Chapter-III
Results and Discussions
3.1. Structure Determination of Mycobacterial DHH Family Protein ‘MSMEG_2630’ and Comparative Analysis with other DHH Family Members

3.1.1. Introduction

MSMEG_2630 is a member of DHH superfamily of phosphodiesterases with very low sequence homology with any other members of this family. DHH superfamily members exist in all domains of life (archaea, bacteria and eukaryotes) and include 5’-3’ exonucleases (RecJ), 3’-5’ nanoRNases (Nrn), cyclic nucleotide phosphodiesterases (Drosophila and Human prune proteins), pyrophosphatases (Class II inorganic pyrophosphatases) and exopolyphosphatases. These proteins contain a conserved Asp-His-His (DHH) sequence motif in the N-terminal domain of the protein, known as DHH domain. This conserved motif is involved in catalysis and hence is essential for the activity of these proteins (119). C-terminal domain of the protein is variable and on the basis of conservation in C-terminal domain this superfamily has been subdivided into two subfamilies: DHH/DHHA1 subfamily and DHH/DHHA2 subfamily (66) (Figure 1.4).

MSMEG_2630 belongs to DHH/DHHA1 subfamily and contains GGGH conserved sequence motif in C-terminal domain of the protein. Other prominent members of this subfamily are RecJ-like exonucleases and nanoRNases. Top hits from BLAST shows closest homologs of MSMEG_2630 are members of DHH superfamily, namely, phosphodiesterases, RecJ-like exonucleases and exopolyporphosphatase-like enzymes. A BLAST search with Rv2837c, described earlier as a nanoRNase in M. tuberculosis (120), identified MSMEG_2630 as the ortholog in M. smegmatis mc²155 genome with an e-value of 10⁻¹⁶⁸. MSMEG_2630 consists of 340 amino acids and when compared with Rv2837c has sequence identity and similarity of 73% and 83%, respectively. Rv2837c has been reported to be essential for survival of M. tuberculosis (73,74). A biochemical study that appeared during the course of this work showed Rv2837c exhibits 3’-5’ nanoRNase activity on smaller RNA substrates in vitro (120). NanoRNases are
involved at the final step of degradation of small RNA (2-5-mer) species (70). A recent work showed that transposon mutant of MSMEG_2630 in *M. smegmatis* shows growth arrest in presence of DNA damaging agents (Ethidium Bromide, menadione and mitomycin C) that get restored by ectopic expression of MSMEG_2630, indicating a possible role in DNA repair pathways (Srivastav and Kumar et al, 2014). The structural study of MSMEG_2630 would help to understand the molecular mechanism of the enzyme and its interaction with potential substrates in greater details.

### 3.1.2. Sequence Analysis of MSMEG_2630

Sequence alignment of full length MSMEG_2630 with the prominent member of DHH/DHHA1 subfamily, namely, RecJ-like exonucleases shows conservation of DHH motif and GGGH motif in both subfamilies (Figure 3.1A). RecJ-like exonucleases also contain a long stretch of amino acids at the C-terminal of protein, which is absent in MSMEG_2630. Structurally this region constitutes Oligo-nucleotide binding (OB) fold domain, involved in binding with single stranded DNA substrates (42). The overall sequence identity and similarity of MSMEG_2630 with RecJ of E. coli is 12% and 17% while that with NrnA of B. subtilis is 25% and 44% respectively. Sequence alignment of MSMEG_2630 with the other prominent member of DHH/DHHA1 subfamily, namely, nanoRNases show conserved DHH and GGGH motifs and residues involved in Mn$^{2+}$ metal ion co-ordination (67). In addition, N-terminal of MSMEG_2630 contains a unique stretch 16 amino acid residues (Figure 3.1B), not present in other NrnA sequences. Sequence analysis indicates nanoRNases are closer homologs of MSMEG_2630, in terms of sequence length and conservation.
Figure 3.1. Multiple sequence alignment of MSMEG_2630 with members of DHH superfamily; (A) Multiple sequence alignment of MSMEG_2630_Msm (M. smegmatis) with RecJ like exonucleases, RecJ_Ecol (E. coli) and RecJ_Tth (T. thermophilus). Conserved DHH and GGGH motifs are shown in rectangular boxes. (B) Multiple sequence alignment of MSMEG_2630_Msm (M. smegmatis) with selected nanoRNases, NrnA_Bsu (B. subtilis) and Nrn_Bfr (B. fragilis) and Nrn_Sha (S. haemolyticus). Conserved DHH and GGGH motifs are shown in rectangular boxes and additional 16 amino acids residues at N-terminal of MSMEG_2630 in M. smegmatis is highlighted in brown rectangular box. The residues that are involved in metal ion (Mn$^{2+}$) co-ordination in Nrn_Bfr structure (3W5W) are also conserved in MSMEG_2630_Msm.

Further examination of genomic loci ofnrnA from B. subtilis and MSMEG_2630, revealed striking differences in patterns of genic neighbors these genes. Genic neighbours for other sequence homologs of B. subtilis NrnA (sequence homologs were obtained as bidirectional best hits from BLAST), were also analyzed.
Translation initiation factor-2 (infB) (MSMEG_2628) and Ribosome binding factor-A (rbfA) (MSMEG_2629) are upstream genic neighbors of MSMEG_2630 and appear to share a transcription start site with MSMEG_2630 (Figure 3.2A). An open reading frame (ORF) encoding a multidrug and toxic compound extrusion protein (MATE)/ DNA-damage-inducible protein F (dinF) (MSMEG_2631) present downstream of MSMEG_2630 also appears to be transcribed from the same transcription start site. While MATE is a genic neighbor of mycobacterial sequences only, nusA, a transcriptional regulator present further upstream of MSMEG_2630, is also commonly found in this locus in all the examined sequences. On the other hand, catalytic subunit alpha of DNA polymerase III (dnaE), acetyl CoA carboxylase (accA), 6-phosphofructokinase (pfkA) and pyruvate kinase (pyk), are the frequently observed neighbors of nrnA of B. subtilis and many other bacterial species (Figure 3.2B).
Figure 3.2. Synteny conservation suggests two major clusters in nanoRNase like sequences; (A) Genic neighbors of MSMEG_2630, infB (translation initiation factor B), rbfA (ribosome binding factor A), musA (transcription elongation factor A) and MATE/dinF (DNA damage inducible protein F) differs from (B) genic neighbors of NrnA from B. subtilis, dnaE (DNA polIII subunit beta), malic enzyme, accD (acetyl-coenzyme A carboxylase carboxyl transferase subunit beta), accA (acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha), pfkA (6-phosphofructokinase) and pyk (pyruvate kinase), indicating differences in possible roles in vivo.

A phylogenetic tree constructed for NrnA-like and MSMEG_2630-like sequences (on the basis of genic neighbours) showed clear clustering of these genes into two major clades, with infB, rbfA and musA as genic neighbors in one clade and other set of enzymes (dnaE, malic enzyme, accD, accA, pfkA and pyk) as genic neighbors in the other (Figure 3.3). All the sequences appear to club into either of the two clades, based on the specialized function possibly affected by the genic neighbors as likely interacting partners. MSMEG_2630, like Rv2837c, is distinct from B. subtilis NrnA and is part of a separate clade indicating it could have different in vivo role in comparison with NrnA from B. subtilis.
Figure 3.3. Phylogenetic analysis; A phylogenetic tree for selected bacterial nanoRNases indicate grouping of these proteins into distinct clades.

Rv2837c and B. subtilis NrnA have been characterized previously and shown to have 3'-5' nanoRNase activity (65,120). MSMEG_2630 also exhibits 3'-5' nanoRNase activity on 5-mer RNA. In addition, MSMEG_2630 exhibits in vitro 3'-5' exonuclease activity on longer ssDNA substrates (49-mer) and is also involved in rescue from growth arrest due to DNA damaging agents like hydrogen peroxide, menadione, EtBr and mitomycin (Srivastav and Kumar et al, 2014). The presence of
DNA damage inducible protein (\textit{dinF}) in neighborhood indicates the possible role of MSMEG_2630 in DNA damage repair. It also exhibit activity on smaller DNA/RNA species, possibly it may have also some specific roles in regulation of RNA-mediated transcriptional and/or translation-related activities. Sequence comparison, phylogenetic tree analysis and examination of genic neighbours, indicates \textit{MSMEG_2630}, differs from \textit{NrnA} of \textit{B. subtilis} and hence may have some additional roles \textit{in vivo}.

3.1.3. MSMEG_2630 Purification and Crystallization:

MSMEG_2630 was purified from pET28-His10-Smt3 expression vector and concentrated to 30 mg/ml with Amicon concentrator (<3 kDa cut off filter). The purity of protein was confirmed by running the sample at 12 % SDS-PAGE (Figure 3.4A) and further used for setting up crystallization experiments. The NTD construct (MS-NTD-B) was also purified (Figure 3.4B) as per protocol and used for biochemical assays.

\textbf{Figure 3.4. Purification of MSMEG_2630; (A) 12\% SDS-PAGE shows purified MSMEG_2630 (35.1 kDa). (B) Purified N-terminal construct (MS-NTD-B) (38.7 kDa) and native MSMEG_2630 with N-terminal Sumo tag (Ntag_MS_2630) (52.1 kDa).}
3.1.4. Crystallization Trials of MSMEG_2630

Purified MSMEG_2630 (30 mg/ml) was put up for crystallization with commercially available crystallization screens at multiple temperatures (10°C and 24°C). Initial tiny rod-shaped crystals (40-60 µm) (Figure 3.5A) were obtained in crystallization condition 0.2 M LiCl, 0.1 M HEPES pH 7.0 and 26% PEG 6000. These crystals were confirmed as protein crystals by staining them with a commercial staining dye, Izit dye (Hampton Research) (Figure 3.5B). These crystals were also put for test diffraction at home source but diffracted very poorly with low resolution (>8Å).

![Image](A) ![Image](B)

**Figure 3.5. Crystallization of MSMEG_2630; (A) Initial tiny rod shaped crystals (40-60 µm) obtained for MSMEG_2630. (B) MSMEG_2630 crystals were stained with Izit dye; blue color picked up by crystals indicated these are protein crystals.**

3.1.5. Crystallization Optimization

The initial crystals obtained were very tiny (40-60 µm), fused and fragile in nature and hence needed to be optimized for good quality diffraction data. Various crystallization optimization methods (Changes in temperature, concentration of protein and pH gradients) were used. Large, thick and rod shaped (300-400 µm) crystals were finally obtained with micro-seeding experiments (Figure 3.6). These crystals diffracted to 2.16 Å resolution at BM14, ESRF, Grenoble.
Chapter 3

Figure 3.6. Crystals of native MSMEG_2630 were optimized with micro-seeding experiments and large, rod shaped (300-400µm) crystals were obtained.

3.1.6. Data Collection

The diffraction data for native MSMEG_2630 crystals were collected at ESRF, beam line BM14 to 2.16 Å resolution. Initial processing of this data indicated that these crystals belong to $P2_12_12_1$ space group with unit cell dimensions of $a=65.11$, $b=79.87$, $c=111.37$, $\alpha=\beta=\gamma=90^0$. Matthew’s coefficient (Vm) for this crystal was calculated to be $2.03 \AA^3$/Da showing 39% solvent content, assuming a dimer in the asymmetric unit.

For MAD and MIRAS experiments, data for Pt and Hg soaked crystals were also collected at the respective anomalous energies (for Pt 11.568 keV and for Hg 12.317 keV) at ESRF beam-line BM14.

3.1.7. Phasing Strategies

a) Molecular Replacement

The closest structural homologs for MSMEG_2630 in PDB database were 3DMA (25%) from *B. fragilis* and 3DEV (22%) from *S. haemolyticus*. We tried for molecular replacement solution with native diffraction data using these two structures as template. For Molecular Replacement solution, AutoMR wizard
coupled with AutoBuild from Phenix software package were used. The best solution with $R_{\text{free}} = 0.58$ and $R_{\text{work}} = 0.44$ was obtained. The solution obtained did not produce good quality maps and hence could not proceed further with model building.

b) MAD and MIRAS Phasing

Pt and Hg soaked crystals were scanned at the respective anomalous energies for the presence of heavy atoms at BM14 and the scan showed the incorporation of heavy atoms in the crystals (Figure 3.7A & 3.7B). The scan for Au soaked crystals did not show any incorporation for Au. The Pt and Hg anomalous data were collected and processed with $iMOSFLM$ (101,102) and scaled with $SCALA$ (106). Scaled data was further processed with AutoSol wizard in Phenix software package (104) for respective anomalous datasets to obtain phases and initial solution. The Pt-MAD experiment provides an initial solution with $R$-values ($R_{\text{work}} = 0.44/R_{\text{free}} = 0.53$), Figure of Merit (FOM) = 0.22 and Map Correlation Coefficient (Map-CC) = 0.41. Despite several attempts at model building and refinement, map quality did not improve. The best model obtained consisted of 390 amino acid residues out of 680 amino acid residues in 69 fragments but the fragments did not contain any standard secondary structure elements. The map quality was very poor and could not be processed further. Similarly, The Hg-MAD experiment gives an initial solution with $R$-values ($R_{\text{work}} = 0.54/R_{\text{free}} = 0.58$), FOM = 0.10 and Map-CC = 0.14. This model contained 98 amino acid residues out of 680 amino acid residues in 22 fragments but no standard secondary structure element was observed in the model.

R- Factors, Map-CC values and FOM obtained for Hg and Pt anomalous data indicated that phases could not be improved significantly. The attempts for model building could not be carried out further due to poor quality maps. We also tried MIRAS phasing strategy by combining native datasets and Heavy atom derivative datasets using AutoSol wizard but could not provide solution.
Figure 3.7. Extended X-ray Absorption Fine Structure (EXAFS) scans of the crystals at the absorption edge of heavy metal anomalous energy; (A) EXAFS scan for the presence of Pt atoms in crystals at the Pt absorption edge (11.568 keV Energy). (B) EXAFS scan for the presence of Hg atoms in crystals at the Hg absorption edge (12.317 keV Energy).

c) Sulphur SAD Phasing

We collected the sulphur-SAD data with a chromium anode at home source (NII, New Delhi) to 3.7 Å resolution. In addition to low resolution, crystals were anisotropic and diffraction quality in many frames was poor. No sulphur anomalous signal could be detected with this data. We also collected the diffraction data for sulphur-SAD at BM14 at 13.000 keV (remote data set) and 7.1601 keV (sulphur anomalous data) and
at 0 Kappa and 70 Kappa angles for increased redundancy to 3.3 Å resolution. The data was processed with *iMOSFLM* (101,102) and scaled with *SCALA* program (106) of CCP4 Package (103). Scaled data was further processed with AutoSol wizard from Phenix software package (121) to obtain initial solution. The best solution obtained with R-values ($R_{\text{work}} = 0.48/R_{\text{free}} = 0.55$), FOM = 0.19 and Map-CC = 0.47. The model contained 276 amino acid residues out of 680 amino acid residues in 41 fragments. The maps obtained for the initial model were poor and no standard secondary structure elements were built in the model. We tried to build the model from initial solution but due to poor quality maps model could not be improved further.

**d) Se-Met SAD Phasing**

Initial attempts at Se-Met derived MSMEG_2630 crystallization did not succeed. Se-Met derived MSMEG_2630 crystals were finally optimized and diffraction quality crystals (Figure 3.8) were finally obtained by seeding experiments with native seeds.

![Figure 3.8. Se-Met derived protein (MSMEG_2630) crystals](image)

Se-Met incorporation in crystals was confirmed by Mass spectrometric analysis as well as by scanning at selenium peak energy (12.667 keV) at BM14 beamline at ESRF. The EXAFS scan showed the incorporation of selenium in the crystals (Figure 3.9). Data for Se-Met derivatized MSMEG_2630 crystals were collected at the same beamline at selenium peak energy (12.667 keV) to 2.2 Å resolution and further used for structure determination by Se-Met SAD method.
Figure 3.9. Extended X-ray Absorption Fine Structure (EXAFS) scan of Se-Met derived crystals of MSMEG_2630 at the selenium absorption edge to confirm the incorporation selenomethionine in crystals.

The dataset was processed with iMOSFLM (101,102) and scaled using SCALA program(106) of CCP4 Package (103). The crystals belonged to P1 21 1 with unit cell dimensions a=65.87 Å, b=86.10 Å, c=66.03 Å, α=90°, β=118.37, γ=90° (Table 2.3). The scaled dataset was submitted to online Auto-Rickshaw server for structure determination with required information for selenium SAD dataset. Initial model obtained from Auto-Rickshaw contained R-values (R\textsubscript{work}=0.246/R\textsubscript{free}=0.299). Initial model obtained from Auto-Rickshaw solution contained 554 amino acid residues out of 680 amino acid residues.

3.1.8. Structure Refinement and Final Model Building

The initial model obtained from Auto-Rickshaw server was refined with REFMAC5 (109) and BUSTER 2.1.0 (110) software package. The Refinement statistics are given in table 3.1. The final model built with R\textsubscript{work}=0.176 and R\textsubscript{free}=0.225 with average B-factor (Å\textsuperscript{2})=39.0 and contains a dimer in asymmetric unit. The final model consists of a total of 638 amino acid residues out of 680 residues for both chains. First 21 amino acids at the N-terminal were disordered for both chains and hence could not be included in the final model. Ramachandran plot analysis showed 98.0% residues in
most favored region (Molprobity server) (122) (Figure 3.10). The final coordinates for MSMEG_2630 were submitted to RCSB server with PDB ID 4LS9.

**Table 3.1. Structure refinement statistics**

<table>
<thead>
<tr>
<th>Refinement</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>19.38 - 2.20</td>
</tr>
<tr>
<td>Number of reflections working/test</td>
<td>32831 / 1666</td>
</tr>
<tr>
<td>$R_{work}$ (%)</td>
<td>0.176 (0.183)</td>
</tr>
<tr>
<td>$R_{free}$ (%)</td>
<td>0.225 (0.243)</td>
</tr>
<tr>
<td>Total Protein residues/atoms</td>
<td>638/4779</td>
</tr>
<tr>
<td>Total Water molecules</td>
<td>268</td>
</tr>
<tr>
<td>Wilson B factor (Å$^2$)</td>
<td>31.7</td>
</tr>
<tr>
<td><strong>Average B factor (Å$^2$)</strong></td>
<td></td>
</tr>
<tr>
<td>Protein atoms (chain)</td>
<td>39.0</td>
</tr>
<tr>
<td>Water molecules</td>
<td>39.3</td>
</tr>
<tr>
<td><strong>r.m.s.d. from ideal</strong></td>
<td></td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
<td>0.01</td>
</tr>
<tr>
<td>Bond angles (°)</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>Ramachandran plot</strong></td>
<td></td>
</tr>
<tr>
<td>Most Favored regions (%)</td>
<td>98.0</td>
</tr>
<tr>
<td>Disallowed regions (%)</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Figure 3.10. Ramachandran plot analysis** of MSMEG_2630 (Molprobity server) indicates all amino acid residues fall in allowed region.
3.1.9. N-terminal Sequencing

First 21 amino acid residues from N-terminal of protein sequence could not be included into the final model of MSMEG_2630. To confirm the presence of these residues at the N-terminal of protein in the crystals and to avoid the speculation of any possibility of N-terminal degradation, we did the N-terminal sequencing of the protein in the crystals. N-terminal sequencing results indicated the N-terminal of the protein is intact in the crystals and there is no degradation. The major peaks in N-term sequencing were corresponding to amino acid sequence of SMPVTTTDPK as expected. The extra Serine residue was contributed from sumo tag during cloning.

3.1.10. Overall Structure of MSMEG_2630

The overall structure of MSMEG_2630 contains two molecules in the asymmetric unit forming a homo-dimer of identical subunits (Figure 3.11).

![Overall structure of MSMEG_2630](image)

**Figure 3.11. Overall structure of MSMEG_2630**, consisting of two monomers twisted against each other making a homo-dimer. Active site residues have been shown in stick model and Mn^{2+} ion at active site has been shown as magenta sphere. The N-terminal and C-terminal domain have been designated as DHH and DHHA1 domain respectively.

The quaternary structure of MSMEG_2630 was also confirmed as dimer by size exclusion chromatography over Superdex 75 10/300 GL column (GE Healthcare)
The overall structure of each subunit can be further sub divided into two distinct domains, a N-terminal domain (NTD) comprising residues 22-206 and a shorter C-terminal domain (CTD) comprising residues 224-340. The two domains are joined by a short linker region consisting of residues 207-223. Linker region and residues of N-terminal and C-terminal domain near linker region are involved in dimer interface. The unique N-terminal stretch in mycobacterial sequences (Figure 3.1B) comprising residues 1-21 of MSMEG_2630 were disordered in both subunits and hence could not be included in the final model.

The N-terminal DHH domain of MSMEG_2630 consists of a twisted five-stranded parallel β-sheet with helices on both sides (Figure 3.13). The topology resembles the six-stranded parallel β-sheet of nucleotide binding α/β-fold or Rossman-fold with differences in the crossover helices as also confirmed by DALI (for instance, a Z-score of 10.6 and rmsd of 2.8 for 133 Cα atoms was obtained for PDBID: 1DJL, a proton-translocating transhydrogenase). The NTD harbors the conserved DHH motif and hence designated the DHH domain. One Mn²⁺ was located in the electron density map coordinated to the conserved residues in the DHH domain. The CTD consists of a central mixed β-sheet surrounded by helices on both sides. The C-
terminal domain harbors conserved residues $^{313}$GGGH$^{316}$ of the DHHA1 motif and designated DHHA1 domain of MSMEG_2630.

The average B-factor of CTD (49.8 Å$^2$ or 41.6 Å$^2$ in the two subunits) is comparatively higher than the average B-factor of NTD (27.9 Å$^2$ or 34.6 Å$^2$) in both subunits, suggestive of larger flexibility in CTD. The NTD or CTD have similar conformations when superposed with the corresponding domain in the other subunit with low rmsd values of 0.29 Å (185 C$^\alpha$ atoms) and 0.30 Å (114 C$^\alpha$ atoms), respectively (Figure 3.14B & 3.14C). However, superposition of complete subunits (rmsd: 2.59 Å for 295 C$^\alpha$ atoms) reveals different orientations indicating both NTD and CTD can move relative to each other due to flexibility around the linker region and provides multiple snapshots of the enzyme (Figure 3.14A).
Figure 3.14. Superposition of MSMEG_2630 subunits: (A) Structural alignment of individual subunits of MSMEG_2630 (green and cyan). N-terminal DHH domains of each subunit aligns very well while C-terminal DHHA1 domain is present in different conformations in the two subunits due to movements around the flexible linker region. A Mn$^{2+}$ marks the catalytic center (magenta sphere) (B) Individual N-terminal DHH and (C) C-terminal DHHA1 domains of individual subunits however align better. Terminal residues of each chain are indicated.

3.1.11. Active Site

The active site of MSMEG_2630 is located at the domain interface and defined by a Mn$^{2+}$ coordinated to conserved residues in the DHH domain. Mn$^{2+}$ is present in an octahedral geometry coordinated to Asp-51, His-135, Asp-185 of the protein. In addition, Asp-110 is coordinated to Mn$^{2+}$ in a bi-dentate fashion along with two water molecules at distances of 2.33 Å and 2.75 Å to complete the coordination state (Figure 3.15). Although Mn$^{2+}$ is generally present in an octahedral geometry with six coordination state, the seven coordination state (with a bi-dentate coordination with at least one residue) has been previously seen for Mn$^{2+}$ in several hydrolases (PDBID: 1D3V) (123) and phosphatases (PDBID: 1G5B) (124) in PDB.
The catalysis by DHH phosphodiesterase family members including NrnA (67), RecJ (42,68) or pyrophosphatases (125-127) follow a two-metal ion mechanism. However, we could identify only one Mn\(^{2+}\) in each subunit in apo-structure of MSMEG_2630. In the absence of a bound substrate, the position for the second metal ion is not occupied, suggesting that the second metal ion is not structurally required and could be involved in catalysis and hence may bind during catalysis.

![Figure 3.15. Stereo view of Mn\(^{2+}\) coordinated at the active site;](image)

The presence of Mn\(^{2+}\) in the structure was somewhat surprising as Mg\(^{2+}\) was used in protein purification and crystallization. The coordination geometry of both Mg\(^{2+}\) and Mn\(^{2+}\) is very similar and their coordination distances differ only by 0.1 Å (128). The metal ion in the crystal structure was confirmed as Mn\(^{2+}\) through several lines of evidence. Firstly, the local B-factors during refinement were found to be closer to the neighbouring residues for a modelled Mn\(^{2+}\) (B-factor: 30.1 or 36.1 Å\(^2\) in the two subunits) than a modelled Mg\(^{2+}\) (B-factor: 10.1 or 12.1 Å\(^2\)). Moreover, the coordination state of Mn\(^{2+}\) in the active site of Family II pyrophosphatases has been suggested to change from five to six upon substrate binding. Varied coordination
states are possible for transition metals viz., Mn\(^{2+}\) but not Mg\(^{2+}\) (126,129-131). Mn\(^{2+}\) is hence conducive for catalysis due to its alterations between different coordination states, necessary for catalysis (126). Thirdly, Mn\(^{2+}\) has been suggested to accommodate bi-dentate coordination by carboxylate groups (Asp-110 of MSMEG_2630) better than Mg\(^{2+}\) due to its larger ionic radius (127). Lastly, Mn\(^{2+}\) is a stronger Lewis acid than Mg\(^{2+}\) and is likely to activate a water molecule more effectively for a nucleophilic attack for catalysis (127). Exonuclease activity assays indicate that although Mn\(^{2+}\) is the preferred ion for catalysis at low concentrations; at saturating concentrations, both metal ions (Mg\(^{2+}\) and Mn\(^{2+}\)) can substitute each other for exonuclease activity in vitro. However, Ca\(^{2+}\), Ni\(^{2+}\), Co\(^{2+}\), Zn\(^{2+}\) and Cu\(^{2+}\) could not serve as cofactors when used at saturating concentrations and exhibit no activity (Srivastav and Kumar et al, 2014).

3.1.12. Role of DHH Motif

His-135 is the only residue from the conserved DHH\(^{134}\)DHH\(^{136}\) motif that is directly involved in metal coordination with Mn\(^{2+}\) in apo structure of MSMEG_2630 and directly involved in catalysis process. The other two residues, Asp-134 and His-136 appear to be structurally important as O\(\delta 1\) of Asp-134 is hydrogen bonded to N\(\delta 1\) of His-136 and positions His-135 for coordination to Mn\(^{2+}\). Any change in this region is likely to perturb the coordination state of Mn\(^{2+}\) and affect activity (Figure 3.16).

Figure 3.16. Role of Asp134: Asp-134 of DHH motif interacts with His-136 and Asn-139 and plays a role in maintaining the catalytic core of MSMEG_2630. Manganese ion at catalytic centre has been shown as sphere (Magenta) and distances between these residues have been labeled (Å).
Mutation of D134A in MS-D134A led to loss of both exonuclease as well pAp phosphatase activity in vitro (Srivastav and Kumar et al, 2014).

A CD profile of MS-D134A (Figure 3.17) is similar to the CD profile of WT MSMEG_2630 and confirms the overall folded state of the protein. Any potential local perturbations in MS-D134A structure appear to be minor and not directly discernible in the CD profile, suggesting no major structural changes take place in MS-D134A except some perturbation at active leading to the loss of exonuclease activity.

![Figure 3.17. CD analysis; CD spectra of (A) mutants (MS-D134A, closed squares; MS-H316A, closed triangles) and WT (rMSMEG_2630-A, open circles) are indicated. All spectra are indicative of folded states of the proteins.](image)

### 3.1.13. Structural Comparisons

A DALI search (114) reveals top structural hits from all the major members of DHH superfamily (Table 3.2), namely, (i) NanoRNases (DHH/DHHA1 subfamily), (ii) RecJ exonucleases, (also belonging to DHH/DHHA1 subfamily) and (iii) Metal dependant inorganic pyrophosphatases (belonging to DHH/DHHA2 subfamily). The closest structural homologs of MSMEG_2630 are *B. fragilis* nanoRNase structures (PDBIDs: 3W5W, Z-score 29.6, rmsd: 4.5 Å for 340 Cα atoms and 3DMA, Z-score 30.1, rmsd: 4.2 Å for 340 Cα atoms). *S. haemolyticus* exopolyphosphatase (PDBID: 3DEV, Z-score 28.0, rmsd: 3.2 Å for 316 Cα atoms), an uncharacterized protein from...
Northeast Structural Genomics Consortium also shows high scores from DALI and appears to belong to NrnA family.

The overall structure and arrangement of subunits in MSMEG_2630 is quite similar to the GMP-bound NanoRNase (Nrn) structure of *B. fragilis* (PDBID: 3W5W and 3DMA) (Figure 3.18). The superposition of individual subunits as given by DALI, however, is poor. In addition, a short stretch of residues at the N-terminus (residues 22-27 in MSMEG_2630 and 2-11 in *B. fragilis* NrnA) do not align in the structures, indicating variability in the N-terminal region of nanoRNases.

![Figure 3.18. Structural comparison of MSMEG_2630 with B. fragilis NanoRNase; Stereo-view of superposition of MSMEG_2630 (subunits in green and cyan) with NanoRNase of B. fragilis (PDBID: 3W5W) (red) indicates CTD of B. fragilis NanoRNase is slightly shifted towards catalytic center. A bound Mn$^{2+}$ (magenta sphere) marks the active site of MSMEG_2630.](image)

A closer examination of individual subunit structures indicate the C-terminal domain in GMP-bound 3W5W moves upwards towards N-terminal domain shortening the cavity at the domain interface (Figure 3.18). In MSMEG_2630, the cavity at the domain interface remains in a relatively open conformation resulting in large deviations when aligning the structure of MSMEG_2630 with *B. fragilis* Nrn. In the MSMEG_2630 apo structure, the distance between His-316 of the $^{313}$GGGH$^{316}$ DHHA1 motif and Mn$^{2+}$ is 14.6 Å while in the ligand bound ‘closed conformation’ in 3W5W, this distance is 6.6 Å. The individual domains (NTD and CTD), however,
superpose much better (Figure 3.19A & 3.19B) with rmsd of 1.61 Å (173 C\textalpha atoms) and 1.62 Å (110 C\textalpha atoms) for NTD and CTD, respectively, confirming thereby that larger deviations in the overall structures of MSMEG_2630 and \textit{B. fragilis} Nrn are due to domain closure in the GMP-bound state. Overall similarities in the individual domains and conservation of conserved residues in equivalent positions, however, suggest that \textit{B. fragilis} nanoRNase and MSMEG_2630 bind to and degrade nucleotide substrates through similar mechanisms.

![Figure 3.19. Superposition of NTD (A) & CTD (B) alone of the two proteins MSMEG_2630 (Green) and 3W5W (Red) shows better alignment indicating overall similar structures of the individual domains.](image)

A DALI search also revealed structural similarities between MSMEG_2630 and RecJ exonucleases (PDB ID: 1IR6 and 2ZXP), both members of DHH/DHHA1 subfamily (Table 3.2). The structures of the N-terminal DHH domain and the C-terminal DHHA1 domains of both proteins are very similar to each other and superpose well with rmsd of 1.82 Å (for 185 C\textalpha atoms) and 2.83 Å (for 104 C\textalpha atoms), respectively (Figure 3.20A & 3.20B). However, while all other top structural homologs consist of biological dimers, RecJ family exists as a monomer. An additional OB-fold domain present in RecJ (42) but absent in nanoRNases confers binding ability to longer oligonucleotide substrates.
Figure 3.20. **Structural comparison with RecJ;** Superposition of (A) N-terminal DHH domain and (B) C-terminal DHHA1 domain of MSMEG_2630 (cyan) and *T. thermophilus* Rec J (PDBID: 2ZXP) (hot pink) indicates overall similar folds for the two domains.

3.1.14. Differential Subunit Packing in DHH/DHHA1 and DHH/DHHA2 Subfamilies

The third major family showing structural similarities with MSMEG_2630 is family II pyrophosphatases and exopolyphosphatases (Table 3.2). Structural comparison of MSMEG_2630 with *Streptococcus mutans* family II pyrophosphatase (127) provides interesting insights. Both MSMEG_2630 and the *S. mutans* pyrophosphatase form homo-dimers with each monomer folding into two distinct domains. Despite divergent sequences in the N-terminal DHH and the C-terminal DHHA1 or DHHA2 domains of MSMEG_2630 and *S. mutans* pyrophosphatase (PDBID: 1I74), individual subunits (rmsd of 2.92 Å for 309 Cα atoms) and domains (rmsd of 2.19 Å for 146 Cα atoms of NTD and 2.80 Å for 93 Cα atoms of CTD) of these two subfamilies have similar folds and show overall similar structures (Figure 3.21 A & 3.21B), reminiscent of domain movements relative to each other upon substrate binding in *B. fragilis* Nrn.
Table 3.2. DALI server results for structural homologs of MSMEG_2630

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Description</th>
<th>Z-score</th>
<th>RMSD</th>
<th>Aligned residues</th>
<th>% ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DMA</td>
<td>nanoRNase (B. fragilis)</td>
<td>30.1</td>
<td>4.2</td>
<td>309</td>
<td>25</td>
</tr>
<tr>
<td>3W5W</td>
<td>nanoRNase (B. fragilis)</td>
<td>29.6</td>
<td>4.5</td>
<td>309</td>
<td>25</td>
</tr>
<tr>
<td>3DEV</td>
<td>Hypothetical protein (S. haemolyticus)</td>
<td>28.1</td>
<td>3.2</td>
<td>299</td>
<td>22</td>
</tr>
<tr>
<td>1WPM</td>
<td>Pyrophosphatase (B. subtilis)</td>
<td>19.7</td>
<td>10.2</td>
<td>240</td>
<td>16</td>
</tr>
<tr>
<td>1I74</td>
<td>Pyrophosphatase (S. mutans)</td>
<td>19.4</td>
<td>2.6</td>
<td>171</td>
<td>22</td>
</tr>
<tr>
<td>1K20</td>
<td>Pyrophosphatase (S. gordonii)</td>
<td>19.5</td>
<td>2.7</td>
<td>171</td>
<td>22</td>
</tr>
<tr>
<td>1K23</td>
<td>Pyrophosphatase (B. subtilis)</td>
<td>19.2</td>
<td>8.8</td>
<td>237</td>
<td>16</td>
</tr>
<tr>
<td>2EB0</td>
<td>Pyrophosphatase (M. jannaschii)</td>
<td>17.1</td>
<td>9.7</td>
<td>215</td>
<td>17</td>
</tr>
<tr>
<td>1IR6</td>
<td>RecJ Exonuclease core domain (T. thermophilus)</td>
<td>16.6</td>
<td>5.4</td>
<td>259</td>
<td>20</td>
</tr>
<tr>
<td>2ZXP</td>
<td>RecJ Exonuclease full length (T. thermophilus)</td>
<td>16.3</td>
<td>5.0</td>
<td>259</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 3.21. Structural comparison with family II pyrophosphatases; Superposition of (A) NTD and (B) CTD of MSMEG_2630 (cyan) and S. mutans pyrophosphatase (PDBID: 1I74) (orange) indicates overall similar folds for the two domains.
However, there is a striking difference in the subunit arrangement in dimers in these two classes of proteins (Figure 3.22A & 3.22B). The family II pyrophosphatase homo-dimer is formed by the two N-terminal domains packing together making a hydrophobic dimer interface of two anti-parallel β-strands, one contributed by each monomer. Together, the six β-strands from each N-terminal domain, hence, form a single continuous 12-stranded, twisted β-sheet running through the protein (125). In MSMEG_2630, the terminal helix of the NTD (residues 194 to 205) and the first helix of the CTD (residues 227 to 234) apart from the linker region form the dimer interface through interactions with corresponding regions of the other subunit and bury 1985 Å² area at the dimer interface. In contrast to S. mutans pyrophosphatase, MSMEG_2630 subunits are oriented in a perpendicular direction to each other. This orientation presents two independent active sites on opposite faces.

Figure 3.22. Structural comparison of MSMEG_2630 with Class II Pyrophosphatases;
Cartoon representation of subunit arrangement in dimers of (A) MSMEG_2630 and (B) family II Pyrophosphatases.

The linker region in pyrophosphatases is relatively small when compared to nanoRNases and MSMEG_2630 (Figure 3.23). Consequently, the cavity in MSMEG_2630 appears to be enlarged when compared to pyrophosphatases. This
enlarged cavity between the N-terminal DHH domain and C-terminal DHHA1 domain of MSMEG_2630 may be necessary to accommodate comparatively larger substrate (oligonucleotides) as compared to pyrophosphatases.

Figure 3.23. Stereo view of superposition of MSMEG_2630 (cyan) with S. mutans pyrophosphatase (PDBID: 1I74) (orange); The longer linker region in MSMEG_2630 enlarging the cavity at the domain interface to enable binding of larger oligonucleotides substrates are indicated.

Mycobacteria and other microbes harboring nanoRNases hence utilize alternate packaging of the same well-characterized domains and structural folds for recognition and degradation of specific DNA or RNA substrates.

3.1.15. Role of NTD and CTD Domains

In order to investigate the role of individual domains, constructs expressing MS-NTD-A, MS-NTD-B and MS-CTD were prepared (as described in Chapter 2. Materials and Methods, Section 2.2.2.3). Deletion of CTD in MS-NTD-A and MS-NTD-B resulted in loss of in vitro exonuclease activity on synthetic 49-mer ssDNA substrates (Srivastav and Kumar et al, 2014). However, MS-NTD-B exhibited both pAp phosphatase and phosphodiesterase activity in vitro (Figure 3.24A & 3.24B) suggesting that N-terminal DHH domain that contains all critical residues involved in metal co-ordination and activity is capable of catalysis on its own but requires the C-
terminal DHHA1 domain for binding to longer oligonucleotide substrates. Both the phosphodiesterase activity of MS-NTD-B (8.9 pmol/µg/min) and pAp phosphatase activity (7.7 pmol/µg/min) was similar to that of full length Ntag_MSMEG_2630 (9.3 nmol/µg/min or 7.8 pmol/µg/min, respectively).

![Figure 3.24](image)

**Figure 3.24.** Specific activity of MS-NTD-A and MS-NTD-B with (A) bis-pNpp (unfilled bars) and (B) pAp (filled bars) as substrates.

The absence of pAp phosphatase and phosphodiesterase activity in MS-NTD-A was somewhat peculiar. Asp-185 of MSMEG_2630 is required for coordination of Mn$^{2+}$ and maintenance of the active site (Figure 3.15). Asp-185 is present close to the C-terminus of MS-NTD-A construct, which consists of 188 amino acid residues. We presume rearrangements in the C-terminus of MS-NTD-A affects the coordination state of Mn$^{2+}$, thereby leading to loss of activity even on small substrates, unlike MS-NTD-B. CD spectra of N-terminal constructs and Native MSMEG_2630 (Figure 3.25) show folded state of proteins indicating that tag does not affect folding of protein and constructs are structurally stable.

C-terminal DHHA1 domain of MSMEG_2630 is devoid of active site residues. In order to explore the role of CTD, a DALI search with the C-terminal DHHA1 domain alone was carried out. DALI search shows top hits from C-terminal domains of DHH superfamily members (nanoRNases, pyrophosphatases and RecJ exonuclease). Structural similarity was also observed with the C-ala domain of alanyl t-RNA
synthetase involved in single stranded nucleotide binding (PDBID: 3G98, Z-score: 8.8, rmsd: 2.8 Å for 111 Cα atoms) (50) (Figure 3.26).

Figure 3.25. **CD analysis**: CD spectra of NTD constructs (MS-NTD-A, closed triangles; MS-NTD-B, closed squares) and Ntag-MSMEG_2630 (open circles) are indicated and suggestive of folded proteins. MS-NTD-A and MS-NTD-B exhibit somewhat different profiles of secondary structure content due to presence of additional linker domain in MS-NTD-B.

Figure 3.26. **Structural superposition of C-terminal domain (Green) of MSMEG_2630 and C-ala domain of alanyl t-RNA synthetase (Purple)** indicates presence of similar folds in these domains.
Chapter 3

The overall structural similarity of CTD of MSMEG_2630 with the C-ala domain of alanyl t-RNA synthetase suggests its role in binding to longer single-stranded oligonucleotide substrates as suggested earlier for both RecJ (42) or Nrn exonucleases (67). However, no detectable binding to ssDNA or RNA could be identified with the CTD construct of MSMEG_2630 (MS-CTD) under our experimental conditions. The role of CTD and the residues therein was hence investigated by mutagenesis of H316 in the full length protein.

An activity assay with MS-H316A shows no phosphatase activity on pAp or exonuclease activity (Srivastav and Kumar et al, 2014). A GMP ligand binds close to the conserved \(^{313}GGGH^{316}\) motif in the C-terminal domain of \(B.\ fragilis\) Nrn structure, although no direct interactions were observed between GMP and residues in the motif (67). Mutation of equivalent conserved Histidine residue results in decreased exonuclease activity in \(E.\ coli\) RecJ as well (119). A complete loss of exonuclease activity in the MS-H316A in contrast to \(E.\ coli\) RecJ could be due to an additional OB-fold domain available in RecJ for oligonucleotides substrate binding. CD profile of MS-H316A (Figure 3.17) is similar to the CD profile of WT MSMEG_2630 and confirms the overall folded state of the protein.

3.1.16. Proposed Binding Model with Oligonucleotide Substrate (7-mer)

MSMEG_2630 cleaves longer ssDNA (49-mer) and RNA (35-mer) both, \textit{in vitro} in 3'-5' direction (Srivastav and Kumar et al, 2014) indicating its ability to bind both short (<5-mer) and long oligonucleotide substrates. In the absence of OB-fold (present in RecJ exonucleases) it is unclear how MSMEG_2630 binds with long substrates. We propose a model for substrate binding by MSMEG_2630 by performing docking studies with 7-mer Oligonucleotide substrate on the PATCHDOCK server. The final model obtained suggests the path for RNA strand goes through the domain interface (Figure 3.27). The flexibility provided by the long linker region may enable movements to accommodate the longer substrates at domain interface region. In recently solved Nrn structure of \(B.\ fragilis\) (3W5W), a GMP ligand has been located near His-311 residue and has been proposed in
substrate interaction. The equivalent Histidine residue (His-316) is conserved in MSMEG_2630 and it has been shown to be critical for the exonuclease activity of this protein (Srivastav and Kumar et al, 2014). Conserved His316 residue in MSMEG_2630 also comes in proximity with the oligonucleotide substrate (7-mer) in docking results. The other residues that are conserved and fall in the path of substrate oligonucleotide are primarily aromatic or charged residues, namely Tyr-179, Thr-186, Phe-189, Leu-203, Glu-267, Glu-287, Arg298 and Lys-300 and may be involved in substrate binding through stacking or H-bond interactions. Among these proposed residues, Tyr-179, Thr-186, Phe-189, Leu-203, Arg-298 are conserved across nanoRNases (Figure 3.1B) reinforcing a potential role for these residue as suggested.

![Figure 3.27. Binding model of MSMEG_2630 with 7-mer Oligonucleotide substrate, based on docking study by Patch dock Server; The N-terminal domain (Blue) and C-terminal domain (Green) connected by linker region (Red). 7-mer Oligonucleotide substrate (Yellow) spans throughout the domain interface. His-316 which supposed to be involved in substrate interaction has been shown in stick model and is in proximity with the phosphate backbone of the oligonucleotide substrate.](image-url)
3.1.17. Discussion

3.1.17.1. MSMEG_2630 and NrnA belong to Two Distinct Families

Rv2837c is the only bifunctional enzyme with both exonuclease (nanoRNase) and pAp phosphatase activity identified in *M. tuberculosis* (120) so far. We have characterized MSMEG_2630 of *M. smegmatis* and attributed this enzyme with similar biochemical properties. However, mycobacteria and a few other actinobacteria are unique in encoding a nanoRNase, an oligoribonuclease and CysQ in their genomes (65,132,133). Differing substrate specificities in multiple nano-RNA degrading enzymes in bacteria have been suggested earlier (134). In order to seek greater insights into the physiological role of different these proteins in mycobacteria (or Actinomycetes), we revisited the locus of nanoRNases in different organisms for clues to its possible physiological function. Phylogenetic analysis and examination of genic neighbors revealed clustering of sequences into two distinct groups: operons containing *IF2 (infB)* and *rbfA* genes or those lacking these. In addition to *infB* and *rbfA*, MATE/dinF encoding for MATE efflux pump is present on the same operon in both *M. tuberculosis* and *M. smegmatis* genomes. MATE efflux is involved in the efflux of toxic compounds (135). We have also shown rescue from growth arrest in presence of DNA damaging agents by complementation with MSMEG_2630 (Srivastav and Kumar et al, 2014). These combined factors, (i) genic location of MSMEG_2630 and Rv2837c, (ii) presence of a MATE efflux pump on the same operon, (iii) redundant presence of oligoribonuclease and nanoRNase only among Actinobacteria and (iv) rescue from growth arrest by complementation with MSMEG_2630, lead to a somewhat provocative assumption that while the role of oligoribonuclease in mycobacteria may be in generalized nanoRNA degradation, MSMEG_2630 may be specialized to be involved in DNA repair-related activities.

MSMEG_2630 exhibit *in vitro* 3’-5’ exonuclease activity on ssDNA (49-mer) and also has an *in vitro* pAp phosphatase activity. In addition, it shows phosphodiesterase activity on bis-pNpp (Srivastav and Kumar et al, 2014). These multiple activities observed in MSMEG_2630 are carried out by distinct but evolutionarily related
enzymes in a mechanistically similar manner in many bacteria. Presence of multiple phosphodiesterase, phosphatase, ssDNase as well as RNase activities on a single enzyme would make cellular machinery more economic and versatile, and reveals a new mode of mechanistic conservation while acting on distinct substrates.

3.1.17.2. New Packaging of an Old fold for Mechanistic Conservation

Despite a large number of structures in PDB, the total number of folds is limited (136). Proteins use these limited folds in different permutations to achieve a functional and mechanistic convergence for related activities. Functional convergence in several enzymes through conservation of active site residues despite completely different structures is a common feature in bacterial enzymes. For instance, despite completely different sequence and structure, family I and family II pyrophosphatases remain mechanistically analogous to each other through conservation of their active sites (127). Similar examples of convergent evolution to conserve their mechanisms mean that Trypsin- or subtilisin-families use the same catalytic triad for protein hydrolysis (137). Similarly, despite completely different three-dimensional structures, the active sites and catalytic mechanisms of topoisomerases and site-specific recombinases are similar (138).

MSMEG_2630 achieves a functional conservation and multiple activities in a unique way. The three-dimensional structures of either the N-terminal or C-terminal domains of MSMEG_2630, NrnA, RecJ and family II pyrophosphatases, are very similar. The loss of exonuclease activity on DNA/RNA while retaining phosphodiesterase as well as phosphatase activity by the N-terminal DHH domain (residues 1-211) of MSMEG_2630, suggests that the NTD confers activity while CTD confers substrate recognition in these proteins. However, completely different packing of the subunits in the three proteins results in recognition of different substrates in a highly specific manner. This is achieved through several minor alterations: (i) alternate packing of subunits via additional interactions in NTD, CTD and linker region of MSMEG_2630 against β-sheet extension in NTD of pyrophosphatases, giving rise to completely different quaternary structures (ii) Extension of the linker between NTD and CTD of
MSMEG_2630, to enlarge the substrate binding cavity for binding of oligonucleotides. (iii) Insertion of an additional OB-fold between the DHH and DHHA1 domain of RecJ to enlarge domain interface forcing the protein to remain a monomer. Intriguingly, apart from the above modifications, all these proteins harbor structurally similar N-terminal DHH- or C-terminal DHHA1/DHHA2- domains. Hence, through combinations of previously existing DHH and DHHA1 domains, MSMEG_2630 achieve a new quaternary arrangement enabling it to achieve multiple functions in a mechanistically similar manner. This mode of conservation of mechanism is also evolutionarily advantageous to the cell by avoiding encoding of multiple folds in its genome.

3.2. Insights into Topoisomerase IA from M. smegmatis (MSMEG_6157)

3.2.1. Introduction

Topoisomerases, as the name indicates, are a group of enzymes involved in relieving the topological stress of DNA that occurs during various DNA transaction events like replication, transcription, recombination and repair. Depending on the substrates preference, this family has been classified in two groups: Topoisomerase I (substrate ssDNA) and Topoisomerase II (substrate dsDNA) (75).

Further, Topoisomerase I have been subdivided in two sub groups: Topoisomerase IA (E. coli topoisomerase IA, topoisomerase III & Reverse Gyrases) & Topoisomerase IB (Human-topoisomerase I, Virus-topoisomerase I & topoisomerase V) (80). These two sub groups differ in their cleavage polarity (Topoisomerase IA binds to the 5’ end of phosphate group of cleavable DNA while Topoisomerase IB binds to 3’end of phosphate group). Topoisomerase IA requires Mg$^{+2}$ ions for activity and it shows ATP independent reactions. It relieves negatively super coiled DNA, catenate and knot ssDNA. Crystal structure of 67kda N-terminal for E. coli in apo- as well as in complex with DNA oligonucleotides is available and consists of four distinct domains enclosing a hole which is large enough to fit dsDNA (detailed description in chapter 1). Proposed mechanism of action on the basis of available structural information
states that the tyrosine residue of catalytic core binds covalently to the 5'-phosphoryl end of the DNA and creates a transient single stranded break which makes the way for another strand (75).

Besides the above characteristic features of Topoisomerase IA, mycobacterial Topoisomerase IA (110 kDa) shows several unique features which makes it an interesting enzyme for structural studies. Unlike other topoisomerase IA, mycobacterial topoisomerase IA does not have any zinc finger motif or any cysteine residue in its C-terminus. In addition, mycobacterial topoisomerase IA has high sequence specificity to cleavage site (binds to conserved CG/TCTT known as strong topoisomerase site, STS) and is stable at high temperature. Investigations of these unique features of this enzyme will help to understand its role in greater details.

3.2.2. Sequence Analysis of Topoisomerase IA from *M. smegmatis* (MstopoI)

*E. coli* topoisomerase (EctopoI) is one of the most studied prokaryotic topoisomerase. It exhibits the most of the typical characteristic of this class of enzyme. Sequence comparison of MstopoI with EctopoI leads to some interesting observations. The amino terminal (N-terminal) domain of these proteins share significant homology and is highly conserved. This domain contains conserved acidic triad (Asp-108, Asp-110 and Glu-112 in MstopoI), the catalytic tyrosine residue (Tyr-339 in MstopoI) and other conserved residues (Glu-9, Lys-13, Arg-341 and His-365 in MstopoI) that have been proposed to be involved in catalysis (Figure 3.28A & 3.28B). Acidic triad (DxDxE motif) is conserved in eubacterial topoisomerase IA and has been shown to be involved in co-ordination with Magnesium (Mg$^{2+}$) ion (139). In short, N-terminal domain of MstopoI contains all the minimal domains and residues required for DNA relaxation activity.
Figure 3.28. Sequence analysis of N-terminal of topoisomerase IA from *M. smegmatis*; (A) Domain arrangement of N-terminal domain of topoisomerase IA from *M. smegmatis* (Msm_NTD) and *E. coli* (Eco_NTD), Conserved residues have been shown as black bars. Four domains of N-terminal have been represented in different colors. Domain III (Yellow colored) contain the catalytic tyrosine residue, hence is known as catalytic domain. (B) Sequence alignment of N-terminal domain of topoisomerase IA from *M. smegmatis* (Msm_NTD) and *E. coli* (Eco_NTD). The conserved residues have been shown in rectangular boxes.

Carboxy terminal (C-terminal) domain of almost all eubacterial topoisomerase IA contains zinc finger motifs; usually this zinc finger motif differs from the zinc finger motifs of sequence specific DNA binding enzymes. However, C-terminal domain of
Mstopol show very low sequence homology with the respective region from Ectopol (Figure 3.29A & 3.29B) and does not contain any Zinc finger motif.

![Zinc ribbon domain and sequence alignment](image)

**Figure 3.29.** Sequence analysis of C-terminal domain of Topoisomerase IA from *M. smegmatis*; (A) Domain arrangement of C-terminal domain of Topoisomerase IA from *M. smegmatis* (Msm_CTD) and *E. coli* (Eco_CTD), Zinc finger motif has been shown with grey colored box in C-terminal domain of topoisomerase from *E. coli*. (B) Sequence alignment of C-terminal domain of *M. smegmatis* (Msm_CTD) with *E. coli* (Eco_CTD).

From the above sequence analysis, it reflects that C-terminal domain should be dispensable for relaxation activity but experimentally it has been shown that C-terminal domain is indispensable for the relaxation activity and N-terminal alone could not perform the activity of the enzyme (99). In absence of full length structure, the actual role C-terminal could not be determined. With the existing biochemical evidences it seems to be involved in holding the non-scissile strand of DNA during strand passage step of relaxation. C-terminal domain of *M. smegmatis* binds with ssDNA as well as with dsDNA with high affinity in sequence non-specific manner (99).
3.2.3. Purification of *M. smegmatis* topoisomerase IA (MstopoI_WT and MstopoI_Y339F mutant)

MstopoI (MstopoI_WT and Y339F mutant) was purified as described (chapter 2. Materials and methods, Section 2.3.3.2). The protein was concentrated with Amicon (Millipore) concentrator (<10 kDa cut off). The quality of protein was confirmed with 10% SDS-PAGE (Figure 3.30).

![Figure 3.30. Purification of Full length topoisomerase IA from *M. smegmatis*; homogeneously purified full length topoisomerase IA (MstopoI_WT and Y339F mutant) (10 % SDS-PAGE).](image)

3.2.4. Crystallization Trials with Full Length MstopoI

Purified and concentrated topoisomerase IA (MstopoI_WT and MstopoI_Y339F mutant) (10mg/ml) was put up for crystallization trials with commercially available crystallization screens at different concentrations (5mg/ml to 15mg/ml) and multiple temperatures (10°C and 24°C) using sitting drop vapor diffusion method of crystallization. No successful reproducible crystallization hit could be obtained. Co-crystallization trials with oligonucleotides (Table 2.4) were also set up with mutant MstopoI (Y339F) but did not result in crystals for diffraction experiments.

3.2.5. Purification of N-terminal and C-terminal Domain Constructs of Topoisomerase IA from *M. smegmatis*

As the crystallization trials with full length topoisomerase did not produce any crystallization hit, we opted for domain wise crystallization trials, particularly C-terminal domain. The constructs for N-terminal (1-616) and C-terminal (617-938)
domains of the enzyme were provided by Prof. Nagaraja Lab, MCBL, IISc, Bangalore in pRSETA expression vector. The C-terminal domain was expressing very well and hence purified as per protocol for purification of full length enzyme. N-terminal domain construct was expressing very poorly. It was sub cloned in pET28-His10-Smt3 expression vector and expression was optimized. NTD and CTD construct of MstopoI were purified and purity was confirmed by 12 % SDS-PAGE (Figure 3.31). Various crystallization trials were put up with the homogeneously purified and concentrated N-terminal and C-terminal constructs of MstopoI but no successful hits were obtained. The co-crystallization trials were also set up with various oligonucleotides (Table 2.4).

Figure 3.31. Homogeneously purified N-terminal and C-terminal domain of topoisomerase Iα from *M. smegmatis* (NTD & CTD) was analyzed on 12 % SDS-PAGE.

3.2.6. Genome wide Occurrence of Strong Topoisomerase Site (STS)

While a topoisomerase function suggests a generalized activity to relieve topological stress at the affected site in DNA, MstopoI does not nick DNA in random fashion; rather it cleaves DNA at specific sites known as strong topoisomerase sites (STS), CG/TCTTC (96). We wanted to investigate the distribution of STS in mycobacterial genome. In order to look at the genes harboring STS site, we mapped the mycobacterial genome for the occurrence of STS. We also mapped the genome of few other prokaryotes for the occurrence of STS sequence to compare its frequency with mycobacterial genome. We expected, STS sequence to occur more frequently in mycobacteria as compared to other eubacteria as eubacterial topoisomerase I are not known to be specific have STS preference. The average occurrences of STS in the
investigated genomes appear to be same, between the range of 0.06% to 0.2%. In *M. smegmatis* the average occurrence is 0.099% with respect to total genome size. Next, we investigated the presence of STS in promoter regions (defined as 100 bp upstream of the genic region). Surprisingly, STS sequence was distributed evenly across the genic and promoter regions of all the selected organisms. It did not show any clustering at particular region of genome. The total occurrence of STS in genome of these organisms has been given below in Table 3.3.

Table 3.3. Genome wide occurrence of strong topoisomerase site (STS) in various prokaryotes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Accession</th>
<th>Genome size (bp)</th>
<th>String occurrence in both strands (% w.r.t. Genome size)</th>
<th>String occurrence in genes (% w.r.t. Genome size)</th>
<th>No. of genes</th>
<th>String occurrence in promoter (% w.r.t. number of genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Mycobacterium tuberculosis</em></td>
<td>NC_000962</td>
<td>4200000</td>
<td>5322 (0.126%)</td>
<td>2195 (0.052%)</td>
<td>4003</td>
<td>236 (5.90)</td>
</tr>
<tr>
<td><em>Mycobacterium smegmatis</em></td>
<td>NC_008596</td>
<td>6988209</td>
<td>6901 (0.099%)</td>
<td>2892 (0.041%)</td>
<td>6716</td>
<td>301 (4.48)</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>NC_002655</td>
<td>4600000</td>
<td>6507(0.142 %)</td>
<td>2399 (0.052 %)</td>
<td>5282</td>
<td>274 (5.19)</td>
</tr>
<tr>
<td><em>Streptococcus pneumonia</em></td>
<td>AE007317</td>
<td>2000000</td>
<td>1652 (0.083%)</td>
<td>732 (0.037%)</td>
<td>2043</td>
<td>68 (3.33)</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>NC_007795</td>
<td>2830000</td>
<td>1865 (0.066%)</td>
<td>680 (0.024%)</td>
<td>2891</td>
<td>83 (2.87)</td>
</tr>
<tr>
<td><em>Thermus thermophilus</em></td>
<td>NC_005835</td>
<td>2127482</td>
<td>4176 (0.196%)</td>
<td>1962 (0.092%)</td>
<td>1982</td>
<td>216 (10.89)</td>
</tr>
</tbody>
</table>

The promoters of the genes containing STS sites were analyzed for their annotation in database (Table 3.4) by Blast2GO server.

Table 3.4. Blast2Go analysis for the occurrence of STS in promoter regions

<table>
<thead>
<tr>
<th>Blast2go Results</th>
<th><em>M. tuberculosis</em></th>
<th><em>M. smegmatis</em></th>
<th><em>E. coli</em></th>
<th><em>S. aureus</em></th>
<th><em>S. pneumoniae</em></th>
<th><em>T. thermophilus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total input proteins</td>
<td>236</td>
<td>301</td>
<td>274</td>
<td>83</td>
<td>68</td>
<td>216</td>
</tr>
<tr>
<td>Redundant proteins</td>
<td>4</td>
<td>9</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Net in proteins</td>
<td>232</td>
<td>292</td>
<td>263</td>
<td>82</td>
<td>66</td>
<td>200</td>
</tr>
<tr>
<td>Blastp result</td>
<td>230</td>
<td>292</td>
<td>262</td>
<td>82</td>
<td>66</td>
<td>200</td>
</tr>
<tr>
<td>Sequences used for GO mapping</td>
<td>230</td>
<td>292</td>
<td>262</td>
<td>82</td>
<td>66</td>
<td>200</td>
</tr>
<tr>
<td>Sequences with GO terms</td>
<td>197</td>
<td>225</td>
<td>192</td>
<td>59</td>
<td>51</td>
<td>160</td>
</tr>
<tr>
<td>Sequences w/o GO terms</td>
<td>33</td>
<td>67</td>
<td>70</td>
<td>23</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>No. of annotated sequence (GO slim performed)</td>
<td>196</td>
<td>213</td>
<td>190</td>
<td>59</td>
<td>51</td>
<td>147</td>
</tr>
</tbody>
</table>
The distribution of STS with respect to function was next analyzed. The annotated genes were classified into separate groups according to their proposed functions in the respective organisms by GO slim analysis (Figure 3.32).
Figure 3.32. Distribution of promoters of the genes containing STS site into their functional groups; (A) M. tuberculosis, (B) M. smegmatis, (C) E. coli, (D) S. aureus, (E) S. pneumoniae and (F) T. thermophilus.
The above results indicate the STS site occur evenly in promoter region of the genes of various functional categories of selected organisms. The major functional groups that harbor STS in their promoter region are clustered into four different functional groups namely, (i) nucleotide binding, (ii) DNA binding, (iii) transport activity and (iv) protein binding activities. The first and second cluster of nucleotide binding and DNA binding genes contain the maximum STS sites in their promoter region across all the organisms investigated. These genes are highly expressing genes because of their involvement in multiple cellular events like replication, transcription, translation and DNA repair. The topological stress in the part of the genome containing these genes would be naturally higher due their constant high level of expression. The higher occurrences of STS in these regions are expected to maintain the superstructure of DNA so that the expression of these genes could not get affected. Third and fourth clusters are from genes of transporter activity and protein binding genes. These are also highly expressing genes due to their involvement in many cellular events. More than 70% of total STS occurrence in promoter region falls into these four clusters only. Low expressing genes like nucleases, kinases, peptidase, antioxidants activity, sequence specific DNA binding, carbohydrate binding activity and electron carrier related genes contain fewer STS sites. The patterns of distribution for the occurrence of STS are similar in all the selected organisms. As the STS site occurrence in M. smegmatis and E. coli is comparatively same, we expect that E. coli topoisomerase should also bind with STS site more strongly and should show some preference for this sequence as MstopoI. Hence we cloned N-terminal of E. coli topoisomerase IA and performed binding studies with various set of oligonucleotides to compare it with the N-terminal domain of MstopoI.

3.2.7. Binding studies with N-terminal and C-terminal of MstopoI

To confirm the binding specificity of NTD and CTD of MstopoI, we performed binding studies of N-terminal and C-terminal domains from M. smegmatis with various oligonucleotides with or without STS sites (Figure 3.33A) and compared binding affinity of these two domains (Figure 3.33B & 3.33C). The binding affinity (K_a) of N-terminal from M. smegmatis with the 22-mer ssDNA containing STS site is 7.9x10^7 M^-1 while it does not show any binding with the 14-mer non-STS containing
ssDNA. NTD_Mstopol binds with dsDNA and dsDNA with STS overhung with low affinity of 1.57 x10^6 M^-1 and 1.41x10^6 M^-1 respectively. C-terminal domain of Mstopol binds with all oligonucleotides irrespective of their sequence with the highest affinity for dsDNA containing STS overhang i.e. 1.97x10^8 M^-1 (Table 3.5).

![Figure 3.33](image)

**Figure 3.33. Binding of Oligonucleotides with Topoisomerase IA constructs;** (A) Oligonucleotides used for binding studies, STS sites are highlighted. ITC profile for binding of (B) NTD and (C) CTD of topoisomerase IA from *M. smegmatis* with Oligonucleotides of various lengths (With or without STS sites).

**Table 3.5. Binding affinity of oligonucleotides with Topoisomerase IA constructs from *M. smegmatis* and *E. coli.*

<table>
<thead>
<tr>
<th>Protein Constructs</th>
<th>Binding affinity (Ka) with 22-mer STS</th>
<th>Binding affinity (Ka) with 14-mer non-STS</th>
<th>Binding affinity (Ka) with dsDNA non-STS</th>
<th>Binding affinity (Ka) with overhang (22-mer STS+14-mer non-STS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD_Mstopol</td>
<td>7.9x10^7</td>
<td>No binding</td>
<td>1.56x10^6</td>
<td>1.41x10^8</td>
</tr>
<tr>
<td>CTD_Mstopol</td>
<td>1.96x10^7</td>
<td>8.01x10^7</td>
<td>6.86x10^6</td>
<td>1.97x10^8</td>
</tr>
<tr>
<td>NTD_Ectopol</td>
<td>4.26x10^6</td>
<td>1.54x10^6</td>
<td>3.06x10^5</td>
<td>1.17x10^8</td>
</tr>
</tbody>
</table>
We also performed the binding studies with the N-terminal domain from \textit{E. coli} with the same set of oligonucleotides to compare its binding affinities with the N–terminal domain of Mstopol (Table 3.5). In \textit{E. coli}, N-terminal domain is sequence non-specific but it also shows preference for oligonucleotides containing STS sites. The N-terminal domain of \textit{E. coli} binds with 22-mer STS containing ssDNA with high affinity of \(4.2 \times 10^6 \text{ M}^{-1}\) and with 14-mer ssDNA without STS, \(1.54 \times 10^6 \text{ M}^{-1}\). It binds with dsDNA with very low affinity. These results indicate that although NTD of \textit{E. coli} topoisomerase binds with ssDNA in sequence dependent manner but it shows preference for the sequences containing STS sites. The binding preference for STS containing DNA and distribution of STS sites indicate that NTD of \textit{M. smegmatis} and \textit{E. coli} might be quite similar in their mechanism of action.

\subsection*{3.2.8. Discussion}

Topoisomerases are involved in relieving topological stress of DNA and play crucial role in several crucial cellular processes. Unlike other members of topoisomerase IA family enzymes, Mstopol acts in sequence specific manner and acts on particular sequence motif (C/GTCT↓T) known as strong topoisomerase site (STS-site). This phenomenon could not be explained because of lack of structural details. Primary sequence alignment shows catalytic N-terminal domain is highly conserved and C-terminal domain is variable and lacks Zn finger motif, commonly found in C-terminal domain of other Topoisomerase IA. Binding studies shows CTD binds with ssDNA as well as with dsDNA in sequence independent manner while NTD binds to the oligonucleotides containing STS sequence motif with higher affinity and has no binding with non STS containing ssDNA. This indicates while N-terminal provides sequence specificity to the Mstopol, C-terminal may be involved in holding the non-scissile strand and double stranded region of DNA. We also compared the binding affinity of Mstopol NTD with Ectopol NTD with ssDNA (with and without STS) and dsDNA (With STS over-hang and blunt end). NTD_Mstopol clearly showed higher affinity for the sequences containing STS-sites(C/GTCT↓T) but \textit{E. coli} NTD also showed binding preference for STS containing ssDNA as compare to ssDNA without STS. We mapped the genomes of mycobacteria and some other eubacteria for the genome-wide occurrence of STS-sites and interestingly found that STS occurs
randomly throughout the genome of all these organisms. We also looked for the upstream promoter region for the occurrence of STS site throughout the genome of these organisms and it showed same level of occurrence. The sequence distribution for STS sequence showed clustering of these sites in the promoter region of highly expressing genes that are involved in nucleotide binding, transport activity and protein binding in these selected organisms. These striking similarities in binding pattern and distribution of STS sites in the genome of the selected organisms indicate that Topoisomerase IA from all these selected organisms would exhibit sequence specific DNA binding ability. The binding experiments with NTD of E. coli topoisomerase appear to corroborate this finding.