Chapter-I
Introduction, Review of Literature, Aims & Objectives
1.1. Introduction to Mycobacterial Biology and Pathogenesis

1.1.1. Mycobacterial Infections

Mycobacterial infections are one of the major infections caused by organisms of genus *Mycobacterium* from phylum *Actinobacteria*. Out of 50 species of the genus *Mycobacterium*, identified as potential human pathogens, major infections include tuberculosis (causative agent *Mycobacterium tuberculosis*) and leprosy (causative agent *Mycobacterium leprae*) (1). Species other than *M. tuberculosis* and *M. leprae* have been designated as “atypical” or “non-tuberculous” Mycobacteria (NTM). NTM are opportunistic pathogens and generally inhabit in soil and water. These organisms primarily infect immune-compromised individuals. The *M. avium* complex (MAC) is one of the major group of opportunistic pathogens from NTM. MAC causes extensive infiltration of tissues, bacteremia, and prolonged fever in HIV infected persons (2). NTM also causes localized skin infection in immune-compromised persons (3).

Typical mycobacterial infection includes tuberculosis (TB), one of the leading cause of deaths worldwide. TB is a highly infectious disease and primarily affects respiratory system and lungs (Pulmonary TB) but it may also affect other sites (Extra pulmonary) as well. *M. tuberculosis*, causative agent of TB (4) is a highly aerobic, acid fast staining bacteria. *M. tuberculosis* is a fatal pathogen that has caused 8.6 million new infections worldwide in year 2012, out of which about 1.3 million people died (5) (WHO Global Tuberculosis report, 2013). About one-third of world’s population is reported to be infected with *M. tuberculosis* but most of them show no symptoms as the bacteria remains in dormant state. Only 10% of the infected individuals develop active disease. People with compromised immunity caused by malnutrition, diabetes, or smokers have a higher risk of developing active disease (6-8). Even though the short-course direct observed treatment chemotherapy (DOTS) and the Bacille Guerin-Calmette (BCG) vaccine are available for the treatment of tuberculosis in many countries, due to recent emergence of new multi-drug resistant (MDR) and extensively drug resistant strains (XDR) of *M. tuberculosis*, the situation has become more severe. Hence development of new therapeutics for the prevention and treatment of tuberculosis beyond the existing anti tuberculosis therapies is required.
1.1.2. Brief History of Tuberculosis

The first evidence of existence of TB bacteria comes from the remains of a bison around 17000 years ago (9). Recent developments in sequencing and genetic analysis indicate modern *M. tuberculosis* strains are evolved from early progenitors present in East Africa 3 million years ago (10). On the basis of phylo-geographical divisions, *M. tuberculosis* can be divided into six major lineages namely Indo-Oceanic lineage, East-Asian lineage, East-African-Indian lineage, Euro-American lineage, West-African lineage 1 and West-African lineage 2 (11). Although they vary in their global distribution, most of them are present in East Africa region. Fossils of bones with tubercular deformities from countries like Egypt, Italy, Denmark, countries in the Middle East and America suggest that the disease was wide spread throughout the world around 4000 years ago (12). It has been known with various names as *yaksma*, consumption, wasting disease, white plague, Pott’s disease (the tuberculous manifestation of the spine), scrofula (TB of the lymphatic system), *tabes mesentrica* (TB of the abdomen) and *lupus vulgaris* (TB of the skin).

In 1689, Dr. Richard Morton established formation of tubercles as the pathology of pulmonary TB. The term "Tuberculosis" was coined by J.L. Schönlein in 1839. In 1865, Jean-Antoine Villemin, a French army doctor, demonstrated for first time that the disease could be transmitted from tuberculous animals to healthy animals (13). A few years later in 1882, Robert Koch identified and described the bacillus causing tuberculosis. He was awarded Nobel Prize in medicine in 1905 for this remarkable discovery. Later in 1890, Koch prepared a glycerin extract of tubercle bacilli called as tuberculin and presented as a remedy for the TB. Although it was not effective as treatment, later on tuberculin was used for diagnosis. In 1907, Clemens Freiherr von Pirquet set the cut-off of 5 mm for tuberculin test and identified positive reaction in children without symptomatic infection as latent TB (14).

The first vaccine against TB was developed by Albert Calmette and Camille Guérin in 1906 by using attenuated bovine strain of bovine tuberculosis, *Mycobacterium bovis* and named as bacillus of Calmette and Guérin (BCG). This vaccine was first used on humans in 1921 but received wide spread acceptance only after World War II. Development of antibiotics like streptomycin and isoniazid provided effective
treatment against tuberculosis(15). However, the emergence of new multiple drug resistant strains (MDRs) of *M. tuberculosis* proved a major setback to the hopes of complete elimination of tuberculosis. The subsequent resurgence of tuberculosis resulted in the declaration of a global health emergency by the World Health Organization in 1993.

### 1.1.3. Tuberculosis: A Global Threat

In 2012, 8.6 million new cases of tuberculosis were registered worldwide out of which 1.3 million people died (24% were HIV positive) (5) (Global tuberculosis report, WHO, 2013). HIV-TB co-infection shows synergistic interaction between these pathogens with each enhancing the pathogenicity of the other. Despite the availability of treatment, 1.4 million deaths (including 430 thousand deaths among HIV co-infected patients) were recorded due to TB in 2011 (16). Tuberculosis is a global pandemic with developing countries being affected severely. South-East Asia (29%) and Africa (27%) and Western Pacific region (19%) accounts for the larger proportion of all cases reported in 2012 (Global tuberculosis report, WHO, 2013). India and China are majorly affected countries contributing 26% and 12% respectively of total TB cases (5) (Global tuberculosis report, WHO, 2013). African countries suffer most with TB–HIV co-infection accounting for 75% of these cases globally (5) (Global tuberculosis report, WHO, 2013). The number of TB incidence is decreasing worldwide with the availability of new antibiotics. The incidences of MDR-TB, defined as disease caused by *M. tuberculosis* strains that are resistant to at least the two most effective anti-TB drugs, isoniazid and rifampicin have increased drastically in recent years. In 2012, approximately 450,000 new cases of MDR-TB were reported and estimated 170,000 people died from MDR-TB infection (5) (Global tuberculosis report, WHO, 2013). Approximately 60% of these cases are from India, China, Russian Federation, Pakistan and South Africa (16,17). Recent emergence of extensively drug resistant tuberculosis (XDR-TB), defined as the disease caused by *M. tuberculosis* strains which are resistant to all fluoroquinolones and any of three injectable anti-tuberculosis drugs kanamycin, amikacin and capreomycin in addition to isoniazid and rifampicin (17) have made situation more severe. XDR-TB cases account for 9.6% of total MDR-TB cases globally (5) (Global tuberculosis report,
Minimum one case of XDR-TB has been reported in 92 countries in the year of 2012. XDR-TB strain has been found to be spread worldwide according to a recent report of World Health Organization (5) (Global Tuberculosis Report, WHO, 2013) (Figure 1.1).

Figure 1.1. Geographical representation for the occurrence of extensively drug resistant tuberculosis (XDR-TB) cases throughout the world in 2012 (Global Tuberculosis Report 2013, WHO, 2013): the countries where at least one case of XDR-TB has been reported is colour coded in magenta colour.

About 75% of 2.9 million missed cases of tuberculosis (either could not be diagnosed or diagnosed but not reported to National TB Programmes) fall in 12 countries. India (31% of total) remains on the top of this list (5) (Global tuberculosis report, WHO, 2013). Cases of TB infection and death are primarily found in men but it remains the third largest killer for women worldwide. Developing countries being severely affected by TB infection need more attention and recent emergence of MDR-TB and XDR-TB strains indicate that there is need for new therapeutics to fight against this pathogen.

1.1.4. Pathogenesis and Treatment of Tuberculosis

*M. tuberculosis* infection generally occurs through inhalation of aerosols spread by coughing or sneezing by person infected with active TB. After inhalation it can follow
several potential outcomes including early clearance, leaving no immunological or radiological fingerprint; development of disease soon after infection (known as primary tuberculosis); or developing a subclinical or asymptomatic latent infection with the potential to transmit the active disease in later stage (18,19). Successful infection depends on many factors like virulence of strain, number of bacilli infecting, duration of UV exposure to the droplet nuclei carrying the bacteria while in atmosphere, and immunity of the person (20) (CDC, 2013). Inside lungs, bacteria are engulfed by various innate immune cells like neutrophils, macrophages and dendritic cells. Dendritic cells play a crucial role in infection and disseminating the bacteria because of being a better antigen presenter and migratory in nature. Inside macrophages, the bacteria reside in membrane bound vacuoles by altering the protein content of vacuoles and inhibiting subsequent acidification (21). Infected macrophages release chemokines which attract lymphocytes, monocytes and neutrophils. These immune cells form granulomas which further limit the spread and replication of the bacteria. Granulomas contain both activated and inactivated macrophages. Activated macrophages represent mycobacterial antigens to T-cells and activate them. As the cellular immunity responds, infected macrophages are killed and form a caseous center in the granulomas. M. tuberculosis, although unable to replicate in acidic and hypoxic environment of this damaged tissue, can remain in dormant phase for years and the infection can persist as latent asymptomatic infection (22). If immune system fails to check the bacterial growth, the granulomas keep on growing in size and numbers resulting in tissue necrosis. This tissue necrosis allows the granulomas content to leak out into lumen of lung or spread the infection into blood vessels of lung. The destruction of the lung lumen gives rise to symptoms of active tuberculosis – a persistent cough with blood in sputum (23).

The discovery of antibiotics like streptomycin (SM), p-amino salicylic acid (PAS) and isoniazid (INH) revolutionized the chemotherapy of tuberculosis infection. Later on, development of resistance against individual antibiotic led to combinatorial therapy for TB infection starting with all three drugs (SM, PAS, INH) for 3 months followed by PAS and INH for next 9 months (24). Further research in therapeutics led to discovery of other anti-TB drugs such as ethambutol (EMB), rifampicin (RMP), and pyrazinamide (PZA) (25-27). The current anti-TB regimen includes combination of EMB, INH, PZA and RMP for initial two months followed by combination RMP and INH for next 4 months.
EMB and INH inhibit cell wall synthesis (28,29), RMP inhibits transcription (30) and PZA inhibits fatty acid synthase hence hinders fatty acid synthesis (31). Discontinuity in course leads to the emergence of MDR and XDR-TB strains. Emergence of these strains poses a threat to world as no antibiotic therapy works against these strains. In addition, these long term therapies in tuberculosis leads to multiple side effects like hepatitis, dyspepsia, exanthema and arthralgia in patients leading to termination or discontinuation of therapy (32-34). Hence newer and effective drug therapy is required to combat this disease in better and more efficient manner.

Nucleic acid modulating enzymes are involved in many crucial cellular processes like replication, transcription, DNA repair, recombination and nucleoid organization (35-38) etc. These enzymes are one of the most attractive targets for designing inhibitors against these pathogens. These enzymes interact with nucleic acids in various ways. Understanding the mechanism of action of nucleic acid modulating enzymes is essential to know about their functions and mechanism of action in greater details.

1.2. Nucleic Acid-protein Interactions

Nucleic acids (DNA & RNA) are messengers of the cell. They require repertoire of proteins for expressing their message properly. Nucleic acid and protein interactions are one of most crucial cellular phenomenon because of its involvement in multiple cellular processes like replication, transcription, translation, recombination, chromatin organization and DNA repair. Any perturbation in these events could be lethal for the cell. Proteins can interact with nucleic acids either in sequence specific or non-sequence specific manner. Sequence specific interactions depend on interaction with specific bases while non-sequence specific interactions are independent of sequence of bases in nucleic acids. These interactions may be either with DNA or RNA and accordingly modulate the activity of specific pathways within cell.

1.2.1. DNA-protein Interaction

DNA-Protein interactions occur during various DNA transaction events like replication, transcription, recombination, DNA repair and chromatin or nucleoid organization. On the basis of their substrates, these interactions further can be subdivided into two sub-classes:
(a) dsDNA-Protein interactions &
(b) ssDNA-Protein interactions.

1.2.2. dsDNA-protein Interaction

Double-stranded DNA (dsDNA) binding proteins interact with major or minor grooves of dsDNA (39) but primarily interaction have been observed with major grooves of DNA (40). RNA polymerases, various transcription factors (repressors and activators) and proteins involved with genome organization are the major members of this class of proteins. Transcription factors generally exhibit specific DNA-protein interactions while proteins involved in genome packaging and chromatin organization are examples for non-specific DNA-protein interactions. In eukaryotes, these are basic proteins known as Histone proteins while in prokaryotes this function is performed by multiple proteins known as Nucleoid associated proteins (NAPs) (35,36). These proteins contain various structural DNA binding domains such as helix-turn-helix, helix-loop-helix, zinc finger, leucine zipper, HMG box and winged helix (Figure 1.2) that facilitate the binding of these proteins with dsDNA.

Figure 1.2. Cartoon representation of various dsDNA binding motifs; These structural motifs (Helix-turn-helix, Leucine Zipper, Zinc finger motif, Helix-loop-Helix, HMG domain and winged helix) are present in dsDNA binding proteins and facilitate their binding with dsDNA.
1.2.3. ssDNA-Protein Interaction

To make the copy of DNA (replication), to repair damaged DNA or to facilitate recombination or to create a copy of RNA via transcription, dsDNA needs to be unwound resulting in formation of single stranded form of DNA (ssDNA). This function is generally performed by helicases. ssDNA is highly flexible in nature because of free rotation around phosphodiester backbone and also prone to nuclease degradation. This form of ssDNA is stabilized and protected from nuclease by specialized class of proteins. These proteins help in stabilizing and manipulating single stranded form of DNA. Moreover, in many cellular processes free ends of ssDNA are produced as a byproduct of reaction which is needed to be cleared up. This function is carried out by various ssDNA exonucleases. Presence of negatively charged oxygen on phosphate group in the backbone makes ssDNA negatively charged. Hence for binding with ssDNA, proteins needs to contain positively charged amino acids like arginine and lysine on surface. The aromatic bases in DNA (Purine and Pyrimidine) are planar in nature and also make stacking interactions with aromatic amino acids such tyrosine and phenylalanine. The combination of hydrophobic, non-covalent and stacking interactions makes the basis of interactions of proteins with ssDNA.

The proteins interacting with ssDNA contain a common structural motif known as OB fold (Oligonucleotide or Oligosaccharide binding fold) which facilitate the binding of ssDNA (41). There is no conservation at sequence level in OB folds of different proteins but their topology remain conserved. The length of OB fold varies from 75 amino acid residues to 150 residues. OB fold consist of five stranded anti-parallel β-barrels terminating into α-helix (Figure 1.3A). The β-barrels are twisted in such a way that it allows the binding of ssDNA only but not with dsDNA. Bacterial RecJ like proteins (42), E. coli SSB (43), Human mitochondrial SSBs (44), bacteriophage T4 SSB (45) and Oxytrichia nova telomere-binding Protein (46) contain OB folds. The binding affinity of OB fold is not very strong with ssDNA; hence many proteins contain more than one copy of OB fold. OB folds are also present in some ssRNA binding proteins (47). Binding of OB folds with ssDNA, RNA and oligosaccharide
indicates that these folds have evolved from common ancestor and adopted consequently for binding with specific substrates (48).

Another common structural fold that exists in many single stranded nucleic acid binding proteins is Rossmann fold (Figure 1.3B). Rossmann fold contains a central six parallel $\beta$-sheets linked to pairs of $\alpha$-helices (49). RecJ like exonucleases (42) and alanyl-t-RNA synthetase (50) contains Rossman folds in their structures. In addition to the OB and Rossmann fold containing ssDNA binding proteins, many other ssDNA binding proteins like RecA (51), topoisomerase IA (52), ssDNA exonucleases (53) and adenovirus SSB (DNA binding protein) (54) that do not contain these folds also bind with ssDNA efficiently through hydrogen bonding, electrostatic and stacking interactions. These proteins may contain domains like zinc finger, leucine zipper and helix-turn-helix which facilitate the binding of nucleic acids (Figure 1.2). Presence of these diverse proteins with different structural folds indicates although folds may vary but chemical basis of ssDNA recognition remains the main theme of proteins binding to ssDNA.

ssDNA binding proteins like Exonuclease I, topoisomerase IA, RecJ exonuclease, SSBs perform important cellular functions and hence provide interesting targets for designing inhibitors against them. The study of this class of proteins would be helpful for understanding the molecular mechanisms of various DNA transaction events and
DNA repair pathways in details. The genes involved in various DNA repair pathways are highly conserved in *M. tuberculosis* and *M. smegmatis* and hence orthologs from *M. smegmatis* can be used for the studies. This study majorly focuses on structural study of two mycobacterial proteins from this class namely a ssDNA exonuclease, DHH super family member MSMEG_2630 and topoisomerase IA, MSMEG_6157 from *M. smegmatis*.

### 1.2.4. ssDNA Exonuclease

Genomic stability and integrity of the cell is affected by various DNA damaging conditions. DNA damaging agents like reactive oxygen species (ROS), reactive nitrogen intermediates (RNI) cause base modifications, generation of abasic sites and strand breaks (55). The cell has evolved to produce various enzymes and proteins to counteract the effect of these agents and maintain the genomic integrity. Exonucleases, DNA polymerases and various DNA binding proteins are involved in this process. Exonucleases are a group of enzymes that hydrolyze the phosphodiester bond of DNA backbone from free terminal ends of polynucleotide chains. These enzymes are involved in DNA repair events. On the basis of polarity of cleavage, three different kinds of exonucleases are found in prokaryotes and eukaryotes: 5’-3’ exonucleases (e.g. RecJ exonuclease), 3’-5’ exonucleases (e.g. Exonuclease I) and bidirectional exonucleases (e.g. Exonuclease VII).

ssDNA exonucleases are involved in many crucial cellular events like DNA repair pathways, replication and recombination. A number of DNA repair pathways require ssDNA exonucleases to process intermediate species of DNA (56). *E. coli* exonuclease I (Exo I) is 3’-5’ exonuclease that specifically degrades single stranded DNA and creates 5’-phosphate mononucleotide as reaction product (53). Exo I is specific to ssDNA only, doesn’t show any degradation on dsDNA and RNA. RecJ-like exonucleases are another group of enzymes which degrade ssDNA in 5’-3’ direction (57,58). They belong to the DHH superfamily and are involved in DNA repair processes like base excision repair (BER) (59), methyl directed mismatched repair (MMR) (60) and homologous recombination (61). It is also an important component of RecBCD and RecF recombination pathway (58,62). In mycobacteria,
various DNA repair pathways are similar to *E. coli* in many aspects, particularly core processes like nucleotide excision repair (NER), base excision repair (BER) and recombination. However, there are subtle differences in many pathways like non-homologous end-joining and absence of methyl directed mismatch repair process (63,64) in mycobacteria. Although a potential RecJ homologue with conserved DHH motif is present in mycobacteria (MSMEG_2630 from *M. smegmatis* and Rv2837c from *M. tuberculosis*) but whether these genes complement functionally the relevant activity remains to be explored. During the course of this work, a RecJ homologue from *M. tuberculosis* (Rv2837c) was reported and shown to degrade smaller RNA species (nanoRNA) *in vitro*. Rv2837c has been proposed to have nanoRNase like activity similar to NrnA from *B. subtilis* (65) suggesting an alternate role for this enzyme. Moreover, the molecular basis for this activity or the selection of nanoRNA over other potential substrates remains unclear.

### 1.2.5. ‘DHH’ Superfamily

Members of DHH superfamily exist in all domains of life starting from bacteria to human. DHH superfamily members include RecJ like exonuclease, family II pyrophosphatase, exopolyphosphatases, nanoRNases (65) and Drosophila and Human prune proteins (66). Characteristic feature of the superfamily is conserved DHH motif in N-terminal domain of protein. This domain contains residues required for core catalytic activity. The C-terminal domain of this superfamily is variable. On the basis of variable C-terminal region, this superfamily can be divided into two subfamilies: DHH/DHHA1 subfamily and DHH/DHHA2 family. DHH/DHHA1 subfamily is primarily present in prokaryotes and contains conserved GGGH motif at its C-terminal region which is supposed to be involved in substrate recognition (42,67). The major members of this subfamily are RecJ-like exonucleases and nanoRNases (Figure 1.4). DHH/DHHA2 subfamily members contain SELNRKD (Prune proteins)/SRKKQVVP (Family II pyrophosphatase) conserved motifs at C-terminal domain and could be involved in providing substrate specificity to the proteins. The major members of this subfamily are eubacterial metal dependent family II pyrophosphatase and eukaryotic prune proteins (Figure 1.4).
Figure 1.4. Domain arrangement of DHH superfamily members: The conserved amino acid residues in DHH domain and DHHA1/DHHA2 domain have been shown in grey boxes. The terminal residues are indicated.

RecJ exonuclease is one of the most studied protein from DHH/DHHA1 sub-family. The structure for full length RecJ exonuclease from *Thermus thermophilus* reveals an O-shaped structure consisting of four distinct domains (42,68) (Figure 1.5).

Figure 1.5. Cartoon representation of full length structure of RecJ exonuclease from *T. thermophilus*; DHH (Yellow), DHHA1 (Cyan) and OB fold (Red) domain has been marked with different color. Two Mn$^{2+}$ ions are shown at catalytic centre and residues involved in co-ordination are represented as stick mode.

Domain I (DHH domain) consist of conserved residues of DHH motif and residues involved in catalysis. Two Mn$^{2+}$ ions are also located in this domain co-ordinated with conserved histidine and aspartate residues. Domain II (DHHA1 domain) contains
conserved residues of GGGH motif and supposed to be involved in substrate recognition. DHH and DHHA1 domain are linked by a long helical region making the active site between the domain interface. Domain III comprise of residues from N-terminal region (1-46) and approximately 110 residues long OB fold domain which is involved in binding with ssDNA. Domain IV does not show similarity with any structural fold hence may be involved in interactions with other proteins.

NanoRNases (Nrns) also belong to DHH/DHHA1 subfamily. In contrast to RecJ exonuclease (666 amino acid residues in *T. thermophilus*), nanoRNases are much shorter (313 amino acid residues in *B. subtilis*). NanoRNases are involved in degradation of short length oligonucleotides, particularly small fragments of RNA (2-5-mer in length) known as nanoRNA. In *E. coli*, the degradation of nanoRNA is carried out by a highly conserved enzyme oligoribonuclease, Orn, deletion of which is detrimental towards growth (69). Orn is a DEDD family exonuclease and is involved in RNA degradation process in many bacterial groups. Despite its important physiological role, ORN is not present in all bacteria and replaced by redundant nanoRNases in many bacterial groups. Several classes of nanoRNases; NrnA and NrnB in *B. subtilis* and NrnC in *Bartonella hensalae* have been identified (65,70,71) with distinct sequence features. Actinobacteria are only group which contain both nanoRNase and oligoribonuclease. The presence of multiple enzymes (Orn/Nrn) for similar function (nanoRNA degradation) is somewhat intriguing and suggests that these enzymes might have some additional role in vivo.

During the course of this work, a report about structure of nanoRNase from *B. fragilis* shows that it contains common structural fold as RecJ exonuclease except it does not contain OB fold (67). The crystal structure of oligoribonuclease (ORN) from *Xanthomonas campestris* (72) is also available but exhibits no structural similarity with *B. fragilis* NrnA and the two proteins contain completely different structural folds. This large amount of variation in sequence as well as in structure between ORN and nanoRNase suggests independent physiological role of these enzymes. MSMEG_2630 show very low sequence similarity (22%) with nanoRNase from *B. fragilis*. Moreover, structural studies for mycobacterial DHH/DHHA1 subfamily members, is not available and its in vivo role remain unclear. Rv2837c has been
previously reported to be essential for growth of *M. tuberculosis* H37Rv (73,74). Whether this is due to its role in DNA repair or another function is yet to be explored. A transposon mutant of MSMEG_2630 in *M. smegmatis* is not lethal and hence enables to investigate *in vivo* function of this protein (Srivastav and Kumar et al, 2014). The structural study of this MSMEG_2630 would be helpful in deciphering the mechanism of the enzyme to atomic details.

1.2.6. Topoisomerase IA

DNA superstructure must be well maintained for the optimum expression of genes and proper functioning of cell. Various essential and natural cellular processes like replication, transcription and DNA recombination introduce stress on topology of DNA. Topoisomerases, as the name indicates are a class of enzymes involved in relieving topological stress of DNA. DNA topoisomerases relieve topological stress by breaking, passing and rejoining DNA strands (75). Because of its involvement in crucial cellular processes, these enzymes are necessary for the survival of the cell. DNA topoisomerases are broadly divided into two families depending on the substrates: Type I (substrate ssDNA) and Type II (substrate dsDNA). Type I DNA topoisomerases are further classified into three subfamilies with different structure hence different functionality (Table 1.1). Type IA topoisomerase (76) relieves negative super coiling by cutting ssDNA and passing intact strand through the transient gate created by this cut resulting in change of linking number in steps of one (77-80). This mechanism is known as strand passage model. Type IA topoisomerase is found primarily in prokaryotes. Type IB topoisomerase (81) also cleaves ssDNA but differs in polarity of cleavage and follows controlled rotation model of strand passage mechanism in which cleaved strand rotates around the intact strand relieving the topological stress of DNA (82-84). This class of enzyme is primarily found in Eukaryotes. Type two topoisomerases are also classified into two subfamilies on the basis of different quaternary structure. Type IIA topoisomerase (85) acts on dsDNA, creates a transient nick in dsDNA and allows the passage of another intact dsDNA through this transient gate (86-88). This class of enzyme uses ATP molecules as source of energy to perform its function and changes linking number in steps of +/- 2. Type IIB topoisomerase (89) is similar in mechanism and domains as type IIA but
differs in quaternary structure. Type IA, Type IIA and Type IIB contain two common domains required for DNA binding. First one is toprim domain, named as toprim because it is not only found in topoisomerases but also found in bacterial primases and few nucleases (90). Toprim domain resembles with Rossman like fold and also contains few conserved residues. It also creates catalytic triad to co-ordinate Mg$^{2+}$ ion (91). Second domain is 5Y-CAP domain, named as 5Y-CAP because it contains tyrosine residue that makes covalent linkage to the 5'-end of the scissile DNA strand and it is also found in many DNA binding protein like *E. coli* Catabolite Activator Protein (CAP). It consists of “winged helix” fold (92).

**Table 1.1. Classification of Type I Topoisomerase family**

<table>
<thead>
<tr>
<th>Enzyme Characteristics</th>
<th>Prokaryotes</th>
<th>Viruses</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Topo I</td>
<td>Topo III</td>
<td>Dam</td>
</tr>
<tr>
<td>Type of enzyme</td>
<td>IA</td>
<td>IA</td>
<td>IA</td>
</tr>
<tr>
<td>Size</td>
<td>97 kDa</td>
<td>74 kDa</td>
<td>108 kDa</td>
</tr>
<tr>
<td>Gene</td>
<td><em>topA</em></td>
<td><em>topB</em></td>
<td><em>topRG</em></td>
</tr>
<tr>
<td>Cleavable Substrate (DNA)</td>
<td>ss</td>
<td>ss</td>
<td>ss</td>
</tr>
<tr>
<td>Covalent Bond</td>
<td>5'-P</td>
<td>5'-P</td>
<td>5'-P</td>
</tr>
<tr>
<td>Dependence on Mg$^{2+}$</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Dependence on ATP</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Catalytic activity</td>
<td>relaxes “−” scDNA</td>
<td>cleavage of two ssDNA</td>
<td>At 60-80°C relaxes “−” scDNA</td>
</tr>
<tr>
<td>Source</td>
<td><em>E. coli</em></td>
<td><em>E. coli</em></td>
<td><em>Desulfurococcus amylolyticus</em></td>
</tr>
</tbody>
</table>
Most well studied enzyme of Topoisomerase IA class is *E. coli* topoisomerase IA, a monomeric metal binding 97 kDa protein. It preferably binds ssDNA, cleaves it and relieves the negative super-coiling followed by ligation of cleaved strand. No full length structure of enzymes from this class has been solved until now, except top IA from *Thermotoga maritima*. But in *T. maritima*, C-terminal is short as compared to other enzymes of this class (93). For *E. coli* topoisomerase IA, structure of catalytic N-terminal domain has been solved in apo form (52) (Figure 1.6A) as well as in complex with oligonucleotides (94) (Figure 1.6B). N-terminal catalytic domain is made up of four domains that make a toroidal shaped quaternary structure. Domain I also known Toprim domain consist of four parallel β-strands sandwiched between four α- helices, like Rossman fold. Domain II consists of three anti-parallel β-strands crossing across another set of three anti-parallel β-strands making the upper surface of the toroidal structure. Domain III consists of four helices packing around the one central helix. It contains many conserved residues including active site Tyrosine-319 residue. Domain IV consists of helical region. The structure of covalent complex of DNA and Topoisomerase shows domain rearrangements to accommodate the covalent bond formation. Domain I bound to DNA showed the maximum movement as compared to apo structure. Higher B-factor of domain I in complex structure as compared to average B-factor of protein is indicative of more flexibility in this region.

On the basis of all the above observations, a mechanism of action for topoisomerase IA has been proposed (Figure 1.7). Whole reaction can be analyzed in four steps: cleavage and opening of transient gate, strand passage, religation of DNA strand followed by separation of enzyme and DNA. Two transesterification reactions are involved in catalysis by topoisomerase IA. The first reaction starts with nucleophilic attack by –OH group of catalytic Tyrosine residue on phosphodiester backbone of ssDNA, resulting in a covalent bond formation with 5’-end of scissile strand. Second reaction is religation of the DNA strand followed by release of topoisomerase enzyme. The relaxation process takes place in between these two reactions. Role of C-terminal in this whole reaction is still not clear in absence of full length structure of Topoisomerase IA.
Figure 1.6. Structure of N-terminal domain of *E. coli* topoisomerase IA; (A) Apo structure of N-terminal domain from *E. coli* topoisomerase IA (PDB ID: 1ECL). Domains I (residues 32–63, 72–157), II (214–278, 406–433, 438–475), III (279–405, 434–437) and IV (64–71, 158–213, 476–590) are colored as yellow, red, cyan and blue respectively. (B) Structure of oligonucleotide bound form of N-terminal domain of *E. coli* topoisomerase IA. Domain I (Yellow) shows the maximum deviation as compared to the Apo- structure and is directly involved in binding with the scissile strand. (PDB ID: 3PX7). Cartoons for structures were prepared in Pymol.
Figure 1.7. Proposed mechanism of action for topoisomerase IA: Topoisomerase IA mediated DNA relaxation takes place through four different stages (Strand cleavage and opening of the gate, strand passage, rejoining of the cleaved strand followed by DNA release) as represented in the above figure.

Mycobacterial topoisomerase IA (MSMEG_6157 from *M. smegmatis* and Rv3646c from *M. tuberculosis*) exhibits some interesting features in addition to the above mentioned typical topoisomerase IA characteristics. Topoisomerase IA from *M. smegmatis* binds and cleaves at preferred site (CG/CTCTT) known as strong topoisomerase site (STS). The C-terminal domain of mycobacterial topoisomerase does not show any sequence similarity with the corresponding region of other eubacterial topoisomerase and unlike other eubacterial topoisomerases it does not contain Zinc finger motifs (95-98). *M. smegmatis* topoisomerase IA binds efficiently with both dsDNA as well as with ssDNA containing STS sites. The relaxation activity of this enzyme gets inhibited in presence of STS containing oligonucleotides (96). The current structure of topoisomerase IA from *E. coli* does not explain the role of C-terminal of the protein. In mycobacteria, The C-terminal domain remains indispensable for the relaxation activity of this protein (99). Structural and biophysical studies of this protein would provide deeper insight into the mechanism of action of this enzyme. Despite a general topological role the presence of STS in mycobacterial topoisomerase IA is unclear.
1.3. Aims and Objectives of the Study

Nucleic acid-Protein interactions are one of the most interesting class of interaction because of its involvement in various critical cellular processes (replication, transcription, genome organization, recombination and DNA repair) crucial for the survival and proper functioning of the cell. Structural studies of the enzymes involved in these processes would be helpful to understand the mechanism of these events and hence would be helpful in designing inhibitors against them. This thesis primarily focuses on structural studies on a non-sequence-specific ssDNA binding protein belonging to DHH/DHHA1 phosphodiesterase subfamily, namely, MSMEG_2630 and a sequence-specific ssDNA binding protein (type IA topoisomerase), namely, MSMEG_6157 from *Mycobacterium smegmatis*.

Following below are the objectives of this study to address structural insights into ssDNA binding protein from *Mycobacteria*:

i) **Cloning, expression and purification of mycobacterial ssDNA binding enzyme**: *MSMEG_2630* and *topoisomerase IA* (*MSMEG_6157*) gene from *M. smegmatis* will be cloned in expression vector and expression and purification of these proteins would be standardized to obtain homogenously pure protein for biophysical studies like crystallization and binding.

ii) **Crystallization trials of apo protein as well as in complex form**: To get the structural insights into these proteins the crystallization experiments will be performed with the apo- form as well as with complex form with the preferred substrates.

iii) **Three dimensional structure determination of mycobacterial enzyme**: Crystals obtained after crystallization trials will be used for the three dimensional structure determinations by X-ray diffraction methods.

iv) **Study of DNA binding and interaction, biochemical and biophysical approaches**: After structural determination of these proteins, mechanism of interaction and binding with the substrate will be studied with various biophysical methods.
v) **Comparative analysis with other enzymes of same class:** The structure of the protein will be compared with the other proteins of their respective classes to get into details of the mechanism of action of the enzyme.