Journal of Applied and Natural Science
14(2), 531 - 542 (2022)
ISSN : 0974-9411 (Print), 2231-5209 (Online)
journals.ansfoundation.org

Research Article

Unveiling the structural features of CysE: a novel target for therapeutic interventions against persistent mycobacteria

Sunita Gupta
Department of Biotechnology, Jaypee Institute of Information Technology, A-10, Sector-62, Noida-201309 (Uttar Pradesh), India
Vibha Gupta*
Department of Biotechnology, Jaypee Institute of Information Technology, A-10, Sector-62, Noida-201309 (Uttar Pradesh), India

*Corresponding author. Email: vibha.gupta@jiit.ac.in

Article Info
https://doi.org/10.31018/jans.v14i2.3461
Received: April 9, 2022
Revised: June 1, 2022
Accepted: June 5, 2022

How to Cite
Gupta, S. and Gupta, V. (2022). Unveiling the structural features of CysE: a novel target for therapeutic interventions against persistent mycobacteria. *Journal of Applied and Natural Science*, 14(2), 531 - 542. https://doi.org/10.31018/jans.v14i2.3461

Abstract
World Health Organization (WHO) reports that one-third of the world's population is infected with a persistent form of *Mycobacterium tuberculosis* (*M.tb*), the causative bacterium responsible for causing the dreaded tuberculosis disease. Targeting mycobacterial persisters is important for achieving WHO's End TB target. The *de-novo* cysteine biosynthetic pathway is a novel target for addressing *M.tb* persistence. The two-step pathway comprises of serine acetyltransferase/CysE and O-acetyl-serine-sulfhydrylase/OASS/CysK. The present study is an attempt to understand the structural features of mycobacterial CysE by investigating the divergence amongst orthologous through phylogenetic analysis. Mapping of mycobacterial CysE sequences on the whole orthologous (COG1045) tree segregated the species into four clusters and several isoforms leading to their descendants identification. Interestingly the analysis revealed that the extended C-terminal α-helix believed unique to *M.tb* is also present in other organisms such as: *Campylobacter ureolyticus, Bacillus cereus, Geminocystis herdmanii* and *Paenibacillus borealis*. Further, the Hidden Markov model search against the whole Uniprot database suggests a plausible role of C-terminal α-helix of CysE in strengthening the substrate and/or co-factor binding. In addition, phylogenetic analysis of CysE sequences from the *Mycobacteriaceae* family facilitates grouping them under ten well-formed and six monophyletic clades, each based on characteristic features with respect to domain architecture, oligomeric assembly, C-terminal tetra-peptide tail, regulatory and feedback mechanism etc. Employing molecular phylogeny in conjunction with structural analysis has provided detailed insights for mycobacterial CysEs as drug target.

Keywords: Serine acetyltransferase, *Mycobacterium tuberculosis*, Phylogenetic analysis, HMM profile, 3D-structure

INTRODUCTION

Mycobacterium tuberculosis (*M.tb*) is the causative agent of one of the most devastating human diseases, estimated to kill 1.48 million people around the globe in 2020 (Global Tuberculosis Report, 2021). The pathogen is successful due to its ability to persist in human host for decades and can recur depending on the immune status of the patient. Tuberculosis (TB) scientific community believes that the central issue to be addressed for eradicating this disease is to target drug-resistant and persistent forms of infection. Regrettably, latent/persistent tubercle bacteria are resistant or tolerant to killing by antibiotics (Handwerger and Tomasz, 1985) and there is an urgent need to develop an anti-tubercular drug with strong bactericidal activity against latent bacilli positioned in the hypoxic areas of granulomas. In this context, quite a few amino acid biosynthesis pathways have been portrayed to be essential for the survival and pathogenesis of *M.tb* (Yelamanchi and Surolia, 2021). One such pathway is the de-novo cysteine biosynthetic pathway, an essential metabolic pathway responsible for producing cysteine from inorganic sulfur (Nakamori *et al.,* 1998; Smith and Thompson, 1969).

The produced cysteine is incorporated in biomolecules such as proteins, coenzymes, and mycothiols, a crucial metabolite for maintaining the redox-defense mechanism, thus facilitating the adaptation and survival of pathogen (Jean Kumar *et al.,* 2013). The reference pathway has been found to be upregulated during persistent tuberculosis infection, indicative of

its importance for the bacterium's successful persistence in its host. Moreover, cysteine is produced by a different route in humans and the de-novo cysteine biosynthetic pathway enzymes are absence is humans making the pathway a more lucrative target for antitubercular drug development (Rengarajan *et al.,* 2005; Sassetti and Rubin, 2003).

The cysteine biosynthetic pathway involves two enzymes, serine acetyltransferase (SAT/CysE; EC: 2.3.1.30) which catalyzes the production of O-acetyl-L-serine (OAS) from acetyl-CoA and L-serine that further gets converted to L-cysteine by O-acetyl serine sulfhydrylase (OASS/CysK; EC: 2.5.1.47) enzyme (Devayani *et al.,* 2008; Hampshire *et al.,* 2004). CysE belongs to the O-acetyltransferase subfamily and is characterized by the presence of a unique hexapeptide repeat sequence ([LIV]-[GAED]-X_2-[STAV]-X) that results in the formation of a left-handed parallel β-helix (LβH) domain at the C-terminal and an N-terminal α–helical domain (SATase_N) (Johnson *et al.,* 2004). The catalytic site is present in the cleft between the two monomers where the transfer of acetyl moiety from acetyl-CoA to L-serine takes place. The multifaceted CysK is a pyridoxal 5'-phosphate (PLP) dependent enzyme that catalyzes the β-substitution of OAS with disulphide to produce acetate and L-cysteine (recently reviewed by Joshi *et al.,* 2019). This enzyme is a homodimer, where each subunit comprises ~310 amino acids and is composed of two domains, each adopting an α/β fold. Crossover of 1 -143 residues of the smaller N-terminal domain into the larger C-terminal domain of each subunit creates an active site pocket buried deep within the protein. The interface of the two domains also houses the coenzyme (PLP) binding site covalently bonded via ε-amino group of K44 residue of *M.tb* CysK. The 3-dimensional crystal structure (PDB-ID: 2Q3C) (Schnell *et al.,* 2007) along with some inhibitors have been reported for *M.tb* CysK enzyme (Poyraz *et al.,* 2013; Ullas *et al.,* 2013), but other than a model structure (Gupta and Gupta, 2020), mycobacterial CysE has not been characterized structurally and functionally till date. Nevertheless, the crystal structures of CysEs from other homologues such as *Escherichia coli* (PDB-ID:1T3D) (Pye *et al.,* 2004), *Haemophilus influenzae* (PDB-ID:1S80,1SSQ) (Gorman and Shapiro, 2004; Olsen *et el.,* 2004), *Entamoeba histolytica* (PDB-ID:3P1B) (Kumar *et al.,* 2011), *Brucella abortus* (PDB-ID:4HZC) (Kumar *et al.,* 2014), *Brucella melitensis* (PDB-ID:3MC4), *Vibrio cholerae* (PDB-ID:4H7O), *Yersinia pestis* (PDB-ID:3GVD), *Bacteroides vulgates* (PDB-ID:3F1X) *Glycine max* (PDB-ID:4N69) (Yi *et al.,* 2013) and *Klebsiella pneumonia* (PDB-ID:6JVU) (Verma *et al.,* 2020) contribute to structure function understanding of this gateway enzyme in the *de novo* cysteine biosynthetic pathway. CysEs exist as hexamer except for *E. histolytica* which supports a trimeric form (Kumar *et al.,* 2011), and *B.abortus* which exists in both hexameric and trimeric forms (Kumar *et al.,* 2014).

The N-terminal region is involved in homo-dimerization of the two trimers to form a hexameric complex whereas the C-terminal takes part in hetero-oligomerization with CysK to form a functional cysteine synthase complex (CSC) for regulating *de-novo* cysteine biosynthesis. For CSC formation, the last C-terminal tetrapeptide from CysE competes with OAS to bind to active site of CysK and inhibit its activity (Huang *et al.,* 2005). In most of the organisms, *M.tb* (DFSI), *G.max* (DYII), *E.coli* (GDGI), *B. abortus* (GDGI), *Y. pestis* (GDGI) and *H.influenza* (NLNI), the terminal amino acid of the tetrapeptide is isoleucine and is found crucial for complex formation. *E. histolytica* SAT1(SPSI) is an exception where despite the presence of Ile, it does not form CSC complex with CysK thereby switching off a regulatory mechanism and ensuring cysteine availability in the organism for its survival (Kumar *et al.,* 2011). In our previous study, we pointed that *M.tb* CysE differs from other homologs by virtue of an extended C-terminal and a shorter N-terminal that supports a trimeric form rather than hexamer. Additionally, we also predicted an α-helix at the C-terminal, which till date has, only been reported in *M.tb,* but its exact role remains unknown (Gupta and Gupta, 2020). In lieu of the above evidences, the present study proposes that the length of the N- and C-terminal domains, the characteristic of the C-terminal tetrapeptide tail and the mutations decide the oligomeric state as well as the regulatory mechanisms of the enzyme (Kumar *et al.,* 2013; Kumar *et al.,* 2011). This leverages a scope to study the evolutionary tie among CysEs from different descendants, and to understand the lineage-specific distribution along species.

The evolutionary and comparative studies rely on the analysis of orthologous genes (homologous genes that diverged at a speciation event), and thus on their robust annotation (Trachana. *et al.,* 2014). Orthologs tend to share comparable molecular functions. Therefore, orthology identification has served as a pivotal technique for transferring knowledge from experimentally annotated enzymes in model species to other unknown species (Altenhoff *et al.,* 2012). Moreover, identifying orthologous groups is useful for genome annotation, evolutionary conservation, comparative genomics, and studying the variability in molecular sequences (Jensen *et al.,* 2008; Li *et al.,* 2003). Tatusov *et al*. launched the database containing orthologs clusters such as Cluster of Orthologous groups/euKaryotic Orthologous Groups (COGs/KOGs), which systematically represent the relationships between genes found in different species (Tatusov *et al.,* 2001; Vasudevan *et al.,* 2003). Many such sources of orthologous groups are available such as EggNOG (Jensen *et al.,* 2008), OrthoDB (Kriventseva *et al.,* 2015), OrthoMCL (Li *et al.,* 2003), Inparanoid (Berglund *et al.,* 2008), Ensembl Compara

(Hubbard *et al.,* 2007), OMA (Roth *et al.,* 2008), etc. However, inadequate annotations of the identified orthologous groups hinder holistic interpretation of subsequent results (Jensen *et al.,* 2008). EggNOG 5.0 database ('evolutionary genealogy of genes: Non-supervised Orthologous Groups') is a public resource in which thousands of genomes are analyzed at once to establish orthology relationships between all their genes. The orthologous groups are constructed from Smith–Waterman alignments through the identification of reciprocal best matches and triangular linkage clustering (Huerta-Cepas *et al.,* 2019). Due to its added advantage of performing functional annotations for the orthologous groups (Jensen *et al.,* 2008), eggNOG was used to carry experiments performed in the present study, that aims to understand the probable role of C-terminal α-helix in mycobacterial CysE.

MATERIALS AND METHODS

CysE ortholog analysis and taxonomic distribution

In an attempt to understand the distribution of CysE enzyme across species, an eggNOG 5.0 database (http://eggnog5.embl.de/#/app/home) search was performed (Huerta-cepas *et al.,* 2019). EggNOG5.0 is a public database of orthology relationships, gene evolutionary histories and functional annotations for eukaryotes, prokaryotes and viral sequences. It allows users to further explore the history of speciation and duplication events within each orthologous group (OG), infer pairwise orthology relationships between specific species, and trace functional changes therein.

Sequence retrieval and multiple alignment studies

The presence of a unique insertion in CysE from *M.tb* fueled the curiosity to identify orthologous genes on the basis of functionality, i.e., the sequences that diverge from common ancestry as the result of a speciation event. The eggNOG5.0 database was used to generate orthologous groups of sequences. It yielded 5,094 orthologous sequences covering 3570 bacterial species (COG1045). The sequences belonging to COG1045 varied from ~40 to 380 amino acid residues. The complete set of sequences was aligned globally using MAFFT v7.3.10 (https://mafft.cbrc.jp) (Katoh *et al.,* 2013) progressive method algorithm along with normalized similarity matrix using default parameters and was further visualized using Jalview 1.8.3 (http://www.jalview.org) (Waterhouse *et al.,* 2009). Keeping in mind the 229 amino acid length of the *M.tb* CysE enzyme, another independent sequence set was prepared with 220-240 amino acid long sequences considered for local alignment (531 sequences) using MAFFT v7.3.10. The sequences that led to a patchy alignment were removed manually, resulting in 482 eligible sequences further aligned globally using default parameters.

Conservation pattern using HMM profile

The conserved part of the aligned sequences in the second set (482 sequences) was seeded to generate the Hidden Markov Model (HMM) profile by using hmmbuild module of the HMMER3.1b2 (http://hmmer.org/) package, with default parameters (Eddy, 2009). The hmmsearch module was searched against the whole Uniprot database (5,61,568 sequences; dated 9-03-2020) at an E-value cut-off of 0.1. The hmmsearch program allows each sequence to pass through three scoring algorithms, namely- MSV, Viterbi, and forward, thereby enhancing the accuracy of sequence selection. Sequences with the highest similarity with the profile were given a domain score and an 'Expect value' (E) by HMMER program. The E-value of a sequence with a score z indicates the number of sequences that are expected to score z by chance when searching a sequence database with the given size.

Phylogenetic tree of COG1045

All bacterial orthologous sequences (COG1045; 5094 in number) were retrieved from the eggNog 5.0 database, followed by their global alignment using MAFFT program. The alignment file was seeded to generate a phylogenetic tree using the FastTree 2.1 program that uses the CAT approximation to model rate heterogeneity across different sites. It deploys the Jones-Taylor-Thorton (JTT) models for amino acids evolution and can handle alignments for a million of sequences speeding up to 100-1000 folds than other methods. It also uses a heuristic variant of neighbor-joining method and stores profiles instead of the matrix to reduce the memory requirements (Price *et al.,* 2010). To visualize the phylogenetic tree, Archeopteryx package was used (Han and Zmasek, 2009).

Sequence retrieval and phylogenetic studies of CysE from *Mycobacteriaceae* family

The CysE enzyme was searched in the Uniprot database (https://www.uniprot.org/) dated 14 January 2021 to retrieve all the sequences along with the *Mycobacteriaceae* family (289 sequences) (Bateman *et al.,* 2015). The identical set of sequences was removed from the dataset and the remaining set was aligned using MAFFT v7.3.10 global alignment, using maximum iteration 1000. The phylogenetic tree was constructed using the above alignment file via FastTree 2 and visualized using the Archaeopteryx package.

Structural prediction analysis

To further classify the CysE enzyme of diverse phylogenetic clades into structural groups, the variation patterns at the residue level in the alignment file were thoroughly analyzed and unique sequences corresponding to different clades were fetched out. Structural prediction of few representative sequences of unique

clades and a few monophyletic clades with the unknown 3D structure were performed using an iterative threading assembly refinement (I-TASSER) server, which deploys threading approaches by identifying the template from PDB (https://zhanggroup.org/I-TASSER) (Zhang, 2009; Yang *et al.,* 2014). The 3D-structures of the modeled enzymes were minimized by employing the steepest descent (100 steps), conjugate gradient (20 steps) algorithm and the ff99SB force field in UCSF Chimera 1.11 (Pettersen *et al.,* 2004). Following energy minimization, several protein structure validation tools, such as ERRAT (Colovos and Yeates, 1993) and ProSA (Wiederstein and Sippl, 2007) were used to identify errors in predicted protein structures. ERRAT gives an "overall quality factor" for non-bonded atomic interactions, with higher scores indicating higher quality. The generally accepted range is >50 for a high-quality model. For the current 3D models, the overall quality factor predicted by the ERRAT server was in the range of 80-95%. The ProSA program analyses the interaction energy of each amino acid of the predicted structure and authenticates the stability of the structure on the basis of z-score and energy.

RESULTS AND DISCUSSION

CysE ortholog analysis and taxonomic distribution

The taxonomic distribution of CysE along different taxa is shown in Fig. 1. Off all CysEs reported till date, only 2% is distributed amongst Archea, with the majority of it lying in the bacterial kingdom. Further bifurcation of the bacterial kingdom gave a maximum share to Proteobacteria (43.0%), followed by Fermicutes (25.7%), Actinobacteria (11.3%) and so on.

Multiple sequence alignment

In our previous study, we proposed the presence of an additional structural element at the C-terminal end of *M.tb* CysE, forming an α-helix (Gupta and Gupta, 2020), apart from the known SATase_N and LβH_SAT domain. The MAFFT global alignment of all the 482 sequences having 220-240 amino acids shows that many other species possess an additional α-helix at the C-terminal. Few representative sequences with their extended C-terminal region are depicted in Fig. 2. The top most sequence with Uniprot-ID P95231 is of *M.tb* CysE (H37Rv) and the red boxed residues (^{187}VGASLDSLLTRVARLEALGGGPQAA^{211}) possess the region having C-terminal α-helix. This extensive sequence alignment divulges that a clear pattern of an amphipathic helix is present in other orthologs such as *Campylobacter ureolyticus, Bacillus cereus, Geminocystis herdmanii, Blautia product, Bacillus thermoamylovorans* and *Paenibacillus borealis*, although it is not conserved at the residue level. The ensuing section investigates the possible role of this extended C-terminal helix present in some CysEs.

Conservation pattern using HMM profile

Searching the HMM profile of CysE C-terminal sequence-specific motif across the Uniprot database revealed that apart from serine acetyltransferases, few other transferases have similar motifs, namely WCAB (Putative colanic acid biosynthesis acetyltransferase), DAPH(2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase), GLMU (Bifunctional acetyltransferase/uridyltransferase), LPXD (UDP-3-O-acylglucosamine N-acyltransferase) and DAPD(2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase).The E-values of respective hits were 9.4e-29, 1.5e-17, 5.6e-05, 4.8e-08,and7.4e-04 respectively. The presence of an α-helix at the C-terminal regions of *E.coli* DapD tetrahydrodipicolinate N-succinyl transferase (THDPDST) as opposed to *M.bovis* homolog, is found to be responsible for significant conformational changes upon acetyl-CoA and substrate binding. In the apo-form, the C-terminal α-helix blocks the acetyl-CoA binding site, whereas, in acetyl-CoA and substrate-bound form, it undergoes large rearrangement and contributes to substrate co-operativity (Nguyen *et al.,* 2008). Another independent study on *M.tb* GlmU enzyme which possesses additional 24 amino acids at the C-terminal end (some of which form helix α9) as compared to *S.pneumoniae* and *E.coli* homologues, is reported to provide additional stability to the biological trimer assembly along with acetyl-CoA binding site through domain swapping mechanism (Zhang *et al.,* 2009).

Extensive structural analysis of the extended C-terminal α-helix observed in *M.tb* CysE revealed its strong amphipathic characteristic, which forms a number of hydrophobic interactions with the LβH domain of the adjacent subunit (Val151 and Leu201, Ile136 and Leu204). Additionally, the hydrophilic surface residue Glu202 also forms a salt-bridge with the residue Lys152. These residues reside at the periphery of the acetyl-coA binding site and their engagement may alter the specificity of the cofactor towards the apo-enzyme whereas, in bound form C-terminal α-helix may undergo rearrangement upon substrate/cofactor binding and possibly strengthen their binding or may provide stability to the quaternary assembly.This reinforces our previous report regarding the inability to build a trimeric assembly of full-length *M.tb* CysE due to steric clashes displayed by the extended C-terminal stretch, which could only be achieved after deleting the last 44 residues of the enzyme (Gupta and Gupta, 2020).

Phylogenetic tree analysis of COG1045

The complete set of bacterial orthologous sequences (COG1045; 5094 sequences) were aligned using MAFFT global alignment program. The alignment file

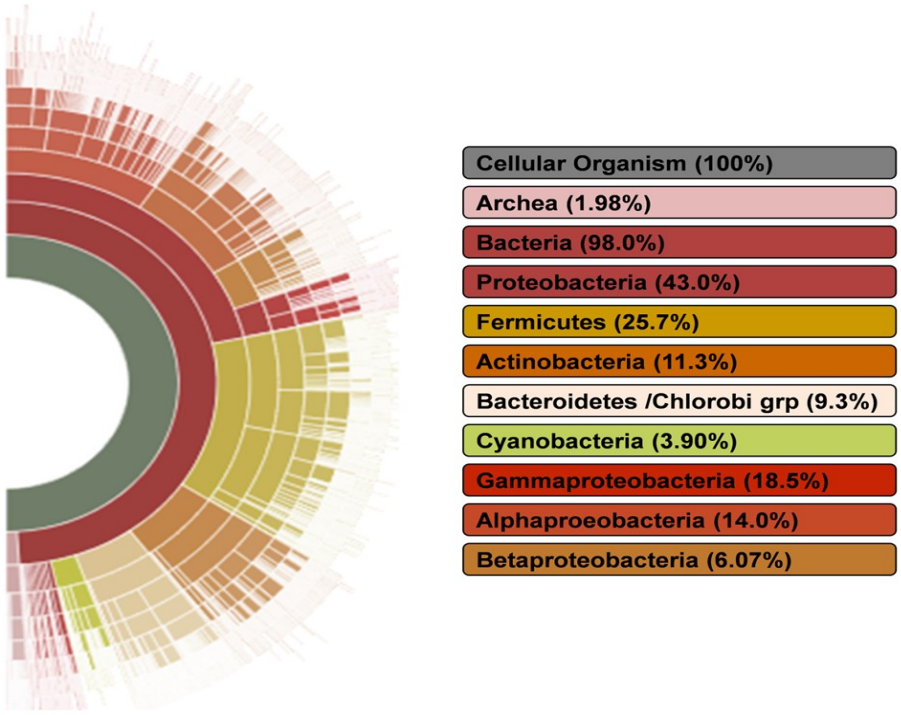

Fig. 1. *Taxonomic distribution of all known CysEs produced using eggNOG 5.0. Different colors indicate distinct taxa*

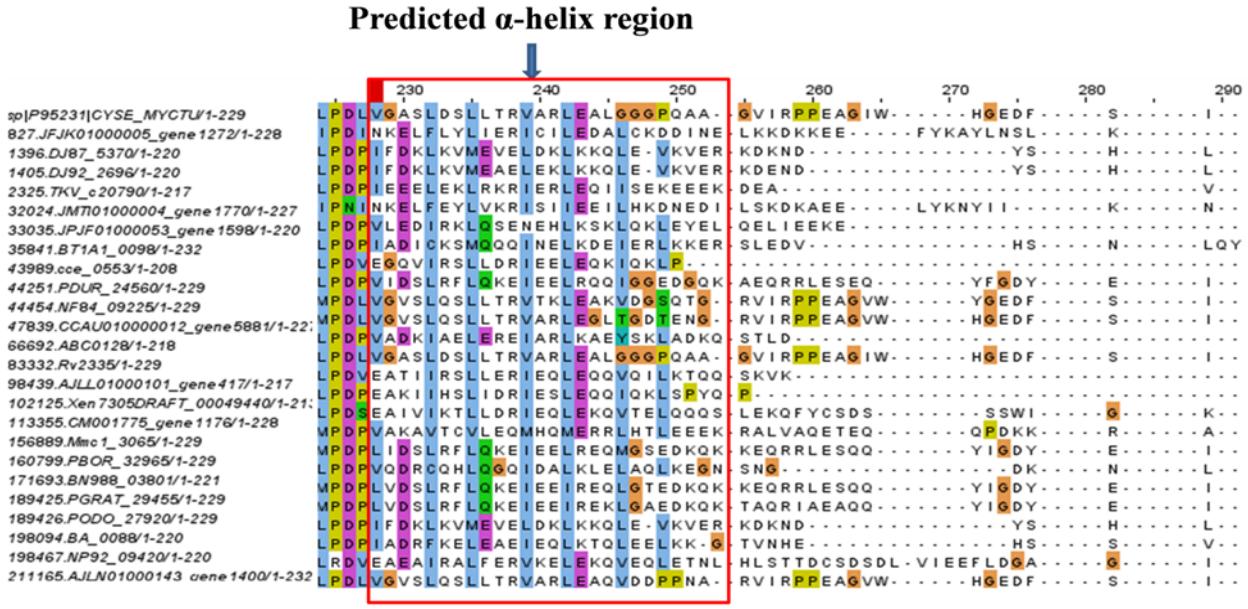

Fig. 2. *MAFFT global alignment of the orthologous sequences depicting the presence of amphipathic α-helical pattern throughout. The topmost sequence with Uniprot-ID P95231 is of M.tbCysE (H37Rv)*

was seeded to generate a phylogenetic tree using FastTree 2 program which depicts four different clades in which *Mycobacteriaceae* sequences lie (Fig. 3B-3E). Fig. 3A represents the main tree in circular form (blue) and the clades highlighted in different colors depicting *Mycobacteriaceae* sequences. The inset view of the clade with a maximum number of *Mycobacteriaceae* sequences is shown in Fig.3B, which includes most pathogenic mycobacteria, including *M.tb* (H37Rv/ H37Ra), *M.leprae, M.marinum, M.ulcerens, M.kansasii, M.triplex, M.genavense, M.avium, M.scrofulaceum, M.intracellulare, M.colombiense* and *M.asiaticum*. CysE from above *Mycobacteriaceae* species are most closely related to CysEs of *Frankia sp.,* corroborating with the known evolutionary linkage (Gao and Gupta, 2012). Few acidophilic (*F. acidiphilum* and *A. ferrooxidans*) and thermophilic (*A. thermophila* and *A. aerophila*) bacteria are also closely related to mycobacteria belonging to the above clade. The clade 3C includes other mycobacteria species, namely, *M.rhodesiae,*

M.aromaticivorans, M.thermoresistibile and *M sp.* (gene -id: 875328.JDM601_1668, 1370120.AUWR010 00022_gene2776, 1172186.KB911468_gene910 and 1449048.JQKU01000008_gene918). The sequences corresponding to CysEs of all the above organisms of this clade are shorter than the *M.tb* sequence with terminal residues AMYI in place of DFSI. As shown in Fig. 3C the sequences are closely related to *S.rotundus* and *S.rugosus.* These two species are distinguished from other Actinomycetes due to their unique mycolic acid pattern and low G+C content as compared to mycobacteria. They are involved in pulmonary infection, bronchiolitis, pneumonia and are often misidentified as mycobacteria (Keikha, 2018). The next interesting clade (3E) consists of only a single mycobacterial CysE sequence corresponding to *M.smegmatis* as an out-group. This is a non-pathogenic bacteria with elongation (~22 amino acids) at the N-terminal and shorter C-terminal as compared to mycobacterial CysE. The terminal tetrapeptide is DAQA and is evolutionarily related to Rhodococcus species. The last clade 3D consists of species like *M.kansasii, M.colombeinse, M.marinum, M.ulcerans* and *M.asiatucum*. They seem to be false negatives which are way shorter (amino acids ~186) than the pathogenic isoforms of the same organisms. The C-terminal amino acid is not isoleucine in any of the sequences and this clade is closer to *N.concave* and *T.sulfidiphilus.* These results indicate the segregation of *Mycobacteriaceae* species along whole bacterial orthologs, their isoforms and the related descendents.

Sequence retrieval and Phylogenetic studies for mycobacterial CysE enzyme

The 289 protein sequences corresponding to *Mycobacteriaceae* family were retrieved from the Uniprot database in FASTA format, of which duplicate sequences were removed, reducing the final count to 240 sequences. These were later aligned using MAFFT global alignment to generate a phylogenetic tree using FastTree 2 (Fig. 4). The phylogenetic tree is segregated into ten well-supported clades (I-X) and six monophyletic clades (a-f) based on their sequence length and architecture. Here, the main tree is divided into three sub-trees; the first sub-tree possesses two sequences (*M.sp.1245111.1* (Uniprot-ID A0A1A3Q6X8) and *M.sp*.E1386 (Uniprot-ID A0A1A2YP50) (clade-I), the second sub-tree is a monophyletic clade of *M.sp*.E2497 (Uniprot-ID A0A1A2X3E1) is an outgroup (a). The third sub-tree is the largest one and includes rest of the sequences. A monophyletic clade of *M.bohemicum* is another outgroups (b). The Uniprot data consists of three isoforms of *M.tb* CysE sequences; first, one is more similar to *M.colombeinse* and lies in the pink clade (X). The rest two differ with a point mutation (resid-202 E/D) that lies in the common clade with other pathogenic species namely, *M.orygis, M.canetti, M.bovis,* and *M.decipiens* (Cyan) (clade-II). The adjacent clade in blue color (clade-III) appears to be more divergent than the rest of the clades and the sequence length varies from 181-186 with terminal residues different from isoleucine in any of the sequences. The sequence *M.sp*.KBS0706 (Uniprot-ID A0A554U6H0) is an outlier with 282 amino acids and DFVI as the terminal tetrapeptide (c). Another small clade (IV) appeared, having two closely related sequences *M.sphagni* and *M.sp*.ELW1 (Uniprot-ID A0A5C1R8R4) having sequence length of 229 amino acids. The sequence belonging to the black clade (V) is also diverged, length ~194 amino acids and terminal residues being AMYI. A few representative species of this clade are *M. aromaticivorans, M. thermoresistibile, M .algericus, M. aichiense, M. holsaticum* and *M. insubricum. M. smegmatis* CysE has a much diverged sequence (203 amino acids) and DAQA as the terminal residues that are evolutionary closed to the above species and have emerged as a monophyletic clade (e). The monophyletic clade (f) represents *M. mucogenicum* CysE sequence having 189 amino-acids and FYVI as the terminal tetrapeptide tail. The grey clade consists of sequences from *M. heidelbergense* and *M. lacus* with an extra insertion at the C-terminal (Fig.5). The clades in yellow, green, orange and pink are not very diverse and possess CysE sequences of ~229-235 amino acids with DFSI as the terminal tetrapeptide. The phylogenetic tree discussed above led us to select 26 representative sequences (two from each clade and all monophyletic clades). Fig. 5 depicts the multiple sequence alignment of the selected sequences using Clustal Omega (Sievers and Higgins, 2014). The residues involved in the binding of L-serine and acetyl-CoA are highlighted in red and green color respectively. The rectangular box (violet) represents the region where C-terminal α-helix is being reported. The N-terminal of *M.sp*.KBS0706 CysE sequence is ~89 amino acids lengthier and ~36 amino acids shorter at C-terminal than the *M.tb* CysE. This suggests that this *mycobacteriaceae* specie will form eight α-helices at the N-terminal, contrary to *M.tb* CysE, which forms only four. It will also not possess the extra α-helices at the C-terminal. The presence of additional 4 α-helices at the N-terminal will favor its oligomeric assembly as hexamer, different than that of the trimeric assembly proposed for the *M.tb CysE* (Gupta and Gupta, 2020). In the above alignment figure, three sequences (*M.smegmatis, M.branderi and M.kyorinense*) reflect the diversity in sequence length as well as architecture. The lack of isoleucine as the terminal amino acid suggests that these three species will not form CSC. The conservation of active site residue is less for *M.branderi and M.kyorinense*, along with a much shorter C-

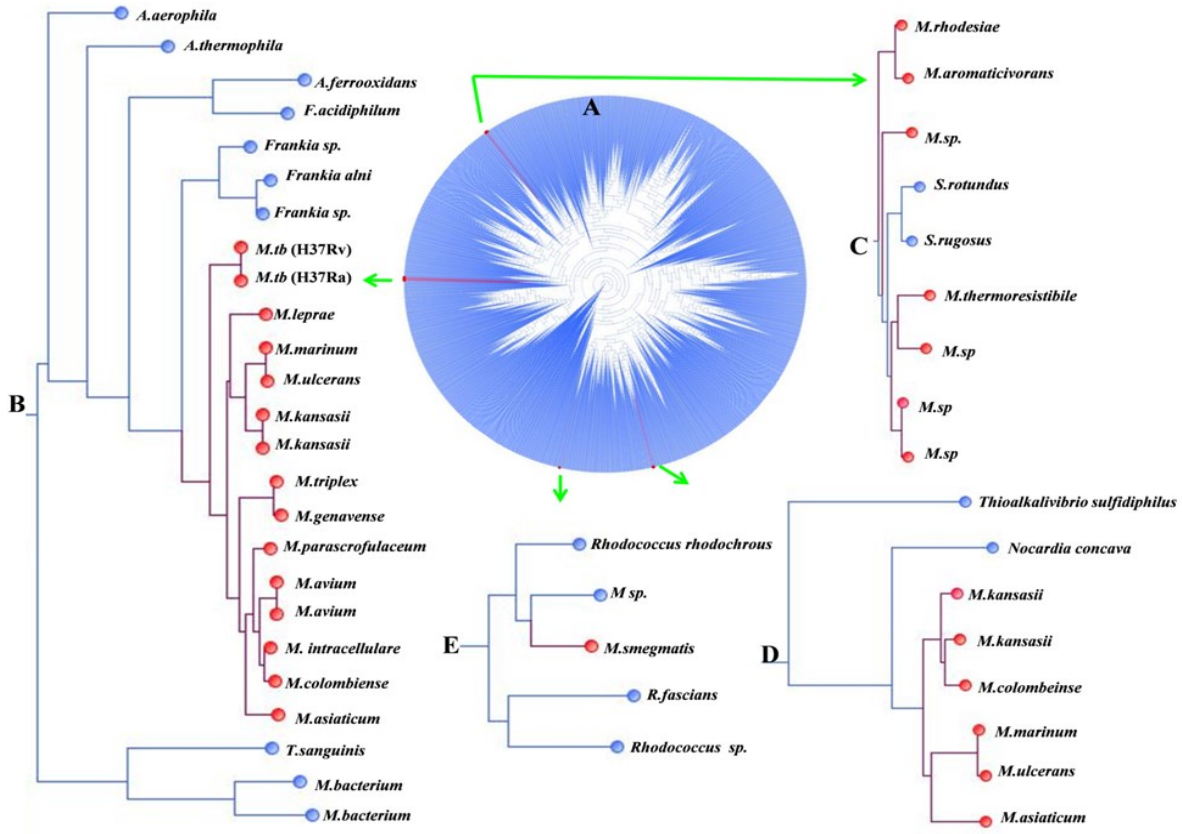

Fig. 3.*(A) Mapping of Mycobacteriaceae sequences on the phylogenetic tree of the bacterial sequences of COG1045. The circular representation of the phylogenetic tree covering all the sequences are shown in blue color and the sub-tree possessing Mycobacteriaceae species are highlighted in pink. (B-E) represents the inset view of the clades where Mycobacteriaceae sequences are lying. The Mycobacteriaceae and its descendents nodes are shown in red and blue color, respectively*

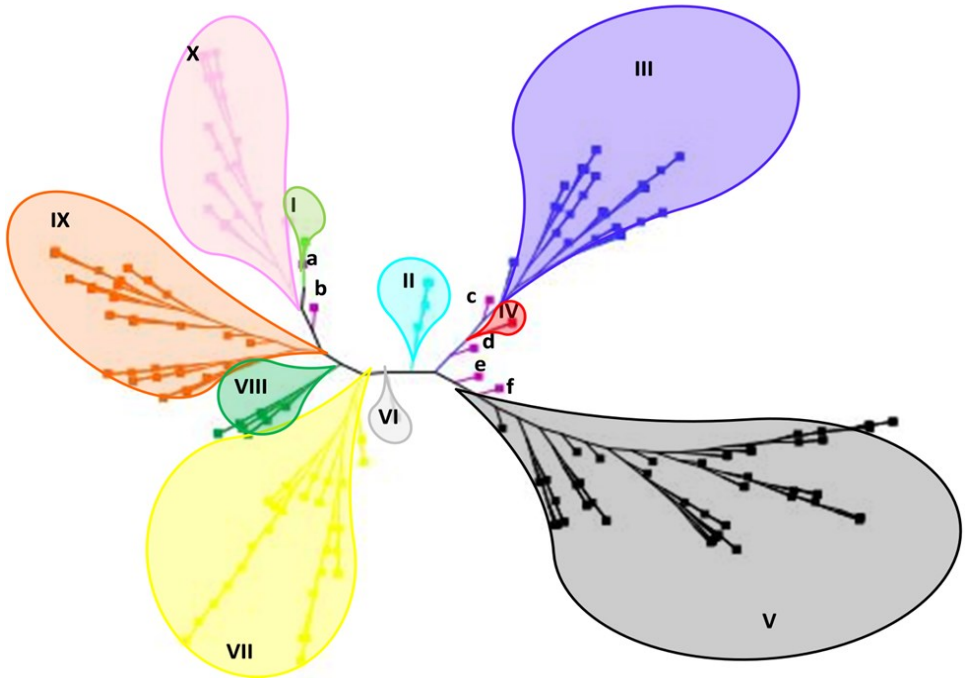

Fig. 4. *Fern representation of the phylogenetic tree for 240 Mycobacteriaceaesequences retrieved from Uniprot database. (I-X) represents the ten major clade and (a-f) represents the six monophyletic clades- a (M.sp.E2497), b (M.bohemicum), c (M.sp.KBS0706), d (M.sp.:Uniprot-Id-A0A5C7WWH0), e (M.smegmatis), f (M.mucogenicum 261Sha1)*

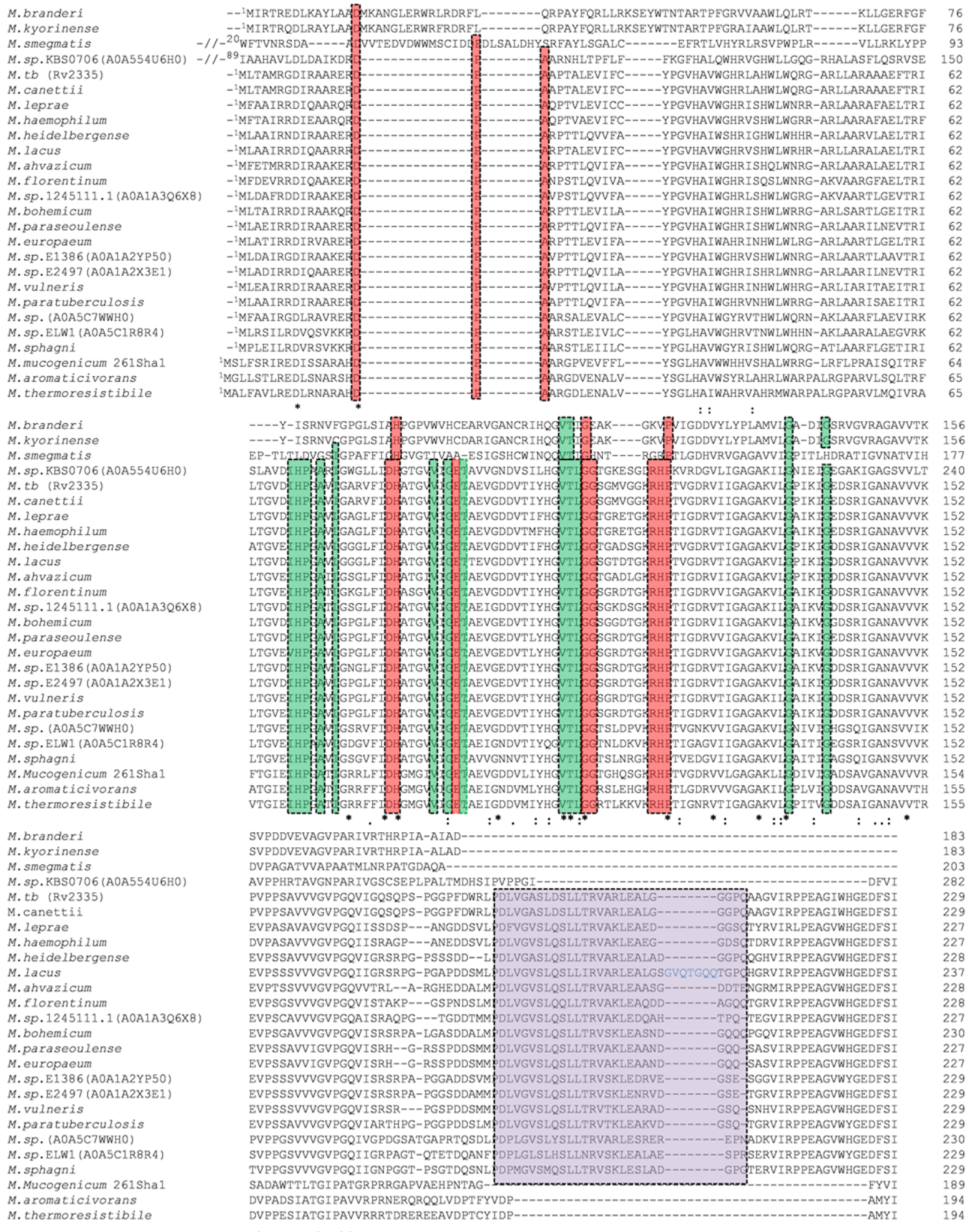

Fig. 5. *Multiple Sequence Alignment (MSA) of twenty-six representative sequences from the Mycobacteriaceae phylogenetic tree. The conserved substrate and acetyl-coA binding residues are highlighted in red and green colored boxes. The asterisks at the baseline of the alignment indicate identical amino acids in a given sequence position, while double and single dots refer to highly and moderately conserved (chemically similar) residues, respectively. The rectangular box (violet) represents the region where C-terminal α-helix is being reported.*

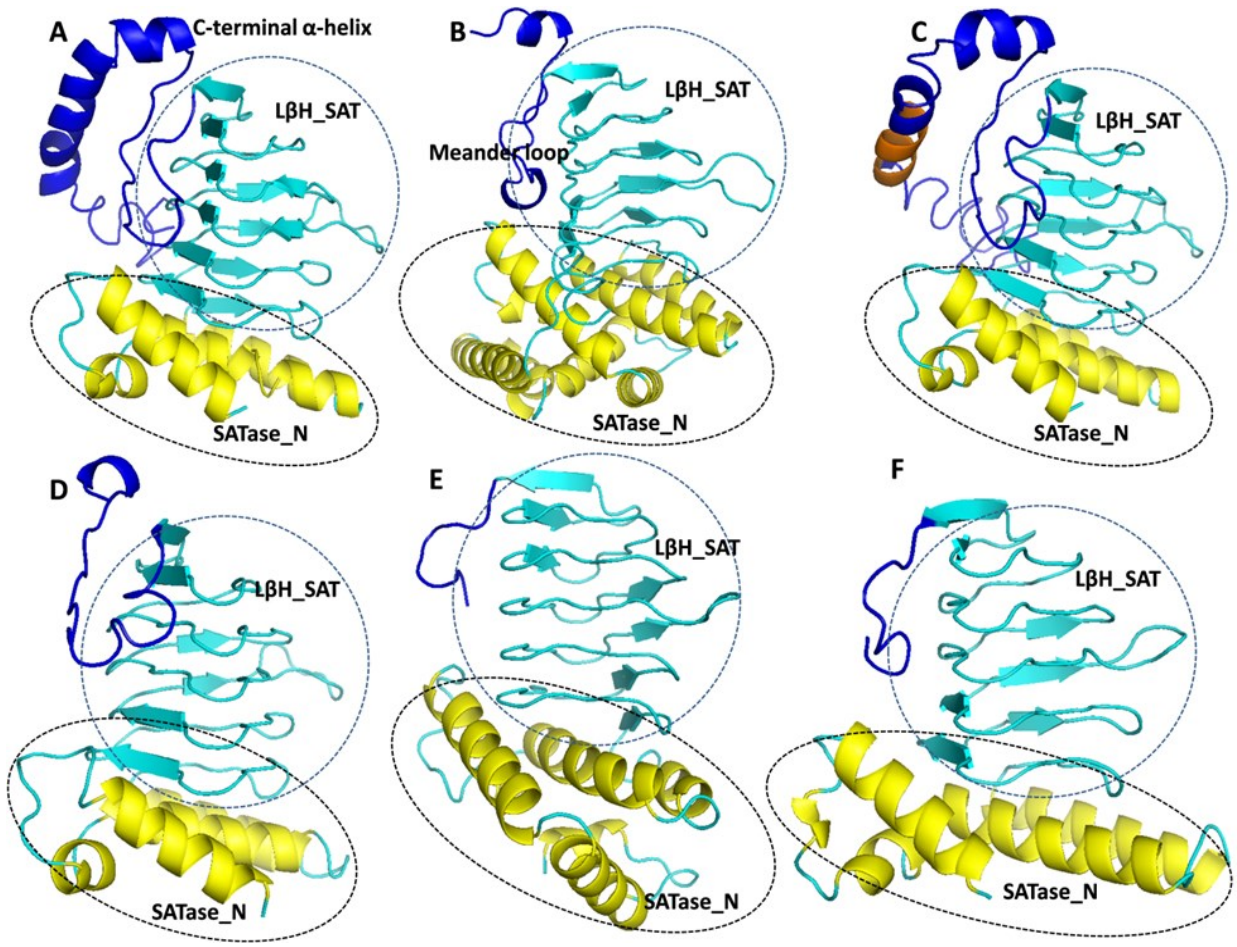

Fig. 6. *3-D modeled structure of six unique species of Mycobacteriaceae family: A (M.tuberculosis-H37Rv), B (M.sp.KBS0706), C (M.lacus), D (M.aromaticovorans), E (M.smegmatis) and F (M.branderi). The SATase_N and LβH domain is shown in yellow and cyan color, respectively. The terminal region after LβH is shown in blue color, which appears differently in all the species*

terminal, making clade-III the most diverse. The two representatives of clade-V (*M.aromaticivorans and M.thermoresistibile*) are shorter by ~30 amino acids at the C-terminal, which concludes that the species lying in clade-V will not possess additional α-helices at the C -terminal but may involve in CSC regulatory mechanism. The sequences across our selection showed lots of variation at the residue level; hence, structural analysis of the few unique sequences revealed diverse structural domains dictating its oligomeric assembly as well as the regulatory mechanism.

Structural prediction analysis

The alignment file was analyzed thoroughly for the residue level variation pattern, which led us to select six unique sequences (*M.tb*, *M.sp*.KBS0706, *M.lacus*, *M.aromaticovorans, M.smegmatis*, and *M.branderi*) for structure prediction analysis (Fig. 6A-6F). All the six monomer structures have the LβH domain intact and only variation is observed at the N and C-terminal ends. Among all the structures, only *M.tb* and *M.lacus* CysE possess an extra helix at the C-terminal (Fig. 6A and 6C). Later, one also possesses an insertion (orange color) at the C-terminal helical portion. All the *Mycobacteriaceae* species have shorter N-terminal, except for *M.sp*.KBS0706, which has ~89 amino acids longer N-terminal shaping up into eight α-helixes and a meander loop at the C-terminal (Fig. 6B), resembling *E.coli* CysE structurally, though it shows 95% sequence identity to the CysE of *Inquilinus limosu*s, which is a slow-growing, gram-negative and non-fermentative bacillus isolated from the respiratory tracts of patients with cystic fibrosis. The structure of *M.aromaticovorans* is similar to that of *M.tb* CysE except for lacking α-helix at C-terminal (Fig. 6D). The *M.smegmatis* (Fig. 6E) and *M.branderi* (Fig. 6F) possess the most diverse sequence due to which N-terminal helical alignment is very different from the rest of the species, which may alter its oligomeric assembly; also, the absence of isoleucine at the C-terminal tetramer tail will not allow it to form CSC complex hindering its regulatory mechanism. Thus, this structural prediction analysis has given a clearer picture of understanding the different structural organizations among the CysE of *Mycobacteriaceae*

species, predicting their oligomeric assembly and possible regulatory mechanism.

Conclusion

The sequence and phylogenetic analysis of the whole ortholog sequences helped in understanding the evolutionary relationship of the mycobacterial species to the other orthologs and their descendants. This work categorically shows for the first time presence of C-terminal α-helix in other orthologs such as *Campylobacter ureolyticus, Bacillus cereus, Geminocystis herdmanii* and *Paenibacillus borealis* other than exclusively observed in *M.tb* CysE. Further, Hidden Markov Model (HMM) profile search against the whole Uniprot data revealed GLMU (Bifunctional acetyltransferase/ uridyltransferase), LPXD(UDP-3-O-acylglucosamine N-acyltransferase) and DAPD (2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase) as couple of hits with similar α-helical extensions, which undergo rearrangement to strengthen the binding of substrate (L -ser) and/or co-factor (acetyl-CoA) or trimeric association. Applying the same to *M.tb* CysE, the possible role of its extended C-terminal α-helix appears in strengthening the binding of substrate (L-ser) and/or co-factor (acetyl-CoA) or trimeric association. The phylogenetic tree created solely from *Mycobacteriaceae* CysE segregates the species into ten well-formed clades and six monophyletic clades on the basis of diverse sequence length and C-terminal tetra-peptide tail.Thus, executing molecular phylogeny in combination with structural study has provided insights into the relatedness of CysE drug target across species in general and of *Mycobacteriaceae* family, in particular, that may pave the way for the rational design of species-specific lead molecules.

ACKNOWLEDGEMENTS

This complete work has been done during the Covid-19 global pandemic, which would not have been possible without the exceptional support of Prof. Andrew M Lynn (School of Computational and Integrative Sciences, Jawaharlal Nehru University) for his unstinting and unconditional generosity in providing access to computational resources via remote login. SG and VG express their heartfelt gratitude for his expert advice and constant encouragement in the successful culmination of this study. SG wishes to acknowledge the research grant from the Department of Science and Technology (DST), Govt. of India for support throughout WOS-A program (SR/WOS-A/LS-1442/2015). Special thanks to Dr. Poonam Vishwakarma for technical help and brilliance in the protocols. SG also thanks Monika and Deepansh Mody for helping in organizing the data and critical acumen of the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

1. Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. & Dessimoz, C. (2012). Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology*, *8* (5), 1-10 https://doi.org/10.1371/journal.pcbi.1002514
2. Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., …& Zhang, J. (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, *43* (D1), D204–D212. https://doi.org/10.1093/nar/gku989
3. Berglund, A. C., Sjölund, E., Östlund, G. & Sonnhammer, E. L. L. (2008). InParanoid 6: Eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research*, *36*(SUPPL. 1). https://doi.org/10.1093/nar/gkm1020
4. Colovos, C. & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, *2*(9), 1511–1519. https://doi.org/10.1002/PRO.5560020916
5. Devayani P. Bhave, Wilson B. Muse III. & Kate S. Carroll. (2008). Drug targets in mycobacterial sulfur metabolism. *Infectious Disorders - Drug Targets*, *7*(2), 140–158. https://doi.org/10.2174/187152607781001772
6. Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, *23* (1), 205–211. https://doi.org/10.1142/9781848165632_0019
7. Gao, B. & Gupta, R. S. (2012). Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. *Microbiology and Molecular Biology Reviews*, *76*(1), 66–112. https://doi.org/10.1128/mmbr.05011-11
8. Global Tuberculosis Report (2021). Geneva: World Health Organization. Licence: CC BY-NC-SA 3.0 IGO. Retrieved from https://www.who.int/publications/i/item/9789240037021
9. Gorman, J. & Shapiro, L. (2004). Structure of serine acetyltransferase from Haemophilus influenzae Rd. *Acta Crystallographica Section D: Biological Crystallography*, *60*(9), 1600–1605. https://doi.org/10.1107/S0907444904015240
10. Gupta, S. & Gupta, V. (2020). Homology modeling, structural insights and in-silico screening for selective inhibitors of mycobacterial CysE. *Journal of Biomolecular Structure and Dynamics*, *39*(5), 1547–1560. https://doi.org/10.1080/07391102.2020.1734089
11. Hampshire, T., Soneji, S., Bacon, J., James, B. W., Hinds, J., Laing, K., …& Butcher, P. D. (2004). Stationary phase gene expression of Mycobacterium tuberculosis following a progressive nutrient depletion: A model for persistent organisms? *Tuberculosis*, *84*(3–4), 228–238. https://doi.org/10.1016/j.tube.2003.12.010
12. Han, M. V. & Zmasek, C. M. (2009). PhyloXML: XML for evolutionary biology and comparative genomics. *BMC*

Bioinformatics, *10*(1), 1–6. https://doi.org/10.1186/1471-2105-10-356
13. Handwerger, S. & Tomasz, A. (1985). Antibiotic Tolerance Among Clinical Isolates of Bacteria. *Reviews of Infectious Diseases*, *7*(3), 368–386. https://doi.org/10.1093/CLINIDS/7.3.368
14. Huang, B., Vetting, M. W. & Roderick, S. L. (2005). The active site of O-acetylserine sulfhydrylase is the anchor point for bienzyme complex formation with serine acetyltransferase. *Journal of Bacteriology*, *187*(9), 3201–3205. https://doi.org/10.1128/JB.187.9.3201-3205.2005
15. Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., …& Birney, E. (2007). Ensembl 2007. *Nucleic Acids Research*, *35*(SUPPL. 1). https://doi.org/10.1093/nar/gkl996
16. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., …& Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085
17. Jean Kumar, V. U., Poyraz, Ö., Saxena, S., Schnell, R., Yogeeswari, P., Schneider, G. & Sriram, D. (2013). Discovery of novel inhibitors targeting the Mycobacterium tuberculosis O-acetylserine sulfhydrylase (CysK1) using virtual high-throughput screening. *Bioorganic and Medicinal Chemistry Letters*, *23*(5), 1182–1186. https://doi.org/10.1016/j.bmcl.2013.01.031
18. Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. & Bork, P. (2008). eggNOG: Automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, *36*(SUPPL. 1). https://doi.org/10.1093/nar/gkm796
19. Johnson, C. M., Huang, B., Roderick, S. L. & Cook, P. F. (2004). Kinetic mechanism of the serine acetyltransferase from Haemophilus influenzae. *Archives of Biochemistry and Biophysics*, *429*(2), 115–122. https://doi.org/10.1016/j.abb.2004.06.006
20. Joshi, P., Gupta, A. & Gupta, V. (2019). Insights into multifaceted activities of CysK for therapeutic interventions. *3 Biotech*, *9*(2), 0. https://doi.org/10.1007/s13205-019-1572-4
21. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436
22. Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010
23. Keikha, M. (2018). Importance of the identification of Segniliparus species from pulmonary infection. *New Microbes and New Infections*. Elsevier Ltd. https://doi.org/10.1016/j.nmni.2018.05.002
24. Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simão, F. A., Pozdnyakov, I. A., …& Zdobnov, E. M. (2015). OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, *43*(D1), D250–D256. https://doi.org/10.1093/nar/gku1220
25. Kumar, S., Kumar, N., Alam, N. & Gourinath, S. (2014). Crystal structure of serine acetyl transferase from Brucella abortus and its complex with coenzyme A. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1844*(10), 1741–1748. https://doi.org/10.1016/j.bbapap.2014.07.009
26. Kumar, S., Mazumder, M., Dharavath, S. & Gourinath, S. (2013). Single Residue Mutation in Active Site of Serine Acetyltransferase Isoform 3 from Entamoeba histolytica Assists in Partial Regaining of Feedback Inhibition by Cysteine. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0055932
27. Kumar, S., Raj, I., Nagpal, I., Subbarao, N. & Gourinath, S. (2011). Structural and biochemical studies of serine acetyltransferase reveal why the parasite Entamoeba histolytica cannot form a cysteine synthase complex. *Journal of Biological Chemistry*, *286*(14), 12533–12541. https://doi.org/10.1074/jbc.M110.197376
28. Li, L., Stoeckert, C. J. & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503
29. Nakamori, S., Kobayashi, S. I., Kobayashi, C. & Takagi, H. (1998). Overproduction of L-cysteine and L-cystine by Escherichia coli strains with a genetically altered serine acetyltransferase. *Applied and Environmental Microbiology*, *64*(5), 1607–1611. https://doi.org/10.1128/aem.64.5.1607-1611.1998
30. Nguyen, L., Kozlov, G. & Gehring, K. (2008). Structure of Escherichia coli tetrahydrodipicolinate N-succinyl transferase reveals the role of a conserved C-terminal helix in cooperative substrate binding. *FEBS Letters*, *582*(5), 623–626. https://doi.org/10.1016/j.febslet.200 8.01.032
31. Olsen, L. R., Huang, B., Vetting, M. W. & Roderick, S. L. (2004). Structure of serine acetyltransferase in complexes with CoA and its cysteine feedback inhibitor. *Biochemistry*, *43*(20), 6013–6019. https://doi.org/10.1021/bi0358521
32. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084
33. Poyraz, Ö., Jeankumar, V. U., Saxena, S., Schnell, R., Haraldsson, M., Yogeeswari, P., …& Schneider, G. (2013). Structure-guided design of novel thiazolidine inhibitors of O -Acetyl serine sulfhydrylase from mycobacterium tuberculosis. *Journal of Medicinal Chemistry*, *56*(16), 6457–6466. https://doi.org/10.1021/jm400710k
34. Price, M. N., Dehal, P. S. & Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490
35. Pye, V. E., Tingey, A. P., Robson, R. L. & Moody, P. C. E. (2004). The structure and mechanism of serine acetyltransferase from Escherichia coli. *The Journal of Biological Chemistry*, *279*(39), 40729–40736. https://doi.org/10.1074/jbc.M403751200
36. Rengarajan, J., Bloom, B. R. & Rubin, E. J. (2005). Genome-wide requirements for Mycobacterium tuberculosis adaptation and survival in macrophages. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(23), 8327–8332. https://doi.org/10.1073/

pnas.0503272102
37. Roth, A. C. J., Gonnet, G. H. & Dessimoz, C. (2008). Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-518
38. Sassetti, C. M. & Rubin, E. J. (2003). Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22), 12989–12994. https://doi.org/10.1073/PNAS.2134250100/SUPPL_FILE/4250TABLE3.XLS
39. Schnell, R., Oehlmann, W., Singh, M. & Schneider, G. (2007). Structural insights into catalysis and inhibition of O -acetylserine sulfhydrylase from Mycobacterium tuberculosis. Crystal structures of the enzyme alpha-aminoacrylate intermediate and an enzyme-inhibitor complex. *The Journal of Biological Chemistry*, *282*(32), 23473–23481. https://doi.org/10.1074/JBC.M703518200
40. Sievers, F. & Higgins, D. G. (2014). Clustal Omega. *Current Protocols in Bioinformatics*, *2014*, 3.13.1-3.13.16. https://doi.org/10.1002/0471250953.bi0313s48
41. Smith, I. K. & Thompson, J. F. (1969). The synthesis of O-acetylserine by extracts prepared from higher plants. *Biochemical and Biophysical Research Communications*, *35*(6), 939–945. https://doi.org/10.1016/0006-291X(69)90715-3
42. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., …& Koonin, E. V. (2001). The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, *29*(1), 22–28. https://doi.org/10.1093/nar/29.1.22
43. Trachana., K., Forslund., K., Larsson, T., Powell, S., Doerks, T., Von Mering, C. & Bork, P. (2014). A Phylogeny-Based benchmarking test for orthology inference reveals the limitations of Function-Based validation. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0111122
44. Ullas, V., Poyraz, Ö., Saxena, S., Schnell, R., Yogeeswari, P., Schneider, G. & Sriram, D. (2013). Discovery of novel inhibitors targeting the Mycobacterium tuberculosis O -acetylserine sulfhydrylase (CysK1) using virtual high-throughput screening. *Bioorganic & Medicinal Chemistry Letters*, *23*(5), 1182–1186. https://doi.org/10.1016/j.bmcl.2013.01.031
45. Vasudevan, S., YI, W., JJ, Y., DA, N., RL, T., ND, F., …& AV, S. (2003). The COG database: an updated version includes eukaryotes.. *BMC Bioinformatics*. Retrieved from http://www.biomedcentral.com/1471-2105/4/41
46. Verma, D., Gupta, S., Saxena, R., Kaur, P., Rachana, R. R., Srivastava, S. & Gupta, V. (2020). Allosteric inhibition and kinetic characterization of Klebsiella pneumoniae CysE: An emerging drug target. *International Journal of Biological Macromolecules*, *151*, 1240–1249. https://doi.org/10.1016/j.ijbiomac.2019.10.170
47. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033
48. Wiederstein, M. & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(SUPPL.2), W407–W410. https://doi.org/10.1093/nar/gkm290
49. Zhang, Y. (2009). I-TASSER: fully automated protein structure prediction in CASP8. *Proteins*, *77 Suppl 9*(Suppl 9), 100–113. https://doi.org/10.1002/PROT.22588
50. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nature Methods*. https://doi.org/10.1038/nmeth.3213
51. Yelamanchi, S. D. & Surolia, A. (2021). Targeting amino acid metabolism of Mycobacterium tuberculosis for developing inhibitors to curtail its survival. *IUBMB Life*, *73*(4), 643–658. https://doi.org/10.1002/IUB.2455
52. Yi, H., Dey, S., Kumaran, S., Lee, S. G., Krishnan, H. B. & Jez, J. M. (2013). Structure of soybean serine acetyltransferase and formation of the cysteine regulatory complex as a molecular chaperone. *Journal of Biological Chemistry*, *288*(51), 36463–36472. https://doi.org/10.1074/jbc.M113.527143
53. Zhang, Z., Esther, M. M., Bunker, R. D., Baker, E. N. & Squire, C. J. (2009). research papers Structure and function of GlmU from Mycobacterium tuberculosis research papers, 275–283. https://doi.org/10.1107/S090744490901036