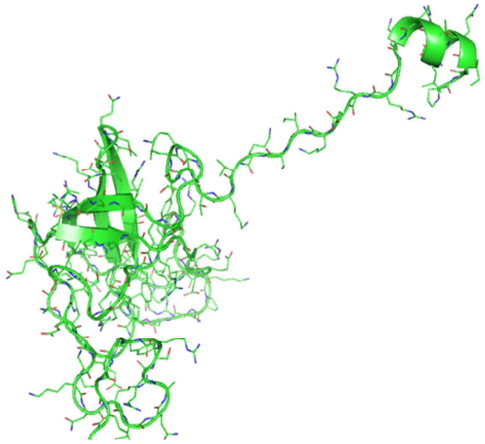


H37Rv_s9_protein

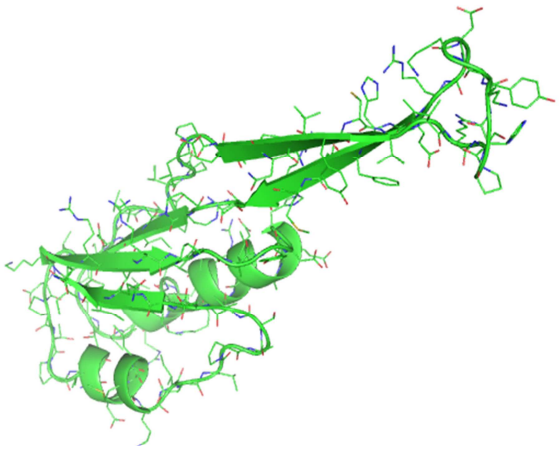


H37Rv_s10_protein

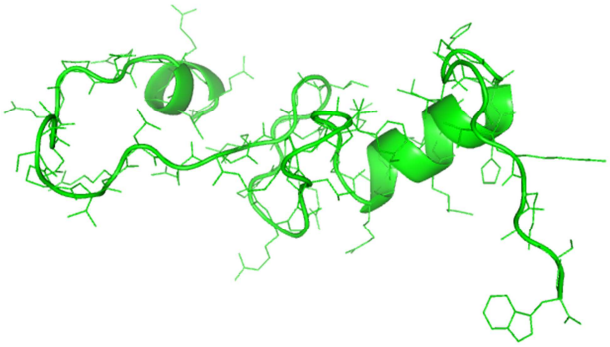
H37Rv_s11_protein



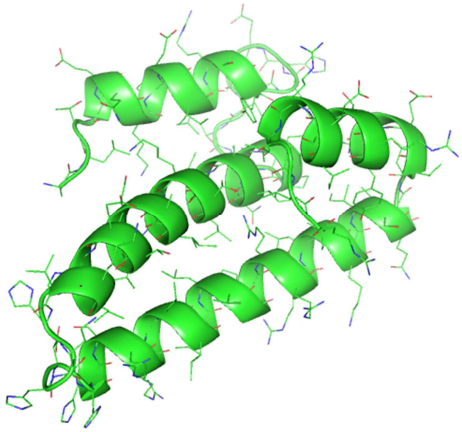
H37Rv_s12_protein



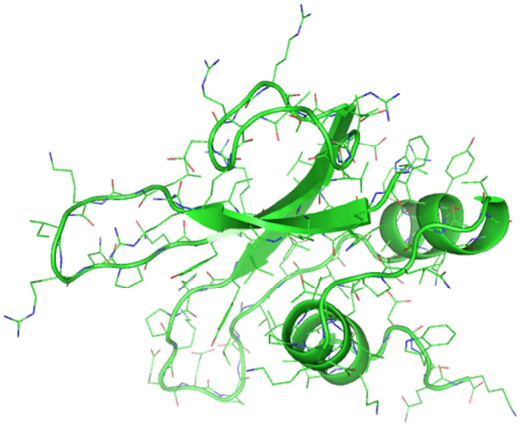
H37Rv_s13_protein



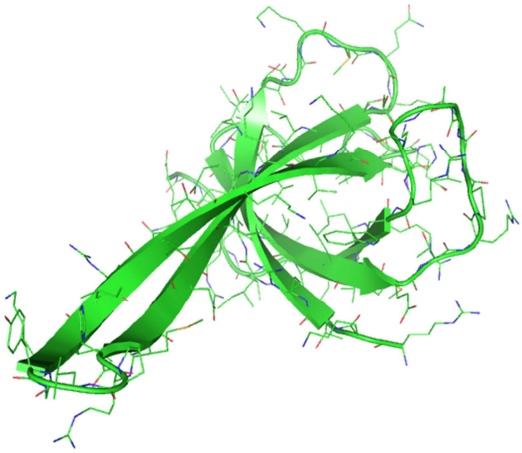
H37Rv_s14_protein



H37Rv_s15_protein

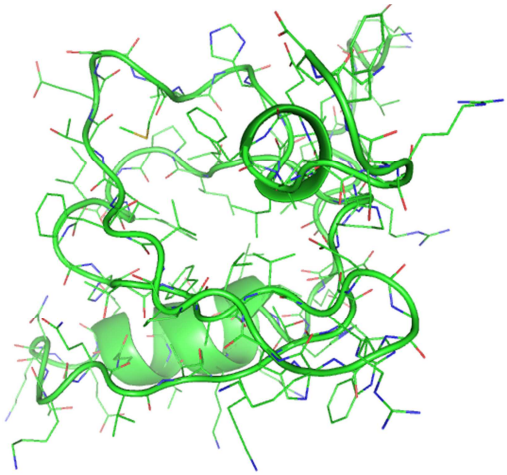


H37Rv_s16_protein

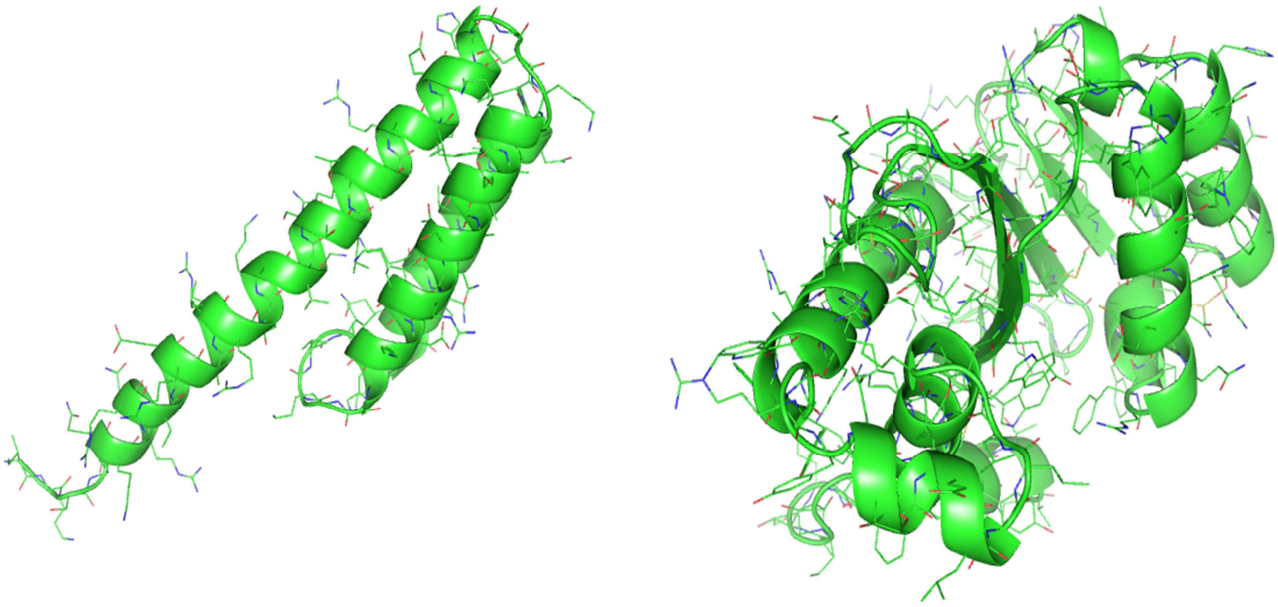


H37Rv_s17_protein

H37Rv_s18_protein



H37Rv_s19_protein



H37Rv_s20_protein

H37Rv_s21_protein

Figure 4:1 S2 to S21 protein found in the small (30S) ribosomal subunit of H37Rv *Mycobacteria tuberculosis* ribosome.

Table 2 showing the Z-score of the protein modeled and the root mean score deviation of the proteins

Number	Z-score	Number of equivalent residue	Root mean square deviation
S2	2.1	44	6.3
S3	33.6	205	0.4
	5.3	66	2.4
	4.8	72	2.7
S4	33.3	200	1.0
S5	28.4	148	0.1
	6.0	57	2.1
	5.9	57	2.1
S6	21.0	96	0.1
	5.3	71	3.3
	5.1	69	2.9
S7	27.7	155	0.1
S8	26.2	130	0.3
S9	23.0	126	0.4
S10	17.5	97	0.1
S11	24.5	116	0.1
S12	21.4	122`	0.1
S13	0.1	38	3.4
S14	0.2	135	3.5
S15	18.3	87	0.1
	2.6	49	3.7
	2.6	49	3.6
S16	17.1	80	1.4
S17	15.4	79	0.1
S18	9.6	146	2.4
S19	17.8	78	0.1
S20	14.8	83	0.1
	6.3	75	2.5
	3.4	61	2.3
	2.0	49	3.1

The Z score shows the general conformation of the structure modeled in relation to the original structure, in this case since there is no original structure model of the *M. tuberculosis* it is compared with the structure of the *E. coli* hence different variation. The room mean square deviation of the structure is also derived by comparing the model structure with the template.

4.1.2 RNA Results

16S rRNA of the *M. tuberculosis* was modeled in domains and then joined together. The different domain structure image have been given starting with the 5 prime , central, 3 major and the 3 minor domains. The whole structure of the RNA is also given. The table of the energy minimization process of the RNA structure is also given with analysis of the most shifted atoms

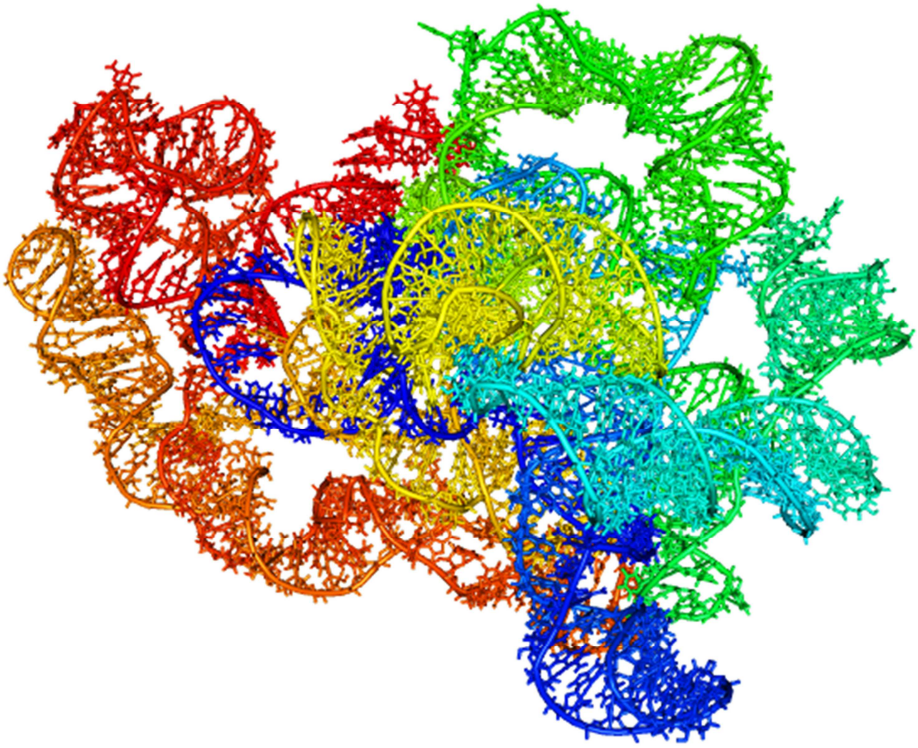


Figure 4:2 H37Rv_16sRNA_5prime_Domain

Figure 4:3 H37Rv_16sRNA_Central_Domain

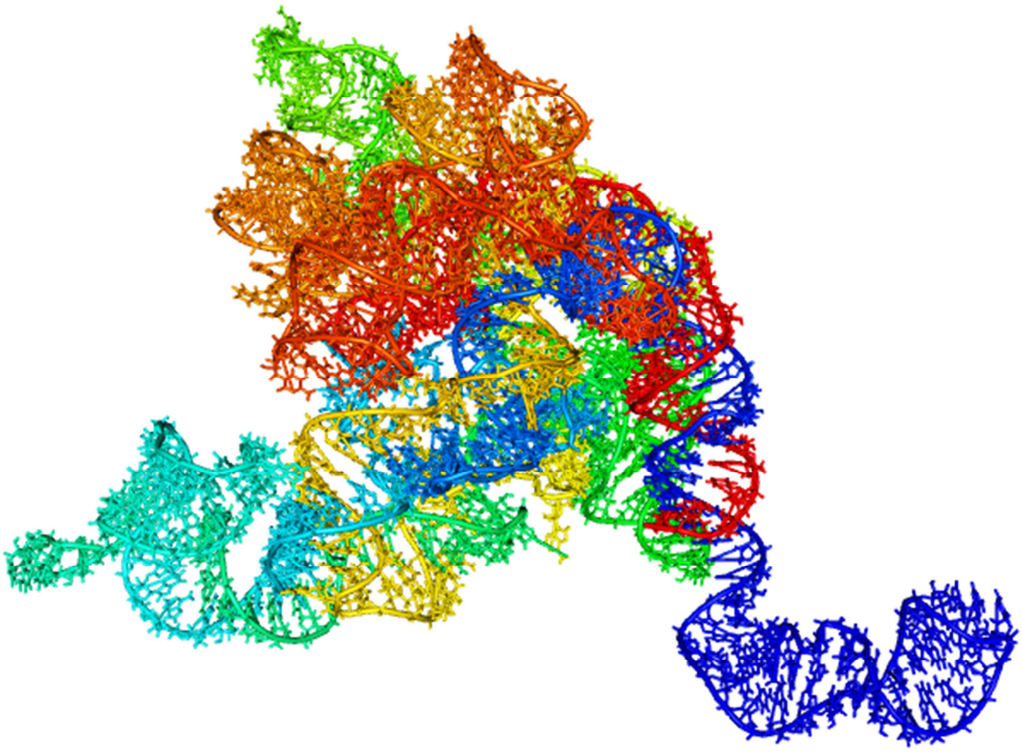


Figure 4:4 H37Rv_16SRNA_3Major_Domain

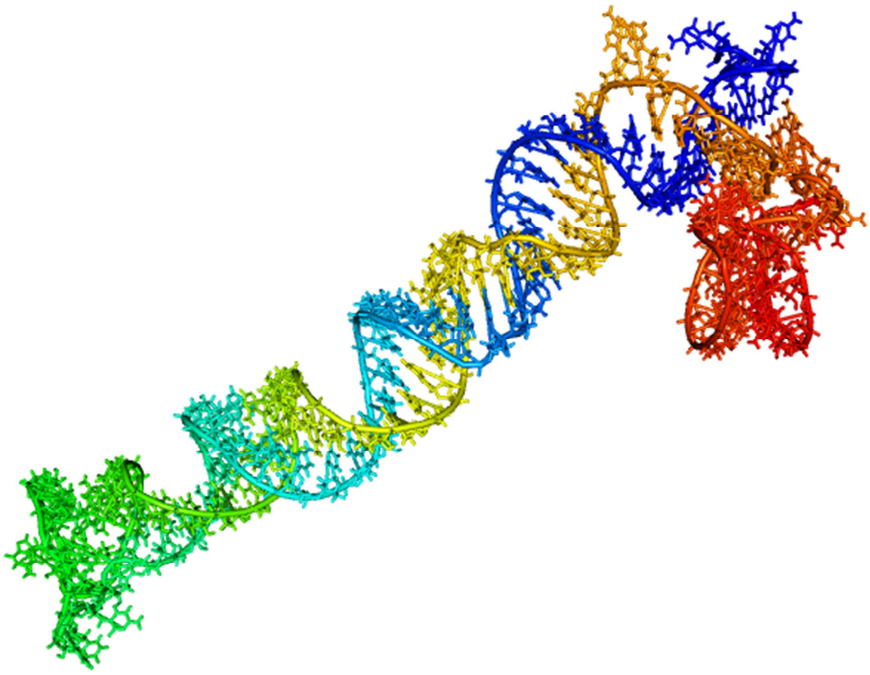


Figure 4:5 H37Rv_16SRNA_3Minor_Domain

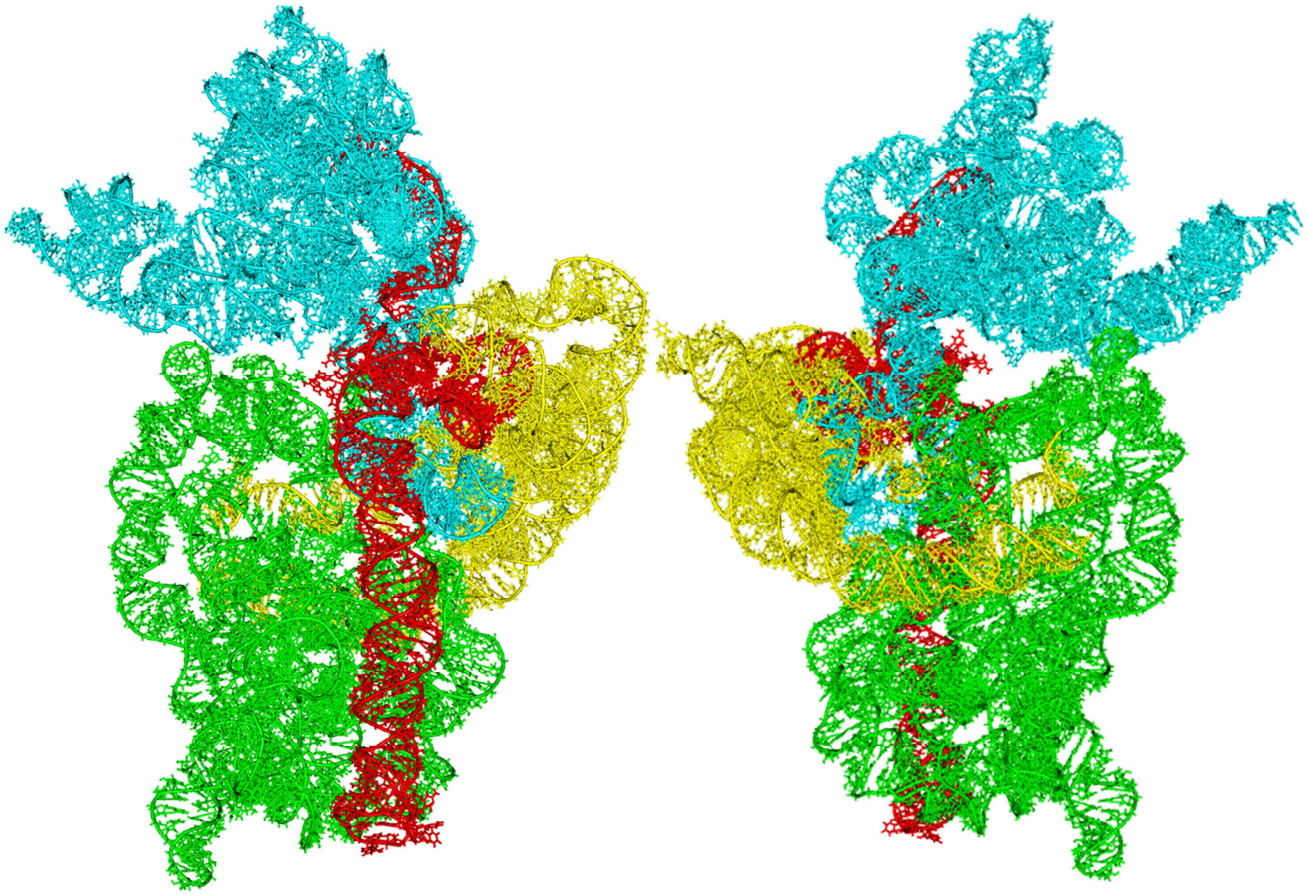


Figure 4:6 showing the combined H37Rv_16S_ribosomal RNA. Different domains are colored differently for identification. The Green color represents the 5prime domain, the Yellow part is the central domain, the 3mojor domain is represented by the cyan color and the Red color represents the three minor. Image on the left represents the right sided image and the image in the right represents the left sided image.

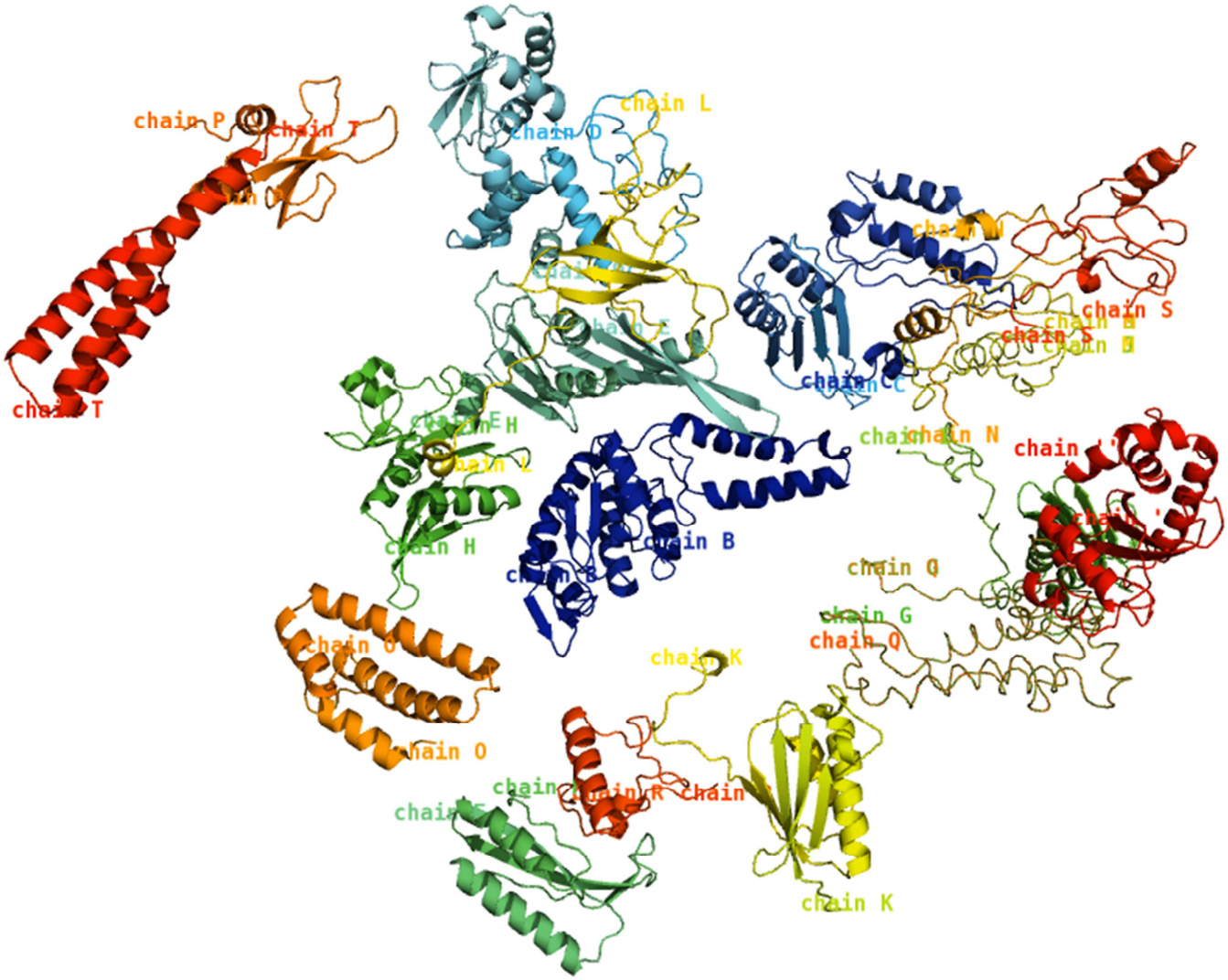


Figure 4:7 20 ribosomal protein of the H37Rv *M.tuberculosis* ready for the H37Rv 16S ribosomal RNA. The proteins are named according to the chain

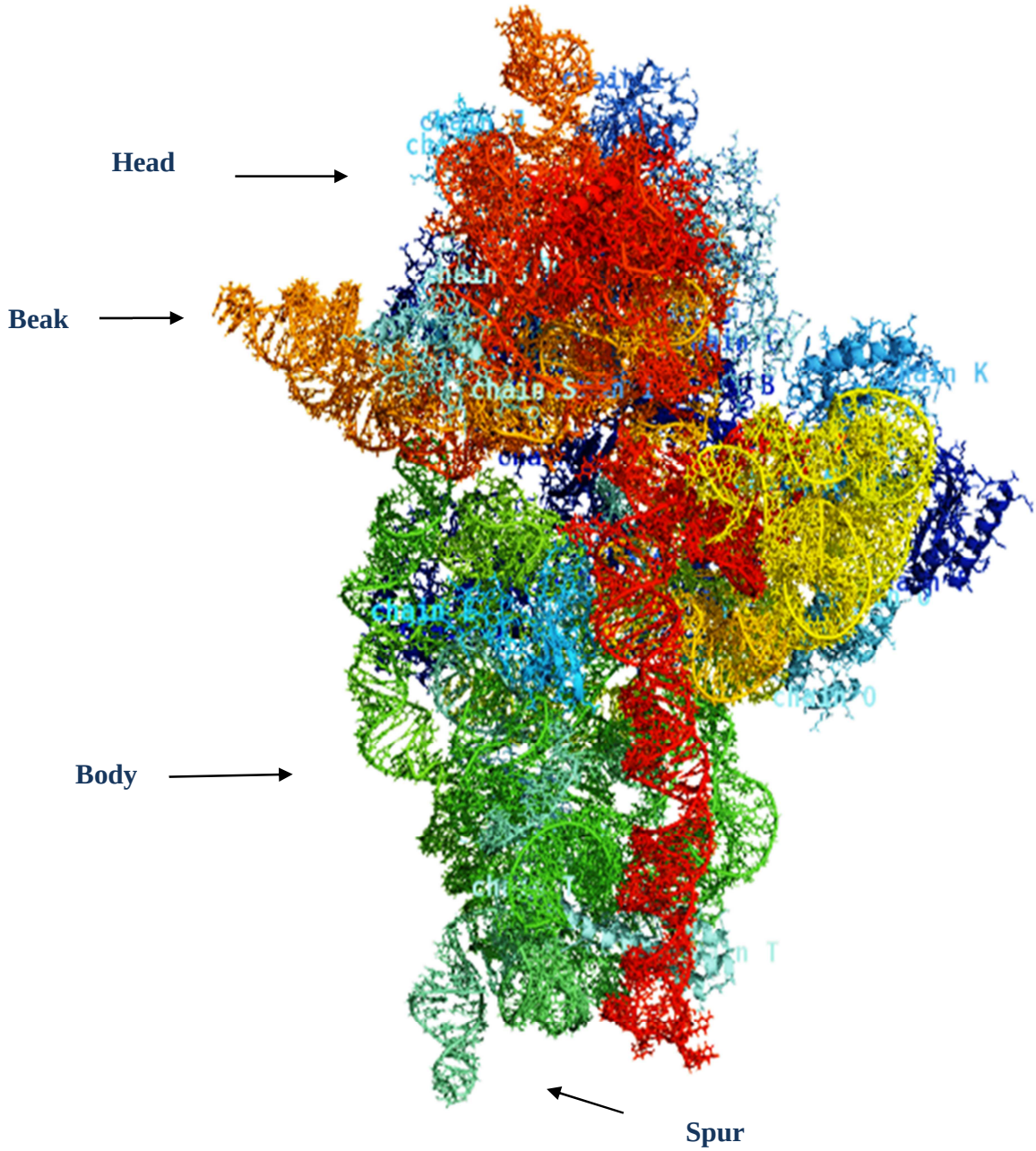


Figure 4:8 Image showing different parts of the 30s ribosome subunit of the *Mycobacteria tuberculosis;* The Head, Beak, Body and the Spur. The parts are consistent with other 30s ribosomal subunit of other prokaryotes

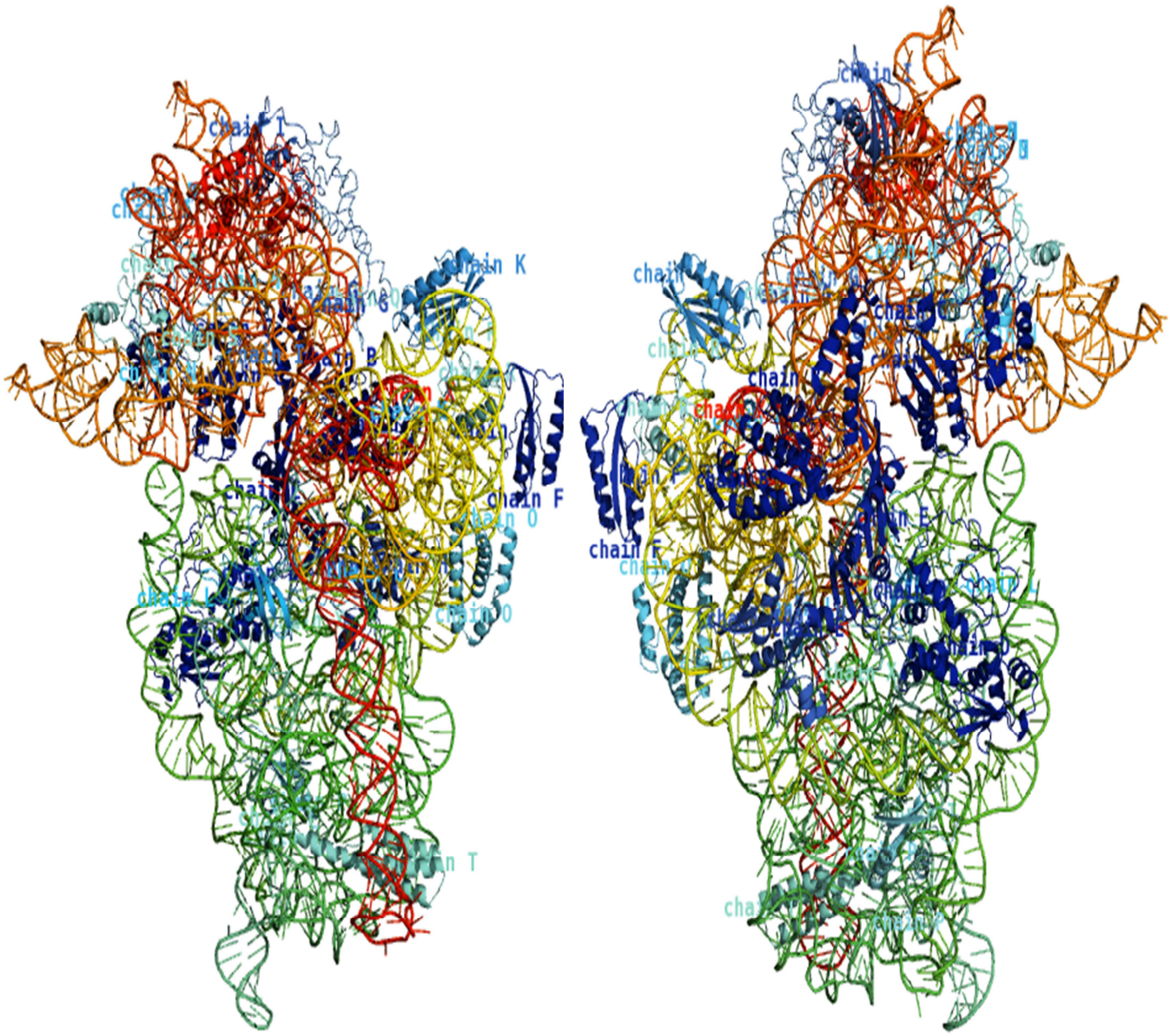


Figure 4:9 H37Rv_30s_ribosomal sub unit. The 16S rRNA is in cyan, and the protein colored different attached to the ribosome. Image on the left is the front side of the unit and the picture on the right is the rear side of the image.

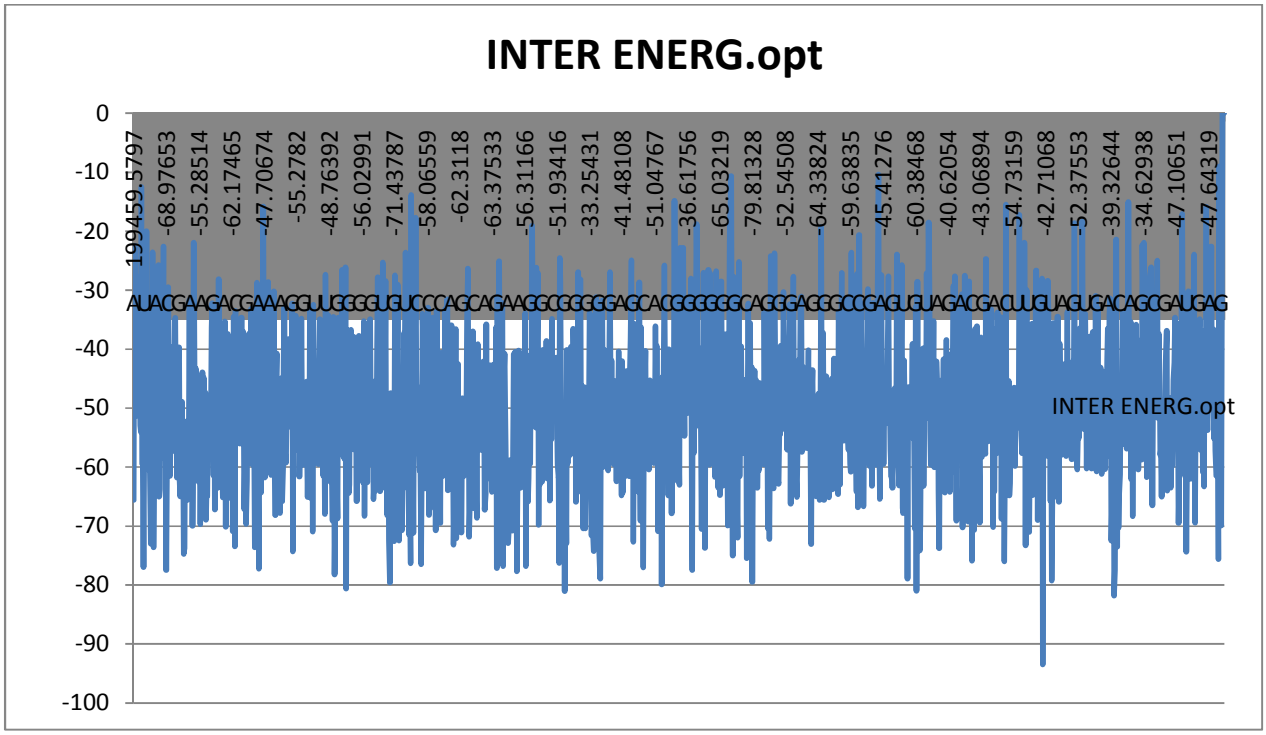


Figure 4:10 graph of the shift of inter energy for the optimized ribosomal structure

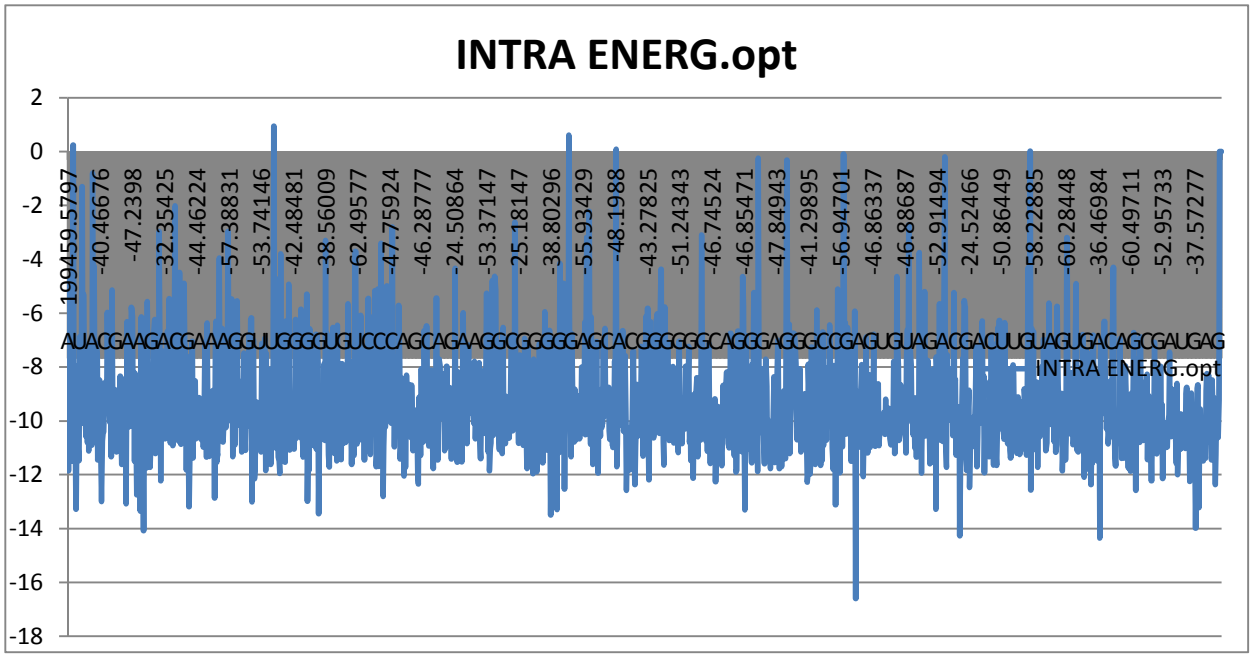


Figure 4:11 graph of the shift of energy in the intra energy for the optimized ribosomal structure

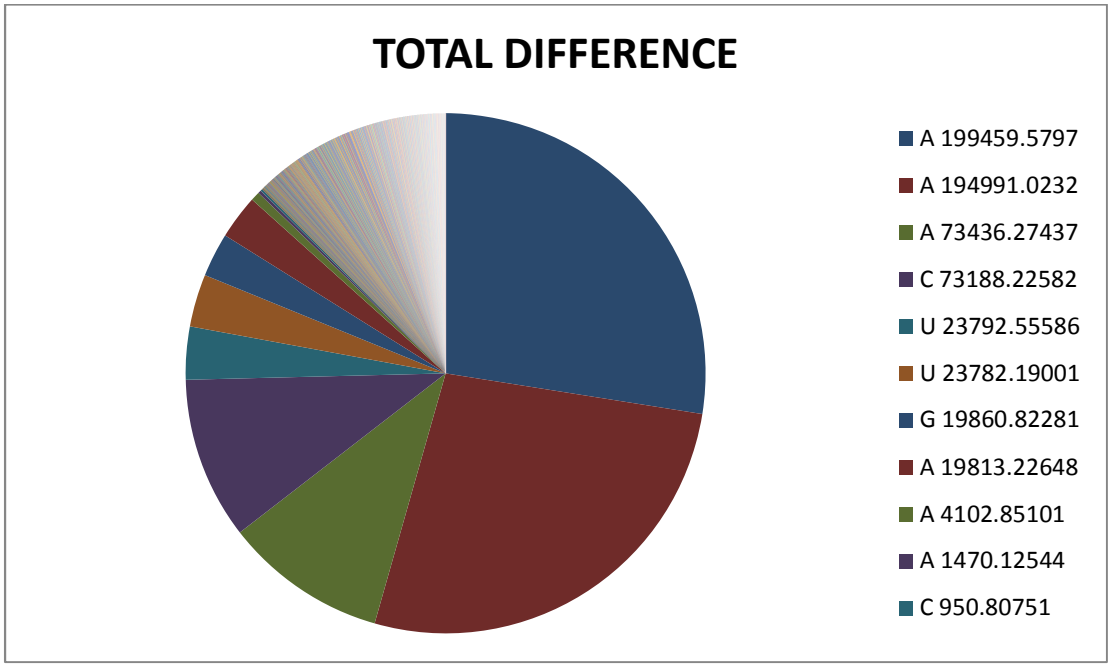


Figure 4:12 analysis of the most shifted molecules in the structure during the process of energy minimization

Table 3 The order of the most shifted atoms and their position in the 30S ribosomal unit

Order of most shifted	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Atoms	**A**	**A**	**A**	**C**	**U**	**U**	**G**	**A**	**A**	**C**
Position of the atom	**1389**	**1390**	**1495**	**870**	**546**	**1523**	**799**	**1392**	**1496**	**912**

5 CHAPTER FIVE

5.1 DISCUSSION

Validation of homology modeling method was done by first conducting a proof of concept. Whereby, X-ray crystallography structures of known prokaryotes were modeled using homology modeling method. The results obtained were evaluated to test and validate the method to be used in this work. The template structure used in the proof of concept had a 3Å (angstrom) and the result of the modeled structure had a root mean square deviation (RMSD) of an average 4Å (angstrom) in the four domains.

RMSD is a measure obtained after a process that matches atoms of the original structure to the atoms of the modeled structure and finds the average distance that all molecules of the model differ from the template. A low RMSD indicates that the two structures are pretty close to each other since the deviation from the template is minimal, whereas a very high RMSD indicates the model deviates from the template with a larger extent. Hence RMSD of 4Å in the proof of concept suggests that the structures modeled by the homology modeling technique are highly similar to the structures modeled by the primary method X-ray crystallography. Results during proof of concept show the global score of the whole ribosome model was close to the original template structure. When the two models structures were compared; the model by homology modeling and the X-ray crystallography model the results indicates that they are similar to each other.

In the proof of concept the 30S ribosomal subunit of *Escherichia coli* was used as the query while that of Thermus *thermophilus* was used as the template structure. This was so because, both the template and the query are prokaryotes and they both have a model structure through X-ray crystallography a primary method of modeling. In addition, their 16S ribosomal RNA's secondary structure and are both similar. In the actual project the choice of the *E. coli* as a template for modeling *Mycobacteria tuberculosis* was influenced by the fact that the secondary structures of the two were similar to each other compared to *T. thermophilus*

Since the method used in the actual project is the same as the one used in the proof of concept the problems encountered in the proof of concept are similar to the ones found in the actual work, therefore we have confidence in the structure obtained (30S ribosomal subunit of H37Rv

Mycobacteria tuberculosis) through homology modeling process as the final result of the modeling.

Throughout the process some mismatches were shown to affect the modeling in great length. In some instances one miss match can lead to inclusion of different coordinates in the structure hence leading to a distorted final structure. In addition gaps in the sequence affected the final structure and this was mostly lethal when the gap falls in an area where there are hairpins, manual realignment had to be done to reduce the gap and address the problem. A gap of more than two nucleotides is dangerous in the process regardless of where it falls, if ignored it leads to bigger gaps between the atoms in the final structure, and this leads to a broken structure.

The results obtained from the protein modeling process are pleasing, checking at the proof of concept where the result obtained are very similar to the original with some having an RMSD of as less as 1 Å (angstrom), which shows how similar the two structures are to each other. The 30S ribosomal sub unit has 21 proteins according to the template used in the study and the query that we were able to model has got 20 proteins. The protein just like the RNA seems to have similar folding although with minor differences which implies that they have the same function as they are in the same family. Table 2 shows the RMSD of the proteins that were obtained and are within the acceptable range; in addition the Z-score is within the acceptable range too. The naming of the query proteins from the first protein to the last protein the S1 to S21 shows that the protein might be having the same function to the template and so does the folding of the protein indicate as the protein functions is more to its folding than it is to the sequence.

Joining of the whole ribosomal structure is perfect in that the protein attaches in the right places just like in the template structure this is usually controlled by the energy they have. Once a PDB file of the protein structure was made, different protein tend to take different conformation space a gap in the middle was created where 16S ribosomal RNA fits in. The ribosomal RNA and the proteins join together to form a 30S ribosomal subunit structure showing the different part of the ribosome, (the head, beak, body and the spur). Just like a structure obtained by the primary method such as X-ray crystallography the results obtained of the 30S subunit by homology modeling shows the different grooves of the structure, reserved areas and also the different folding. During the process of energy minimization it was evident the way the structure energy changed from positive to the negative as given in the table that the structure had more biological

conformation. The process ensured a good structure that is compact and more biologically acceptable.

In the process of joining the whole structure it was noted that there was a way that proteins and the ribosomal RNA of the same species interact. When a similar protein from a different organism was added to the structure, this resulted to a gap between the protein and the other 30S ribosomal subunit, through this energy minimization process we have confidence on the structure as it is compact and has a good biological conformation.

Throughout the project work there were issues that came up especially during the process of RNA modeling; fast the computation power for the machine used in the modeling process is very crucial if one wants to get a successful model, in our case some of the machine used (core 2 due, 2.2, 2.2. and 4Gb Ram) took up to 72 hours to model one domain while others with higher specs (core I3intel 4Gb) took 18 hours to do the same task. During the modeling process it was noted that the software *RNA 123* could not start working on any domain with sequence starting with a Uracil (U) and so one more nucleotide had to be added or it being removed in the starting position of a given domain. During modeling of a domain like the central some atoms had big gap which lead to a distorted structural domain. This was as a result of poorly generated secondary structure which had to be manually corrected so as to get a good structure.

5.2 CONCLUSION

30S ribosomal subunit structure was modeled successfully using homology and *de novo* modeling method. Through the proof concept the structure is shown to be consistent with the structure by the primary method of modeling. Due to the rising need of Insilco three dimensional structure of the protein and the RNAs so as to bridge the knowledge gap, homology modeling is the way to go, this is because the method is fast, reliable and the result gotten are similar to the one obtained by primary method of modeling.

5.3 RECOMMENDATION

Due to the rising need of three dimensional structure of protein, RNA and some organelle, homology modeling is the best method to be used as its first and efficient. But for this to be achieved better machine and servers has to be in place as the software involved require high computational power. In addition there has to be repository for the result generated by the homology method for tracking changes and also to be used in the improving the method.

6 CHAPTER SIX

6.1 REFERENCES

ADURI, R., PSCIUK, B. T., SARO, P., TANIGA, H., SCHLEGEL, H. B. & SANTALUCIA JR, J. 2007. AMBER force field parameters for the naturally occurring modified nucleosides in RNA. *Journal of Chemical Theory and Computation,* 3**,** 1464-1475.

AKBERGENOV, R., SHCHERBAKOV, D., MATT, T., DUSCHA, S., MEYER, M., WILSON, D. N. & BÖTTGER, E. C. 2011. Molecular basis for the selectivity of antituberculosis compounds capreomycin and viomycin. *Antimicrobial agents and chemotherapy,* 55**,** 4712-4717.

ALTSCHUL, S. F. & GISH, W. 1996. Local alignment statistics. *Methods Enzymol,* 266**,** 460-80.

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

AMABILE-CUEVAS, C. F. 2012. Antibiotic Resistance: From Darwin to Lederberg to Keynes. *Microb Drug Resist*.

ARNOLD, K., BORDOLI, L., KOPP, J. & SCHWEDE, T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics,* 22**,** 195-201.

BASHAN, A. & YONATH, A. 2008. The linkage between ribosomal crystallography, metal ions, heteropolytungstates and functional flexibility. *J Mol Struct,* 890**,** 289-294.

BEN-SHEM, A., GARREAU DE LOUBRESSE, N., MELNIKOV, S., JENNER, L., YUSUPOVA, G. & YUSUPOV, M. 2011. The structure of the eukaryotic ribosome at 3.0 A resolution. *Science,* 334**,** 1524-9.

CANNONE, J. J., SUBRAMANIAN, S., SCHNARE, M. N., COLLETT, J. R., D'SOUZA, L. M., DU, Y., FENG, B., LIN, N., MADABUSI, L. V. & MÜLLER, K. M. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics,* 3**,** 2.

CHEN, V. B., ARENDALL, W. B., 3RD, HEADD, J. J., KEEDY, D. A., IMMORMINO, R. M., KAPRAL, G. J., MURRAY, L. W., RICHARDSON, J. S. & RICHARDSON, D. C. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr,* 66**,** 12-21.

COLE, S. T., BROSCH, R., PARKHILL, J., GARNIER, T., CHURCHER, C., HARRIS, D., GORDON, S. V., EIGLMEIER, K., GAS, S., BARRY, C. E., 3RD, TEKAIA, F., BADCOCK, K., BASHAM, D., BROWN, D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R., DEVLIN, K., FELTWELL, T., GENTLES, S., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., KROGH, A., MCLEAN, J., MOULE, S., MURPHY, L., OLIVER, K., OSBORNE, J., QUAIL, M. A., RAJANDREAM, M. A., ROGERS, J., RUTTER, S., SEEGER, K., SKELTON, J., SQUARES, R., SQUARES, S., SULSTON, J. E., TAYLOR, K., WHITEHEAD, S. & BARRELL, B. G. 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature,* 393**,** 537-44.

CYMERMAN, I., FEDER, M., PAWŁOWSKI, M., KUROWSKI, M. & BUJNICKI, J. 2008. Computational methods for protein structure prediction and fold recognition. *Practical bioinformatics***,** 1-21.

DAVIS, I. W., LEAVER-FAY, A., CHEN, V. B., BLOCK, J. N., KAPRAL, G. J., WANG, X., MURRAY, L. W., ARENDALL III, W. B., SNOEYINK, J. & RICHARDSON, J. S. 2007. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research,* 35**,** W375-W383.

DE BACKER, A. I., MORTELE, K. J., DE KEULENAER, B. L. & PARIZEL, P. M. 2006. Tuberculosis: epidemiology, manifestations, and the value of medical imaging in diagnosis. *JBR-BTR,* 89**,** 243-50.

DYE, C. 2000. Tuberculosis 2000-2010: control, but not elimination The Comstock Lecture. *The International Journal of Tuberculosis and Lung Disease,* 4**,** S146-S152.
DYE, C. 2006. Global epidemiology of tuberculosis. *Lancet,* 367**,** 938-40.
ENGER, E. D., ROSS, F. C. & BAILEY, D. B. 2003. *Concepts in biology*, McGraw-Hill.
GARDNER, D. P., REN, P., OZER, S. & GUTELL, R. R. 2011. Statistical potentials for hairpin and internal loops improve the accuracy of the predicted RNA structure. *J Mol Biol,* 413**,** 473-83.
GE, J., CROSBY, S. D., HEINZ, M. E., BESSLER, M. & MASON, P. J. 2010. SnoRNA microarray analysis reveals changes in H/ACA and C/D RNA levels caused by dyskerin ablation in mouse liver. *Biochem J,* 429**,** 33-41.
GLICKMAN, M. S. & JACOBS, W. R., JR. 2001. Microbial pathogenesis of Mycobacterium tuberculosis: dawn of a discipline. *Cell,* 104**,** 477-85.
GUPTA, A., RAHMAN, R., LI, K. & GRIBSKOV, M. 2012. Identifying complete RNA structural ensembles including pseudoknots. *RNA Biol,* 9**,** 187-99.
HACKENBERG, M., HUANG, P. J., HUANG, C. Y., SHI, B. J., GUSTAFSON, P. & LANGRIDGE, P. 2013. A comprehensive expression profile of microRNAs and other classes of non-coding small RNAs in barley under phosphorous-deficient and -sufficient conditions. *DNA Res,* 20**,** 109-25.
HARRIES, A. 1990. Tuberculosis and human immunodeficiency virus infection in developing countries. *Lancet,* 335**,** 387.
HERMANN, T. & WESTHOF, E. 1998. RNA as a drug target: chemical, modelling, and evolutionary tools. *Current opinion in biotechnology,* 9**,** 66-73.
JONES, D. T., TAYLORT, W. & THORNTON, J. M. 1992. A new approach to protein fold recognition.
KIEFER, F., ARNOLD, K., KUNZLI, M., BORDOLI, L. & SCHWEDE, T. 2009. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res,* 37**,** D387-92.
KIM, H. M., RYOU, S. M., SONG, W. S., SIM, S. H., CHA, C. J., HAN, S. H., HA, N. C., KIM, J. H., BAE, J., CUNNINGHAM, P. R. & LEE, K. 2009. Genetic analysis of the invariant residue G791 in Escherichia coli 16S rRNA implicates RelA in ribosome function. *J Bacteriol,* 191**,** 2042-50.
KIM, V. N. & NAM, J. W. 2006. Genomics of microRNA. *Trends Genet,* 22**,** 165-73.
KOLINSKI, A. & BUJNICKI, J. M. 2005. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins,* 61 Suppl 7**,** 84-90.
KRIEGER, E., NABUURS, S. B. & VRIEND, G. 2003. Homology modeling. *Methods of biochemical analysis,* 44**,** 509-524.
KRUPKIN, M., MATZOV, D., TANG, H., METZ, M., KALAORA, R., BELOUSOFF, M. J., ZIMMERMAN, E., BASHAN, A. & YONATH, A. 2011. A vestige of a prebiotic bonding machine is functioning within the contemporary ribosome. *Philos Trans R Soc Lond B Biol Sci,* 366**,** 2972-8.
LASKOWSKI, R. A., RULLMANN, J. A. C., MACARTHUR, M. W., KAPTEIN, R. & THORNTON, J. M. 1996a. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *Journal of biomolecular NMR,* 8**,** 477-486.
LASKOWSKI, R. A., RULLMANNN, J. A., MACARTHUR, M. W., KAPTEIN, R. & THORNTON, J. M. 1996b. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR,* 8**,** 477-86.
LEGER, P., LARA, E., JAGLA, B., SISMEIRO, O., MANSUROGLU, Z., COPPEE, J. Y., BONNEFOY, E. & BOULOY, M. 2013. Dicer-2- and Piwi-mediated RNA interference in Rift Valley fever virus-infected mosquito cells. *J Virol,* 87**,** 1631-48.
LI, K. B. 2003. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics,* 19**,** 1585-1586.
LIND, K. E., DU, Z., FUJINAGA, K., PETERLIN, B. M. & JAMES, T. L. 2002. Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. *Chemistry & biology,* 9**,** 185-193.

LIU, T., TANG, G. W. & CAPRIOTTI, E. 2011. Comparative modeling: the state of the art and protein drug target structure prediction. *Comb Chem High Throughput Screen,* 14**,** 532-47.
LO CONTE, L., AILEY, B., HUBBARD, T. J., BRENNER, S. E., MURZIN, A. G. & CHOTHIA, C. 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Res,* 28**,** 257-9.
MADER, S. S. & BALDWIN, A. 2007. *Essentials of biology*, McGraw-Hill Higher Education.
MEIER, A., KIRSCHNER, P., BANGE, F. C., VOGEL, U. & BOTTGER, E. C. 1994. Genetic alterations in streptomycin-resistant Mycobacterium tuberculosis: mapping of mutations conferring resistance. *Antimicrob Agents Chemother,* 38**,** 228-33.
MOORE, P. B. 1998. The three-dimensional structure of the ribosome and its components. *Annu Rev Biophys Biomol Struct,* 27**,** 35-58.
MOORE, P. B. 2012. How should we think about the ribosome? *Annu Rev Biophys,* 41**,** 1-19.
MOUNT, D. W. 2007. Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc,* 2007**,** pdb top17.
MURZIN, A. G., BRENNER, S. E., HUBBARD, T. & CHOTHIA, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol,* 247**,** 536-40.
PEARSON, N. D. & PRESCOTT, C. D. 1997. RNA as a drug target. *Chemistry & biology,* 4**,** 409-414.
PIEPER, U., ESWAR, N., DAVIS, F. P., BRABERG, H., MADHUSUDHAN, M. S., ROSSI, A., MARTI-RENOM, M., KARCHIN, R., WEBB, B. M., ERAMIAN, D., SHEN, M.-Y., KELLY, L., MELO, F. & SALI, A. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research,* 34**,** D291-D295.
RAMAKRISHNAN, V. 2002. Ribosome structure and the mechanism of translation. *Cell,* 108**,** 557-72.
SCHMEING, T. M. & RAMAKRISHNAN, V. 2009. What recent ribosome structures have revealed about the mechanism of translation. *Nature,* 461**,** 1234-42.
SCHUWIRTH, B. S., BOROVINSKAYA, M. A., HAU, C. W., ZHANG, W., VILA-SANJURJO, A., HOLTON, J. M. & CATE, J. H. D. 2005. Structures of the bacterial ribosome at 3.5 Å resolution. *Science,* 310**,** 827-834.
SCHWEDE, T., KOPP, J., GUEX, N. & PEITSCH, M. C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res,* 31**,** 3381-5.
SHASMAL, M. & SENGUPTA, J. 2012. Structural diversity in bacterial ribosomes: mycobacterial 70S ribosome structure reveals novel features. *PLoS One,* 7**,** e31742.
SIJENYI, F., SARO, P., OUYANG, Z., DAMM-GANAMET, K., WOOD, M., JIANG, J. & SANTALUCIA, J. 2012a. The RNA Folding Problems: Different Levels of sRNA Structure Prediction. *RNA 3D structure analysis and prediction***,** 91-117.
SIJENYI, F., SARO, P., OUYANG, Z., DAMM-GANAMET, K., WOOD, M., JIANG, J. & SANTALUCIA, J. 2012b. The RNA Folding Problems: Different Levels of sRNA Structure Prediction. 27**,** 91-117.
TIPPER, D. J. & STROMINGER, J. L. 1965. Mechanism of action of penicillins: a proposal based on their structural similarity to acyl-D-alanyl-D-alanine. *Proceedings of the National Academy of Sciences of the United States of America,* 54**,** 1133.
WALKER, J. M. 1996. *The protein protocols handbook*, Humana PressInc.
WALKER, J. M. 2005. *The proteomics protocols handbook*, Humana Press.
WALSH, C. 2003. Where will new antibiotics come from? *Nature Reviews Microbiology,* 1**,** 65-69.
WESTBROOK, J., FENG, Z., CHEN, L., YANG, H. & BERMAN, H. M. 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Res,* 31**,** 489-91.
WHEELER, D. L., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHURCH, D. M., DICUCCIO, M., EDGAR, R., FEDERHEN, S., HELMBERG, W., KENTON, D. L., KHOVAYKO, O., LIPMAN, D. J., MADDEN, T. L., MAGLOTT, D. R., OSTELL, J., PONTIUS, J. U., PRUITT, K. D., SCHULER, G. D., SCHRIML, L. M., SEQUEIRA, E., SHERRY, S. T., SIROTKIN, K., STARCHENKO, G., SUZEK, T. O., TATUSOV, R., TATUSOVA, T. A., WAGNER, L. & YASCHENKO, E. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 33**,** D39-45.

YONATH, A. 2005. Antibiotics targeting ribosomes: resistance, selectivity, synergism and cellular regulation. *Annu Rev Biochem,* 74**,** 649-79.
YONATH, A. 2010. Polar bears, antibiotics, and the evolving ribosome (Nobel Lecture). *Angew Chem Int Ed Engl,* 49**,** 4341-54.
ZUO, Z., PENG, D., YIN, X., ZHOU, X., CHENG, H. & ZHOU, R. 2013. Genome-wide analysis reveals origin of transfer RNA genes from tRNA halves. *Mol Biol Evol*.

7 CHAPTER SEVEN

7.1 APPENDICES

7.1.1 Query Sequences Used (H37Rv)

>gi|343206234|ref|NR_044826.1| Mycobacterium tuberculosis strain NCTC 7416 H37Rv 16S ribosomal RNA, complete sequence

UUUGUUUGGAGAGUUUGAUCCUGGCUCAGGACGAACGCUGGCGGCGUGCUUAAC
ACAUGCAAGUCGAACGGAAAGGUCUCUUCGGAGAUACUCGAGUGGCGAACGGGU
GAGUAACACGUGGGUGAUCUGCCCUGCACUUCGGGAUAAGCCUGGGAAACUGGG
UCUAAUACCGGAUAGGACCACGGGAUGCAUGUCUUGUGGUGGAAAGCGCUUUAG
CGGUGUGGGAUGAGCCCGCGGCCUAUCAGCUUGUUGGUGGGGUGACGGCCUACCA
AGGCGACGACGGGUAGCCGGCCUGAGAGGGUGUCCGGCCACACUGGGACUGAGAU
ACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUAUUGCACAAUGGGCGCAA
GCCUGAUGCAGCGACGCCGCGUGGGGGAUGACGGCCUUCGGGUUGUAAACCUCUU
UCACCAUCGACGAAGGUCCGGGUUCUCUCGGAUUGACGGUAGGUGGAGAAGAAG
CACCGGCCAACUACGUGCCAGCAGCCGCGGUAAUACGUAGGGUGCGAGCGUUGUC
CGGAAUUACUGGGCGUAAAGAGCUCGUAGGUGGUUUGUCGCGUUGUUCGUGAAA
UCUCACGGCUUAACUGUGAGCGUGCGGGCGAUACGGGCAGACUAGAGUACUGCA
GGGGAGACUGGAAUUCCUGGUGUAGCGGUGGAAUGCGCAGAUAUCAGGAGGAAC
ACCGGUGGCGAAGGCGGGUCUCUGGGCAGUAACUGACGCUGAGGAGCGAAAGCG
UGGGGAGCGAACAGGAUUAGAUACCCUGGUAGUCCACGCCGUAAACGGUGGGUA
CUAGGUGUGGGUUUCCUUCCUUGGGAUCCGUGCCGUAGCUAACGCAUUAAGUACC
CCGCCUGGGGAGUACGGCCGCAAGGCUAAAACUCAAAGGAAUUGACGGGGGCCCG
CACAAGCGGCGGAGCAUGUGGAUUAAUUCGAUGCAACGCGAAGAACCUUACCUG
GGUUUGACAUGCACAGGACGCGUCUAGAGAUAGGCGUUCCCUUGUGGCCUGUGU
GCAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCC
CGCAACGAGCGCAACCCUUGUCUCAUGUUGCCAGCACGUAAUGGUGGGGACUCGU
GAGAGACUGCCGGGGUCAACUCGGAGGAAGGUGGGGAUGACGUCAAGUCAUCAU
GCCCCUUAUGUCCAGGGCUUCACACAUGCUACAAUGGCCGGUACAAAGGGCUGCG

AUGCCGCGAGGUUAAGCGAAUCCUUAAAAGCCGGUCUCAGUUCGGAUCGGGGUC
UGCAACUCGACCCCGUGAAGUCGGAGUCGCUAGUAAUCGCAGAUCAGCAACGCUG
CGGUGAAUACGUUCCCGGGCCUUGUACACACCGCCCGUCACGUCAUGAAAGUCGG
UAACACCCGAAGCCAGUGGCCUAACCCUCGGGAGGGAGCUGUCGAAGGUGGGAUC
GGCGAUUGGGACGAAGUCGUAACAAGGUAGCCGUACCGGAAGGUGCGGCUGGAU
CACCUCCUUUCU

>gi|15610027|ref|NP_217406.1| 30S ribosomal protein S2 [Mycobacterium tuberculosis H37Rv]

MAVVTMKQLLDSGTHFGHQTRRWNPKMKRFIFTDRNGIYIIDLQQTLTFIDKAYEFVKE
TVAHGGSVLFVGTKKQAQESVAAEATRVGMPYVNQRWLGGMLTNFSTVHKRLQRLK
ELEAMEQTGGFEGRTKKEILGLTREKNKLERSLGGIRDMAKVPSAIWVVDTNKEHIAVG
EARKLGIPVIAILDTNCDPDEVDYPIPGNDDAIRSAALLTRVIASAVAEGLQARAGLGRA
DGKPEAEAAEPLAEWEQELLASATASATPSATASTTALTDAPAGATEPTTDAS

>gi|15607847|ref|NP_215221.1| 30S ribosomal protein S3 [Mycobacterium tuberculosis H37Rv]

MGQKINPHGFRLGITTDWKSRWYADKQYAEYVKEDVAIRRLLSSGLERAGIADVEIERT
RDRVRVDIHTARPGIVIGRRGTEADRIRADLEKLTGKQVQLNILEVKNPESQAQLVAQG
VAEQLSNRVAFRRAMRKAIQSAMRQPNVKGIRVQCSGRLGGAEMSRSEFYREGRVPLH
TLRADIDYGLYEAKTTFGRIGVKVWIYKGDIVGGKRELAAAAPAGADRPRRERPSGTRP
RRSGASGTTATGTDAGRAAGGEEAAPDAAAPVEAQSTE

>gi|15610594|ref|NP_217975.1| 30S ribosomal protein S4 [Mycobacterium tuberculosis H37Rv]

MARYTGPVTRKSRRLRTDLVGGDQAFEKRPYPPGQHGRARIKESEYLLQLQEKQKARF
TYGVMEKQFRRYYEEAVRQPGKTGEELLKILESRLDNVIYRAGLARTRRMARQLVSHG
HFNVNGVHVNVPSYRVSQYDIVDVRDKSLNTVPFQIARETAGERPIPSWLQVVGERQRV
LIHQLPERAQIDVPLTEQLIVEYYSK

>gi|15607861|ref|NP_215235.1| 30S ribosomal protein S5 [Mycobacterium tuberculosis H37Rv]

MAEQPAGQAGTTDNRDARGDREGRRRDSGRGSRERDGEKSNYLERVVAINRVSKVVK
GGRRFSFTALVIVGDGNGMVGVGYGKAKEVPAAIAKGVEEARKSFFRVPLIGGTITHPV

QGEAAAGVVLLRPASPGTGVIAGGAARAVLECAGVHDILAKSLGSDNAINVVHATVAA
LKLLQRPEEVAARRGLPIEDVAPAGMLKARRKSEALAASVLPDRTI

gi|15607195|ref|NP_214567.1| 30S ribosomal protein S6 [Mycobacterium tuberculosis H37Rv]

MRPYEIMVILDPTLDERTVAPSLETFLNVVRKDGGKVEKVDIWGKRRLAYEIAKHAEGI
YVVIDVKAAPATVSELDRQLSLNESVLRTKVMRTDKH

>gi|15607823|ref|NP_215197.1| 30S ribosomal protein S7 [Mycobacterium tuberculosis H37Rv]

MPRKGPAPKRPLVNDPVYGSQLVTQLVNKVLLKGKKSLAERIVYGALEQARDKTGTDP
VITLKRALDNVKPALEVRSRRVGGATYQVPVEVRPDRSTTLALRWLVGYSRQRREKTM
IERLANEILDASNGLGASVKRREDTHKMAEANRAFAHYRW

>gi|15607858|ref|NP_215232.1| 30S ribosomal protein S8 [Mycobacterium tuberculosis H37Rv]

MTMTDPIADFLTRLRNANSAYHDEVSLPHSKLKANIAQILKNEGYISDFRTEDARVGKSL
VIQLKYGPSRERSIAGLRRVSKPGLRVYAKSTNLPRVLGGLGVAIISTSSGLLTDRQAAR
QGVGGEVLAYVW

>gi|15610578|ref|NP_217959.1| 30S ribosomal protein S9 [Mycobacterium tuberculosis H37Rv]

MTETTPAPQTPAAPAGPAQSFVLERPIQTVGRRKEAVVRVRLVPGTGKFDLNGRSLEDY
FPNKVHQQLIKAPLVTVDRVESFDIFAHLGGGGPSGQAGALRLGIARALILVSPEDRPAL
KKAGFLTRDPRATERKKYGLKKARKAPQYSKR

>gi|15607840|ref|NP_215214.1| 30S ribosomal protein S10 [Mycobacterium tuberculosis H37Rv]

MAGQKIRIRLKAYDHEAIDASARKIVETVVRTGASVVGPVPLPTEKNVYCVIRSPHKYK
DSREHFEMRTHKRLIDIIDPTPKTVDALMRIDLPASVDVNIQ

>gi|15610595|ref|NP_217976.1| 30S ribosomal protein S11 [Mycobacterium tuberculosis H37Rv]

MPPAKKGPATSARKGQKTRRREKKNVPHGAAHIKSTFNNTIVTITDPQGNVIAWASSGH
VGFKGSRKSTPFAAQLAAENAARKAQDHGVRKVDVFVKGPGSGRETAIRSLQAAGLEV
GAISDVTPQPHNGVRPPKRRRV

>gi|15077869|gb|AAK83386.1|AF398880_1 ribosomal protein S12 [Mycobacterium tuberculosis]

MPTIQQLVRKGRRDKISKVKTAALKGSPQRRGVCTRVYTTTPKKPNSALRKVARVKLTS
QVEVTAYIPGEGHNLQEHSMVLVRGGRVQDLPGVRYKIIRGSLDTQGVKNRKQARSRY
GAKKEKG

>gi|15610596|ref|NP_217977.1| 30S ribosomal protein S13 [Mycobacterium tuberculosis H37Rv]

MARLVGVDLPRDKRMEVALTYIFGIGRTRSNEILAATGIDRDLRTRDLTEEQLIHLRDYI
EANLKVEGDLRREVQADIRRKIEIGCYQGLRHRRGMPVRGQRTKTNARTRKGPKRTIAG
KKKAR

>gi|15609193|ref|NP_216572.1| 30S ribosomal protein S14 [Mycobacterium tuberculosis H37Rv]

MAKKSKIVKNQRRAATVARYASRRTALKDIIRSPSSAPEQRSTAQRALARQPRDASPVR
LRNRDAIDGRPRGHLRKFGLSRVRVRQLAHDGHLPGVRKASW

>gi|15609922|ref|NP_217301.1| 30S ribosomal protein S15 [Mycobacterium tuberculosis H37Rv]

MALTAEQKKEILRSYGLHETDTGSPEAQIALLTKRIADLTEHLKVHKHDHHSRRGLLLL
VGRRRRLIKYISQIDVERYRSLIERLGLRR

>gi|15610046|ref|NP_217425.1| 30S ribosomal protein S16 [Mycobacterium tuberculosis H37Rv]

MAVKIKLTRLGKIRNPQYRVAVADARTRRDGRAIEVIGRYHPKEEPSLIEINSERAQYWL
SVGAQPTEPVLKLLKITGDWQKFKGLPGAQGRLKVAAPKPSKLEVFNAALAAADGGPT
TEATKPKKKSPAKKAAKAAEPAPQPEQPDTPALGGEQAELTAES

>gi|15607850|ref|NP_215224.1| 30S ribosomal protein S17 [Mycobacterium tuberculosis H37Rv]

MMAEAKTGAKAAPRVAKAAKAAPKKAAPNDAEAIGAANAANVKGPKHTPRTPKPRG
RRKTRIGYVVSDKMQKTIVVELEDRMRHPLYGKIIRTTKKVKAHDEDSVAGIGDRVSLM
ETRPLSATKRWRLVEILEKAK

>gi|15839432|ref|NP_334469.1| 30S ribosomal protein S18 [Mycobacterium tuberculosis CDC1551]

MAKSSKRRPAPEKPVKTRKCVFCAKKDQAIDYKDTALLRTYISERGKIRARRVTGNCVQ
HQRDIALAVKNAREVALLPFTSSVR

>gi|15607845|ref|NP_215219.1| 30S ribosomal protein S19 [Mycobacterium tuberculosis H37Rv]

MPRSLKKGPFVDEHLLKKVDVQNEKNTKQVIKTWSRRSTIIPDFIGHTFAVHDGRKHVP
VFVTESMVGHKLGEFAPTRTFKGHIKDDRKSKRR

>gi|15609549|ref|NP_216928.1| 30S ribosomal protein S20 [Mycobacterium tuberculosis H37Rv]

MANIKSQQKRNRTNERARLRNKAVKSSLRTAVRAFREAAHAGDKAKAAELLASTNRK
LDKAASKGVIHKNQAANKKSALAQALNKL

7.1.2 Template Sequences Used

Fasta sequence of the E.coli (2AVY)

http://www.rcsb.org/PDB/files/fasta.txt?structureIdList=2AVY

>2AVY:A|PDBID|CHAIN|SEQUENCE

AAAUUGAAGAGUUUGAUCAUGGCUCAGAUUGAACGCUGGCGGCAGGCCUAACAC
AUGCAAGUCGAACGGUAACAGGAAGAAGCUUGCUUCUUUGCUGACGAGUGGCGG
ACGGGUGAGUAAUGUCUGGGAAACUGCCUGAUGGAGGGGGAUAACUACUGGAAA
CGGUAGCUAAUACCGCAUAACGUCGCAAGACCAAAGAGGGGGACCUUCGGGCCUC
UUGCCAUCGGAUGUGCCCAGAUGGGAUUAGCUAGUAGGUGGGGUAACGGCUCAC
CUAGGCGACGAUCCCUAGCUGGUCUGAGAGGAUGACCAGCCACACUGGAACUGAG
ACACGGUCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUAUUGCACAAUGGGCGC
AAGCCUGAUGCAGCCAUGCCGCGUGUAUGAAGAAGGCCUUCGGGUUGUAAAGUA
CUUUCAGCGGGGAGGAAGGGAGUAAAGUUAAUACCUUUGCUCAUUGACGUUACC
CGCAGAAGAAGCACCGGCUAACUCCGUGCCAGCAGCCGCGGUAAUACGGAGGGUG

CAAGCGUUAAUCGGAAUUACUGGGCGUAAAGCGCACGCAGGCGGUUUGUUAAGU
CAGAUGUGAAAUCCCCGGGCUCAACCUGGGAACUGCAUCUGAUACUGGCAAGCUU
GAGUCUCGUAGAGGGGGGUAGAAUUCCAGGUGUAGCGGUGAAAUGCGUAGAGAU
CUGGAGGAAUACCGGUGGCGAAGGCGGCCCCCUGGACGAAGACUGACGCUCAGGU
GCGAAAGCGUGGGGAGCAAACAGGAUUAGAUACCCUGGUAGUCCACGCCGUAAA
CGAUGUCGACUUGGAGGUUGUGCCCUUGAGGCGUGGCUUCCGGAGCUAACGCGU
UAAGUCGACCGCCUGGGGAGUACGGCCGCAAGGUUAAAACUCAAAUGAAUUGAC
GGGGGCCCGCACAAGCGGUGGAGCAUGUGGUUUAAUUCGAUGCAACGCGAAGAA
CCUUACCUGGUCUUGACAUCCACGGAAGUUUUCAGAGAUGAGAAUGUGCCUUCG
GGAACCGUGAGACAGGUGCUGCAUGGCUGUCGUCAGCUCGUGUUGUGAAAUGUU
GGGUUAAGUCCCGCAACGAGCGCAACCCUUAUCCUUUGUUGCCAGCGGUCCGGCC
GGGAACUCAAAGGAGACUGCCAGUGAUAAACUGGAGGAAGGUGGGGAUGACGUC
AAGUCAUCAUGGCCCUUACGACCAGGGCUACACACGUGCUACAAUGGCGCAUACA
AAGAGAAGCGACCUCGCGAGAGCAAGCGGACCUCAUAAAGUGCGUCGUAGUCCGG
AUUGGAGUCUGCAACUCGACUCCAUGAAGUCGGAAUCGCUAGUAAUCGUGGAUC
AGAAUGCCACGGUGAAUACGUUCCCGGGCCUUGUACACACCGCCCGUCACACCAU
GGGAGUGGGUUGCAAAAGAAGUAGGUAGCUUAACCUUCGGGAGGGCGCUUACCA
CUUUGUGAUUCAUGACUGGGGUGAAGUCGUAACAAGGUAACCGUAGGGGAACCU
GCGGUUGGAUCACCUCCUUA

>2AVY:B|PDBID|CHAIN|SEQUENCE

ATVSMRDMLKAGVHFGHQTRYWNPKMKPFIFGARNKVHIINLEKTVPMFNEALAELN
KIASRKGKILFVGTKRAASEAVKDAALSCDQFFVNHRWLGGMLTNWKTVRQSIKRLKD
LETQSQDGTFDKLTKKEALMRTRELEKLENSLGGIKDMGGLPDALFVIDADHEHIAIKE
ANNLGIPVFAIVDTNSDPDGVDFVIPGNDDAIRAVTLYLGAVAATVREGRSQDLASQAE
ESFVEAE

>2AVY:C|PDBID|CHAIN|SEQUENCE

GQKVHPNGIRLGIVKPWNSTWFANTKEFADNLDSDFKVRQYLTKELAKASVSRIVIERP
AKSIRVTIHTARPGIVIGKKGEDVEKLRKVVADIAGVPAQINIAEVRKPELDAKLVADSIT
SQLERRVMFRRAMKRAVQNAMRLGAKGIKVEVSGRLGGAEIARTEWYREGRVPLHTL
RADIDYNTSEAHTTYGVIGVKVWIFKGEILGGMAAVEQPEKPAAQPKKQQRKGRK

>2AVY:D|PDBID|CHAIN|SEQUENCE

ARYLGPKLKLSRREGTDLFLKSGVRAIDTKCKIEQAPGQHGARKPRLSDYGVQLREKQK
VRRIYGVLERQFRNYYKEAARLKGNTGENLLALLEGRLDNVVYRMGFGATRAEARQL
VSHKAIMVNGRVVNIASYQVSPNDVVSIREKAKKQSRVKAALELAEQREKPTWLEVDA
GKMEGTFKRKPERSDLSADINEHLIVELYSK

>**2AVY:E|PDBID|CHAIN|SEQUENCE**

AHIEKQAGELQEKLIAVNRVSKTVKGGRIFSFTALTVVGDGNGRVGFGYGKAREVPAAI
QKAMEKARRNMINVALNNGTLQHPVKGVHTGSRVFMQPASEGTGIIAGGAMRAVLEV
AGVHNVLAKAYGSTNPINVVRATIDGLENMNSPEMVAAKRGKSVEEILGK

>2AVY:F|PDBID|CHAIN|SEQUENCE

MRHYEIVFMVHPDQSEQVPGMIERYTAAITGAEGKIHRLEDWGRRQLAYPINKLHKAH
YVLMNVEAPQEVIDELETTFRFNDAVIRSMVMRTKHAVTEASPMVKAKDERRERRDDF
ANETADDAEAGDSEEEEEE

>2AVY:G|PDBID|CHAIN|SEQUENCE

PRRRVIGQRKILPDPKFGSELLAKFVNILMVDGKKSTAESIVYSALETLAQRSGKSELEAF
EVALENVRPTVEVKSRRVGGSTYQVPVEVRPVRRNALAMRWIVEAARKRGDKSMALR
LANELSDAAENKGTAVKKREDVHRMAEANKAFAHYRWLSLRSFSHQAGASSKQPALG
YLN

>2AVY:H|PDBID|CHAIN|SEQUENCE

SMQDPIADMLTRIRNGQAANKAAVTMPSSKLKVAIANVLKEEGFIEDFKVEGDTKPELE
LTLKYFQGKAVVESIQRVSRPGLRIYKRKDELPKVMAGLGIAVVSTSKGVMTDRAARQ
AGLGGEIICYVA

>2AVY:I|PDBID|CHAIN|SEQUENCE

AENQYYGTGRRKSSAARVFIKPGNGKIVINQRSLEQYFGRETARMVVRQPLELVDMVE
KLDLYITVKGGGISGQAGAIRHGITRALMEYDESLRSELRKAGFVTRDARQVERKKVGL
RKARRRPQFSKR

>2AVY:J|PDBID|CHAIN|SEQUENCE

MQNQRIRIRLKAFDHRLIDQATAEIVETAKRTGAQVRGPIPLPTRKERFTVLISPHVNKDA
RDQYEIRTHLRLVDIVEPTEKTVDALMRLDLAAGVDVQISLG

>2AVY:K|PDBID|CHAIN|SEQUENCE

AKAPIRARKRVRKQVSDGVAHIHASFNNTIVTITDRQGNALGWATAGGSGFRGSRKSTP
FAAQVAAERCADAVKEYGIKNLEVMVKGPGPGRESTIRALNAAGFRITNITDVTPIPHNG
CRPPKKRRV

>2AVY:L|PDBID|CHAIN|SEQUENCE

ATVNQLVRKPRARKVAKSNVPALEACPQKRGVCTRVYTTTPKKPNSALRKVCRVRLTN
GFEVTSYIGGEGHNLQEHSVIL
IRGGRVKDLPGVRYHTVRGALDCSGVKDRKQARSKYGVKRPKA

>2AVY:M|PDBID|CHAIN|SEQUENCE

ARIAGINIPDHKHAVIALTSIYGVGKTRSKAILAAAGIAEDVKISELSEGQIDTLRDEVAKF
VVEGDLRREISMSIKRLMDLGCYRGLRHRRGLPVRGQRTKTNARTRKGPRKPIKK

>2AVY:N|PDBID|CHAIN|SEQUENCE

AKQSMKAREVKRVALADKYFAKRAELKAIISDVNASDEDRWNAVLKLQTLPRDSSPSR
QRNRCRQTGRPHGFLRKFGLSRIKVREAAMRGEIPGLKKASW

>2AVY:O|PDBID|CHAIN|SEQUENCE

MSLSTEATAKIVSEFGRDANDTGSTEVQVALLTAQINHLQGHFAEHKKDHHSRRGLLR
MVSQRRKLLDYLKRKDVARYTRLIERLGLRR

>2AVY:P|PDBID|CHAIN|SEQUENCE

MVTIRLARHGAKKRPFYQVVVADSRNARNGRFIERVGFFNPIASEKEEGTRLDLDRIAH
WVGQGATISDRVAALIKEVNKAA

>2AVY:Q|PDBID|CHAIN|SEQUENCE

TDKIRTLQGRVVSDKMEKSIVVAIERFVKHPIYGKFIKRTTKLHVHDENNECGIGDVVEI
RECRPLSKTKSWTLVRVVEKAVL

>2AVY:R|PDBID|CHAIN|SEQUENCE

ARYFRRRKFCRFTAEGVQEIDYKDIATLKNYITESGKIVPSRITGTRAKYQRQLARAIKR
ARYLSLLPYTDRHQ

>2AVY:S|PDBID|CHAIN|SEQUENCE

PRSLKKGPFIDLHLLKKVEKAVESGDKKPLRTWSRRSTIFPNMIGLTIAVHNGRQHVPVF
VTDEMVGHKLGEFAPTRTYRGHAADKKAKKK

>2AVY:T|PDBID|CHAIN|SEQUENCE

ANIKSAKKRAIQSEKARKHNASRRSMMRTFIKKVYAAIEAGDKAAAQKAFNEMQPIVD
RQAAKGLIHKNKAARHKANLTAQINKLA

>2AVY:U|PDBID|CHAIN|SEQUENCE

MPVIKVRENEPFDVALRRFKRSCEKAGVLAEVRRREFYEKPTTERKRAKASAVKRHAK
KLARENARRTRLY

7.1.2.1 Fasta sequence of T. thermophile

>2J00:A|PDBID|CHAIN|SEQUENCE

UUUGUUGGAGAGUUUGAUCCUGGCUCAGGGUGAACGCUGGCGGCGUGCCUAAGA
CAUGCAAGUCGUGCGGGCCGCGGGGUUUUACUCCGUGGUCAGCGGCGGACGGGUG
AGUAACGCGUGGGUGACCUACCCGGAAGAGGGGGACAACCCGGGGAAACUCGGGC
UAAUCCCCCAUGUGGACCCGCCCCUUGGGGUGUGUCCAAAGGGCUUUGCCCGCUU
CCGGAUGGGCCCGCGUCCCAUCAGCUAGUUGGUGGGGUAAUGGCCCACCAAGGCG
ACGACGGGUAGCCGGUCUGAGAGGAUGGCCGGCCACAGGGGCACUGAGACACGGG
CCCCACUCCUACGGGAGGCAGCAGUUAGGAAUCUUCCGCAAUGGGCGCAAGCCUG
ACGGAGCGACGCCGCUUGGAGGAAGAAGCCCUUCGGGGUGUAAACUCCUGAACCC
GGGACGAAACCCCCGACGAGGGGACUGACGGUACCGGGGUAAUAGCGCCGGCCAA
CUCCGUGCCAGCAGCCGCGGUAAUACGGAGGGCGCGAGCGUUACCCGGAUUCACU
GGGCGUAAAGGGCGUGUAGGCGGCCUGGGGCGUCCCAUGUGAAAGACCACGGCUC
AACCGUGGGGGAGCGUGGGAUACGCUCAGGCUAGACGGUGGGAGAGGGUGGUGG
AAUUCCCGGAGUAGCGGUGAAAUGCGCAGAUACCGGGAGGAACGCCGAUGGCGA
AGGCAGCCACCUGGUCCACCCGUGACGCUGAGGCGCGAAAGCGUGGGGAGCAAAC
CGGAUUAGAUACCCGGGUAGUCCACGCCCUAAACGAUGCGCGCUAGGUCUCUGGG
UCUCCUGGGGGCCGAAGCUAACGCGUUAAGCGCGCCGCCUGGGGAGUACGGCCGC
AAGGCUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAAGCGGUGGAGCAUGUG
GUUUAAUUCGAAGCAACGCGAAGAACCUUACCAGGCCUUGACAUGCUAGGGAACC
CGGGUGAAAGCCUGGGGUGCCCCGCGAGGGGAGCCCUAGCACAGGUGCUGCAUGG

CCGUCGUCAGCUCGUGCCGUGAGGUGUUGGGUUAAGUCCCGCAACGAGCGCAACC
CCCGCCGUUAGUUGCCAGCGGUUCGGCCGGGCACUCUAACGGGACUGCCCGCGAA
AGCGGGAGGAAGGAGGGGACGACGUCUGGUCAGCAUGGCCCUUACGGCCUGGGC
GACACACGUGCUACAAUGCCCACUACAAAGCGAUGCCACCCGGCAACGGGGAGCU
AAUCGCAAAAAGGUGGGCCCAGUUCGGAUUGGGGUCUGCAACCCGACCCCAUGAA
GCCGGAAUCGCUAGUAAUCGCGGAUCAGCCAUGCCGCGGUGAAUACGUUCCCGGG
CCUUGUACACACCGCCCGUCACGCCAUGGGAGCGGGCUCUACCCGAAGUCGCCGG
GAGCCUACGGGCAGGCGCCGAGGGUAGGGCCCGUGACUGGGGCGAAGUCGUAACA
AGGUAGCUGUACCGGAAGGUGCGGCUGGAUCACCUCCUUUCU

>2J00:B|PDBID|CHAIN|SEQUENCE

MPVEITVKELLEAGVHFGHERKRWNPKFARYIYAERNGIHIIDLQKTMEELERTFRFIED
LAMRGGTILFVGTKKQAQDIVRMEAERAGMPYVNQRWLGGMLTNFKTISQRVHRLEE
LEALFASPEIEERPKKEQVRLKHELERLQKYLSGFRLLKRLPD
AIFVVDPTKEAIAVREARKLFIPVIALADTDSDPDLVDYIIPGNDDAIRSIQLILSRAVDLII
QARGGVVEPSPSYALVQEAEATETPEGESEVEA

>2J00:C|PDBID|CHAIN|SEQUENCE

MGNKIHPIGFRLGITRDWESRWYAGKKQYRHLLLEDQRIRGLLEKELYSAGLARVDIER
AADNVAVTVHVAKPGVVIGRGGERIRVLREELAKLTGKNVALNVQEVQNPNLSAPLVA
QRVAEQIERRFAVRRAIKQAVQRVMESGAKGAKVIVSGRIGGAEQARTEWAAQGRVPL
HTLRANIDYGFALARTTYGVLGVKAYIFLGEVIGGQKPKARPELPKAEERPRRRRPAVR
VKKEE

>2J00:D|PDBID|CHAIN|SEQUENCE

MGRYIGPVCRLCRREGVKLYLKGERCYSPKCAMERRPYPPGQHGQKRARRPSDYAVRL
REKQKLRRIYGISERQFRNLFEEASKKKGVTGSVFLGLLESRLDNVVYRLGFAVSRRQA
RQLVRHGHITVNGRRVDLPSYRVRPGDEIAVAEKSRNLELIRQNLEAMKGRKVGPWLS
LDVEGMKGKFLRLPDREDLALPVNEQLVIEFYSR

>2J00:E|PDBID|CHAIN|SEQUENCE

MPETDFEEKMILIRRTARMQAGGRRFRFGALVVVGDRQGRVGLGFGKAPEVPLAVQKA
GYYARRNMVEVPLQNGTIPHEIEVEFGASKIVLKPAAPGTGVIAGAVPRAILELAGVTDI
LTKELGSRNPINIAYATMEALRQLRTKADVERLRKGEAHAQAQG

>2J00:F|PDBID|CHAIN|SEQUENCE

MRRYEVNIVLNPNLDQSQLALEKEIIQRALENYGARVEKVEELGLRRLAYPIAKDPQGY
FLWYQVEMPEDRVNDLARELRIRDNVRRVMVVKSQEPFLANA

>2J00:G|PDBID|CHAIN|SEQUENCE

MARRRRAEVRQLQPDLVYGDVLVTAFINKIMRDGKKNLAARIFYDACKIIQEKTGQEPL
KVFKQAVENVKPRMEVRSRRVGGANYQVPMEVSPRRQQSLALRWLVQAANQRPERR
AAVRIAHELMDAAEGKGGAVKKKEDVERMAEANRAYAHYRW

>2J00:H|PDBID|CHAIN|SEQUENCE

MLTDPIADMLTRIRNATRVYKESTDVPASRFKEEILRILAREGFIKGYERVDVDGKPYLR
VYLKYGPRRQGPDPRPEQVIHHIRRISKPGRRVYVGVKEIPRVRRGLGIAILSTSKGVLTD
REARKLGVGGELICEVW

>2J00:I|PDBID|CHAIN|SEQUENCE

MEQYYGTGRRKEAVARVFLRPGNGKVTVNGQDFNEYFQGLVRAVAALEPLRAVDALG
RFDAYITVRGGGKSGQIDAIKLGIARALVQYNPDYRAKLKPLGFLTRDARVVERKKYGK
HKARRAPQYSKR

>2J00:J|PDBID|CHAIN|SEQUENCE

MPKIRIKLRGFDHKTLDASAQKIVEAARRSGAQVSGPIPLPTRVRRFTVIRGPFKHKDSRE
HFELRTHNRLVDIINPNRKTIEQLMTLDLPTGVEIEIKTVGGGR

>2J00:K|PDBID|CHAIN|SEQUENCE

MAKKPSKKKVKRQVASGRAYIHASYNNTIVTITDPDGNPITWSSGGVIGYKGSRKGTPY
AAQLAALDAAKKAMAYGMQSVDVIVRGTGAGREQAIRALQASGLQVKSIVDDTPVPH
NGCRPKKKFRKAS

>2J00:L|PDBID|CHAIN|SEQUENCE

MVALPTINQLVRKGREKVRKKSKVPALKGAPFRRGVCTVVRTVTPKKPNSALRKVAKV
RLTSGYEVTAYIPGEGHNLQEHSVVLIRGGRVKDLPGVRYHIVRGVYDAAGVKDRKKS
RSKYGTKKPKEAAKTAAKK

>2J00:M|PDBID|CHAIN|SEQUENCE

MARIAGVEIPRNKRVDVALTYIYGIGKARAKEALEKTGINPATRVKDLTEAEVVRLREY
VENTWKLEGELRAEVAANIKRLMDIGCYRGLRHRRGLPVRGQRTRTNARTRKGPRKTV
AGKKKAPRK

>2J00:N|PDBID|CHAIN|SEQUENCE

MARKALIEKAKRTPKFKVRAYTRCVRCGRARSVYRFFGLCRICLRELAHKGQLPGVRK
ASW

>2J00:O|PDBID|CHAIN|SEQUENCE

MPITKEEKQKVIQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMM
VGQRRRLLRYLQREDPERYRALIEKLGIRG

>2J00:P|PDBID|CHAIN|SEQUENCE

MVKIRLARFGSKHNPHYRIVVTDARRKRDGKYIEKIGYYDPRKTTPDWLKVDVERARY
WLSVGAQPTDTARRLLRQAGVFRQEAREGA

>2J00:Q|PDBID|CHAIN|SEQUENCE

MPKKVLTGVVVSDKMQKTVTVLVERQFPHPLYGKVIKRSKKYLAHDPEEKYKLGDVV
EIIESRPISKRKRFRVLRLVESGRMDLVEKYLIRRQNYESLSKRGGKA

>2J00:R|PDBID|CHAIN|SEQUENCE

MSTKNAKPKKEAQRRPSRKAKVKATLGEFDLRDYRNVEVLKRFLSETGKILPRRRTGLS
AKEQRILAKTIKRARILGLLPFTEKLVRK

>2J00:S|PDBID|CHAIN|SEQUENCE

MPRSLKKGVFVDDHLLEKVLELNAKGEKRLIKTWSRRSTIVPEMVGHTIAVYNGKQHV
PVYITENMVGHKLGEFAPTRTYRGHGKEAKATKKK

>2J00:T|PDBID|CHAIN|SEQUENCE

MAQKKPKRNLSALKRHRQSLKRRLRNKAKKSAIKTLSKKAIQLAQEGKAEEALKIMRK
AESLIDKAAKGSTLHKNAAARRKSRLMRKVRQLLEAAGAPLIGGGLSA

>2J00:U|PDBID|CHAIN|SEQUENCE

MGKGDRRTRRGKIWRGTYGKYRPRKKK

7.1.3 Tools used in the work

RNA 123 software

RNA 123software is visualization, and a homology and *de novo* modeling tool optimized for RNA work. In homology modeling, the tool generates a 2D secondary structure of the RNA by the use of the structure based sequence alignment algorithm. By 2D structure, one can view the Watson-Crick and non-Watson-Crick base pairing by color-coding in the structure display, and also one can Analyze secondary and tertiary base interactions using Leontis Westhof classification.

Generation of 2D structure is followed by 3d structure prediction by the use of homology modeling. In non-homologous section the software employs *de novo* modeling process. The software allows one to View 3D structures generated on an enhanced Rasmol-based display. It also allows one to manipulate the visual attributes of 3D structures using the intuitive *RNA123* control panel.

When it comes to the analysis of the structure, RNA123 has ability to automatically identify and correct bond length errors, steric clashes, missing atoms, and many other errors that may be found in the modeled structure. The homology model of the software have been tested by modeling 30S ribosome subunit of *Escherichia coli* and of *Thermus thermophiles* and the result gotten are consistence with the result by the X-ray crystallography.

Pymol (visualization tool).

This is a visualization tool used in the viewing of three dimensional structures of protein, RNA or small macromolecules. It is also used in the labeling of a 3D structure differently, by labeling chains, element or atoms of the structure. It also used in taking of a quality picture and making a three dimensional structure movie of the structure under view. Pymol is easy to use and it's open source software with the latest version being version 1.5. It can be accessed at http://www.pymol.org/.

Discovery Studio Accelrys

This is a comprehensive software suite for Life Sciences research. It consists of a set of products that enable researchers to capture, access, and analyze scientific data. By using common

underlying technologies and data models, the software allows the full range of methodologies used in modern research to be seamlessly combined to solve diverse computational problems. The software is used for visualization and analysis of three dimensional structures. Among the many things that Accerlys discovery studio can perfume are the built protein structure, run scripts (in Perl), generate Ramachadria plots. With accerlys one can do the following; one can carry out *in silico* experiment hence being able to test a hypothesis. Hence reducing expenses that would be incurred in the actual experiment and the time that would be taken is reduced to.
It can be used in scientific exploration as from selection identifications up to the lead samples by the use of simulation and modeling tools.
It enables scientist to share information hence making informed decisions. This is as a result of its automated process and integrated data types

Yasara (Yet another Scientific Artificial Reality Application)

Yasara is a software package for visualization, simulation and modeling of molecules. It uses Portable Vector Language (PVL) which allows the visualization of very large proteins and the fact that its portable and one may not need to install it makes it very efficient in its usage. Yasara has capabilities just like the other visualization tools but when viewing the sequence and the structure at the same time it has better viewing view as compared to pymol and the accelrys.

Molprobity (protein structure analysis website)

This is a web server that is used to validate protein and PDB files format. It is also used in the addition of hydrogen in the structure. Molprobity gives information about atom contact analysis i.e. whether the atoms in the structure are crushing, presence of overlapping atoms and others. It also gives information on the dihedral angles, and the can also calculate the hydrogen and the Vander Waals contact in the interface. It has also capability of analyzing the RNA structure on the dihedral angles and the atoms analysis to. Molprobity has algorithm that enables fast processes and improve clarity of the structure. Molprobity is a free online tool and can be accessed at http://molprobity.biochem.duke.edu. The Sever was used in the analysis of the structure formed and also in the addition of hydrogen molecules in the structure.

Procheck (protein structure analysis website)

It's an online server that is used in the validating of geometry and restraints of protein structure. It gives detailed information on restraint violation, summary of statistics. It also checks and analyzes the stereochemistry of the structures modeled by analyzing residue-by-residue geometry and compares that with the overall geometry and then gives report of the final analyses of the structure. The server was used in the analysis of the protein modeled and also in the calculating the Root Mean Square Deviation of the structures. The server can be accessed online through the following link. http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

Swiss model sever

This is an online server that is used in the protein modeling as explained in the in the section 3.1.4. The server can be accessed online through the following link. http://swissmodel.expasy.org/