

# The Rv3799–Rv3807 gene cluster in *Mycobacterium tuberculosis* genome corresponds to the ‘Ancient Conserved Region’ in CMN mycolyltransferases

Hemalatha G. Ramulu, Adindla Swathi and Lalitha Guruprasad

School of Chemistry, University of Hyderabad, Hyderabad 500 046, India.

**Abstract:** We have identified based on gene cluster analysis that the genes between Rv3799–Rv3807 in *M. tuberculosis* have orthologs in Corynebacteria, Mycobacteria and Nocardia (CMN) genomes. Therefore, this gene cluster possibly corresponds to the ‘Ancient Conserved Region’ of CMN mycolyltransferases. The evolutionary trace analysis suggests that twelve amino acid residues; Leu39, Trp51, Pro71, Trp82, Trp97, Phe100, Gly124, Ser126, Asp192, Glu230, Gly260 and Trp264 are ‘absolutely conserved’. These amino acid residues constitute the active site and conserved hydrophobic tunnel in CMN mycolyltransferases.

**Keywords:** Corynebacteria; Mycobacteria; Nocardia; Mycolyltransferases; Gene cluster, Evolutionary trace analysis.

## Introduction

The organisms belonging to the Corynebacterium, Mycobacterium and Nocardia (or ‘CMN’) genera have been grouped together based on factors that distinguish them individually. These include; complex cell wall components, presence and type of mycolic acids, adjuvant activity, presence of cord factor, sulfolipids, iron-chelating compounds, polyphosphate, and serological cross-reactivity. The cell wall of the CMN group organisms consist interconnected peptidoglycan and polysaccharide-mycolate complex and are characterized by the presence of mycolic acid on their surface (Cocito and Delville 1985). Mycolic acids are long chain fatty acids that form a part of the unique cell envelope responsible for the pathogenesis and the survival of the organism inside the host. These mycolic acids are identified by different names depending upon the genus and comprise different carbon chain lengths; corynomycolic acids from the genus Corynebacterium has 22–36 long carbon chain, mycolic acids from the genus Mycobacterium has nearly 60–90 long carbon chain and nocardomycolic acids from the genus Nocardia has 40–60 long carbon chain (Collins et al 1982; Minnikin 1982; Daffé M and Draper 1992).

In *M. tuberculosis*, the antigen 85 complex enzymes constitute three secreted proteins (Wiker and Harboe 1992); Ag85A (gene identifier: Rv3804), Ag85B (Rv1886) and Ag85C (Rv0129) that comprise a signal peptide at the N-terminus followed by a carboxylesterase domain. These are known to catalyse the transfer of mycolic acids to the  $\alpha, \alpha'$ -trehalosemonomycolate (TMM) and arabionogalactan. It has been demonstrated that Ag85 complex enzymes catalyse the transfer of mycolyl residue from one molecule of TMM to another TMM leading to the formation of  $\alpha, \alpha'$ -trehalosedimycolate (TDM) and hence these enzymes are termed mycolyltransferases (Belisle et al 1997). Also, in Corynebacterium and Nocardia, orthologous proteins synthesize trehalosedicorynomycolate (TDCM) and trehalosedino-cardiomycolate (TDNM), respectively. Mycolyltransferases are also termed fibronectin binding proteins, since they are involved in binding fibronectin that aids entry of the organism into host cells (Abou-Zeid et al 1988; Ratliff et al 1988). The compound TDM commonly known as “cord factor” extracted from *M. tuberculosis* has been shown to be toxic in mouse model (Kato 1968). Hence a study related to the analysis of the structure, function and evolution of proteins responsible for the synthesis of TDM in these species is important.

The three dimensional structures of Ag85A (PDB code: 1SFR; Ronning et al 2004), Ag85B (1F0P; Anderson et al 2001) and Ag85C (1DQZ, 1VA5; Ronning et al 2000) are known for both native and substrate bound forms and comprises a  $\alpha/\beta$  hydrolase fold. The catalytic triad constituting S126, G230 and H262 (numbering is according to PDB code: 1F0P) is responsible for the mycolyltransferase activity. The structural comparison

**Correspondence:** Lalitha Guruprasad, School of Chemistry, University of Hyderabad, Hyderabad 500 046, India. Tel: + 91 40 2313 4820; Fax: + 91 40 2301 2460; Email: lgpsc@uohyd.ernet.in

of these mycolyltransferases revealed that the active sites are virtually identical indicating that they share the same substrate. However, in contrast to the high level conservation within the substrate-binding and active site, it was observed that surface residues disparate from the active site are quite variable indicating that all three Ag85 enzymes are needed to evade the host immune system (Ronning et al 2004). The multiple sequence alignment (see Appendix-A in supplementary data) suggests that the three sequences corresponding to Ag85A, Ag85B and Ag85C share more than 69% sequence identity. In our previous work (Adindla et al 2004a), we identified mycolyltransferases in the *C. glutamicum* and *C. efficiens* genome and analyzed the three-dimensional computer models that were constructed based on comparative modeling methods. The mycolyltransferases are restricted to the CMN genera and the complete genome sequences of *M. tuberculosis* (Cole et al 1998), *C. glutamicum* (Kalinowski et al 2003), *C. efficiens* (Kawarabayasi et al 2002), *C. diphtheria* (Cerdeno-Tarraga et al 2003) and *Nocardia farcinica* (Ishikawa et al 2004) are now available. Therefore, we intended to identify and analyse all the mycolyltransferases from various species in order to get an insight into their substrate specificity. Since there are several isoforms in each genome, we intended to understand the origin of their evolution and therefore carried out the evolutionary trace analysis. Further, synteny or colinearity of gene order is observed when a group of genes are present in the same order in two or more genomes as a cluster. Two species that have recently diverged from a common ancestor might be expected to share a similar set of genes present in the same order. During evolution the sequence of each pair of genes is accompanied by changes, such as, gene duplication and gene loss. Genetic analyses reveal that genes with related function are frequently clustered at one chromosomal location in evolution. Despite *M. tuberculosis* being an ancient species compared to *Corynebacterium* and *Nocardia*, evolution has maintained the conservation of proteins involved in the synthesis of the cell envelope. Also, it is interesting to note that there are varying numbers of mycolyltransferases in these different species. We therefore intended to identify the evolutionary origin of CMN mycolyltransferases based on the computational analysis of 'gene neighborhood' and to identify the role of conserved amino acid residues using the evolutionary trace analysis.

## Materials and methods

The amino acid sequences corresponding to the mycolyltransferases; Ag85A, Ag85B and Ag85C were obtained from the website at [www.srs.ebi.ac.uk/](http://www.srs.ebi.ac.uk/). The homologous proteins from *C. glutamicum*, *C. efficiens*, *C. diphtheria* and *N. farcinica* completed genome database were identified using the BLASTP and PSI-BLAST (Altschul et al 1990; 1997) programs available at the website [www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/) and by using the sequence corresponding to Ag85B as query. The blosum62 matrices were used and the results were sorted based on p-value. The analysis of gene clusters was carried out by performing BLAST searches using mycolyltransferases and their neighbouring proteins as query on all the finished and unfinished genomes. The evolutionary trace (ET) analysis was carried out using TraceSuite II server (Innis et al., 2000) available at the website <http://www.cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html> by submitting the sequences corresponding to the carboxylesterase domain of CMN mycolyltransferases and the crystal structure of Ag85B (PDB code: 1F0P). A trace is generated by comparing the consensus sequences for groups of proteins that originate from a common node in a phylogenetic tree and is characterized by a common evolutionary time cut-off (ETC).

## Results and discussion

### Identification of CMN mycolyltransferases

We identified 32 mycolyltransferases in the genera of CMN group; 4 proteins in *M. tuberculosis*, 4 in *C. diphtheria*, 6 in *C. glutamicum*, 5 in *C. efficiens*, and 13 in *N. farcinica*. The four mycolyltransferases corresponding to each of the mycobacterial species; *M. tuberculosis*, *M. leprae* and *M. bovis* are highly similar, therefore only the mycolyltransferases from *M. tuberculosis* is referred to in the discussion. There are between 350–480 amino acid residues in these proteins. However, proteins corresponding to gene identifiers; Ncgl2777, Ce2709 and Dip2193 (in *Corynebacterium*) and Nfa1840 (in *N. farcinica*) are associated with an additional ~300 amino acid residue domain towards the C-terminus that is not a part of the carboxylesterase domain. The carboxylesterase domain that is responsible for the mycolyltransferase activity, corresponds to approximately 280 amino acid residues. The multiple sequence alignment is attached as



**Figure 1.** Schematic representation of genes corresponding to the ‘Ancient Conserved Regions’ in five completed genomes based on gene neighbourhood analysis. ‘/’ indicates insertion of gene. Nfa1810-30 indicates three genes Nfa1810, Nfa1820 and Nfa1830.

supplementary data along with this manuscript in Appendix-A. The *N. farcinica* proteins; Nfa1820 and Nfa1810, comprise a ‘long insertion’ sequence of 22 and 27 amino acid residues that is rich in glycine and serine respectively relative to the other CMN proteins. This region is located between the ‘absolutely conserved’ W82 and W97 residues. The glycine/serine rich sequences are often known to be associated with cell-surface proteins. Another insertion region in some corynomycyltransferases and nocardiomycyl transferases relative to mycolyltransferases is located between the ‘absolutely conserved’ D192 and E230 amino acid residues. We predict this loop to be close to the substrate binding site based on a comparison with the crystal structure of Ag85B.

### Gene cluster analysis

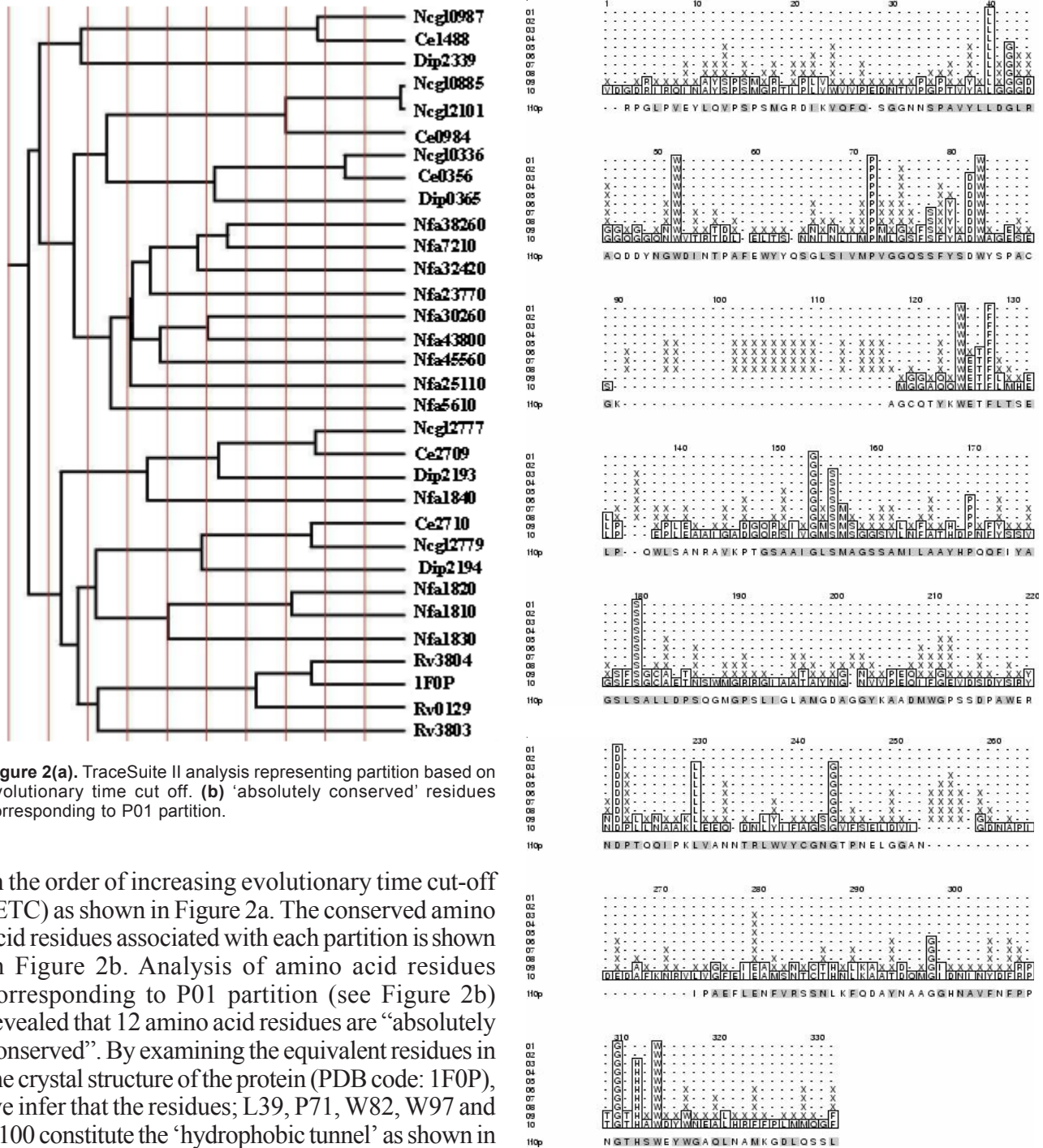
The analysis of all mycolyltransferases and their neighbouring proteins revealed that genes between Rv3799 – Rv3807 in the *M. tuberculosis* genome has corresponding orthologs in Corynebacterium and Nocardia genera as shown in Figure 1. The ten protein orthologs shown in Figure 1 share high sequence similarity in the five different species analyzed. In addition to mycolyltransferase (Rv3804) and its precursor protein (Rv3803) this cluster also comprises propionyl CoA carboxylase (Rv3799), polyketide synthase (Rv3800), acyl CoA synthase (Rv3801), membrane proteins (Rv3806, Rv3807), and hypothetical proteins (Rv3802, Rv3805). We

observed that the Nocardia proteins are arranged in the reverse order relative to the other species. We report that this set of genes represents the only mycolyltransferase comprising gene cluster during divergence of a common ancestral organism into individual genera, such as, Corynebacterium, Mycobacterium and Nocardia (CMN group). Therefore, we propose that this gene cluster corresponds to the “Ancient Conserved Regions - ACR’s” among the mycolyltransferases across the CMN genera. It was reported that Rv3800 (*pks13*) is involved in the final condensation step in mycolic acid synthesis (Damien et al 2004). It was also reported that the genes; Rv3799, Rv3800 and Rv3801 (*accD4-pks13-fadD32*) play an essential role in the biosynthesis of mycolic acids (Gande et al 2004). This indicates that the proteins in this cluster are important for the mycolic acid synthesis and its transfer to trehalose. Since, functionally related genes are often clustered, we suggest that the other “uncharacterized” proteins belonging to the ACR gene cluster may also have a role in associated functions. Further, we observed that the mycolyltransferase gene neighbours; Rv0129 and Rv1886 are conserved among *M. tuberculosis* and *M. bovis* suggesting that gene duplication events have occurred before speciation.

### Evolutionary Trace Analysis

The TraceSuite II server generates a phylogenetic tree split into 10 evenly distributed partitions (P01–P10)



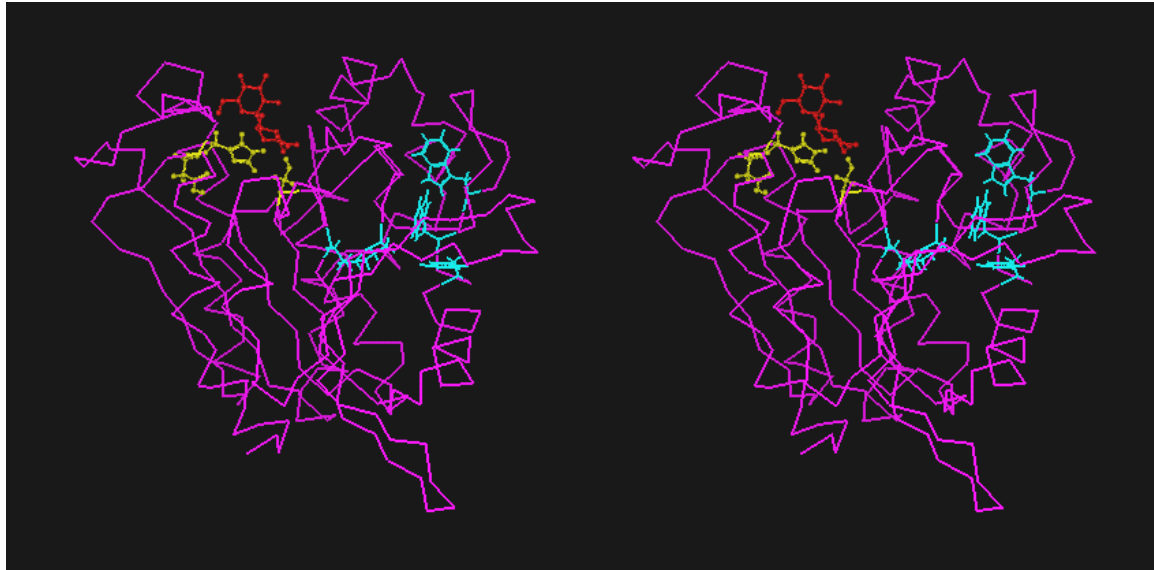


**Figure 2(a).** TraceSuite II analysis representing partition based on evolutionary time cut off. **(b)** ‘absolutely conserved’ residues corresponding to P01 partition.

in the order of increasing evolutionary time cut-off (ETC) as shown in Figure 2a. The conserved amino acid residues associated with each partition is shown in Figure 2b. Analysis of amino acid residues corresponding to P01 partition (see Figure 2b) revealed that 12 amino acid residues are “absolutely conserved”. By examining the equivalent residues in the crystal structure of the protein (PDB code: 1F0P), we infer that the residues; L39, P71, W82, W97 and F100 constitute the ‘hydrophobic tunnel’ as shown in Figure 3a. This figure also indicates the amino acid residues involved in the catalytic triad. The residues in the ‘hydrophobic tunnel’ are needed in order to accommodate the alkyl chain of mycolic acid indicating a functional conservation in these proteins. The invariant S126 and G260 are close to the catalytic active site comprising the amino acid residue E230. The indole side chains of the W51 and W264 are perpendicular to each other and are in proximity to G124 associated with the  $\beta$  5 strand. The amino acid residue D192 is away from the active site

**Figure 2(b).**

indicating that the conservation extends beyond the catalytic site in CMN mycolyltransferases. According to Figure 2a, the 14 proteins indicated in the lower half, from Corynebacterium, Mycobacterium, and Nocardia represent the ‘Ancient Conserved Region’. The 18 proteins in the upper half, comprise only the Nocardia and Corynebacterium. From the multiple sequence

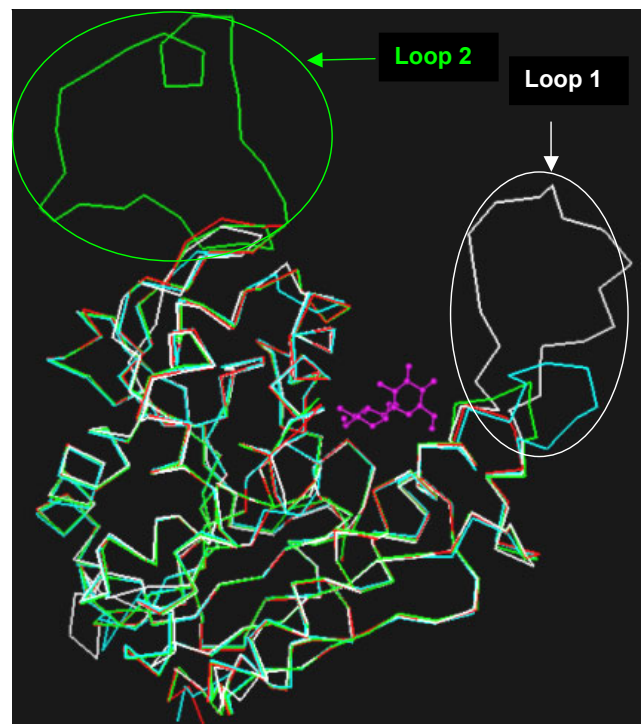


**Figure 3(a).** Stereo-view showing three-dimensional model corresponding to the protein with gene identity Nfa1840. The amino acid residues comprising the catalytic triad (pink), hydrophobic tunnel (blue) and trehalose (red) are also indicated.

alignment, we observed that the proteins in the upper half of Figure 2a are associated with an insertion loop of variable length between 4 to 20 amino acid residues and this loop is close to the active site. The positions of these insertion loops are shown in Figure 3b. Further, the amino acid residues comprising the specificity pockets defined by interactions with trehalose substrate in the protein with PDB code: 1F0P are mutated in these proteins. Primarily, the mutations associated with the substrate binding sites in some *Corynebacterium* (Adindla et al., 2004a) and *Nocardia* proteins accompanied by the presence of ‘insertion loops’ close to the active site suggest that these may interfere with trehalose binding. These *Corynebacterium* and *Nocardia* proteins are possibly a result of divergent evolution accompanied by gene duplication and mutation events in order to accommodate different substrates in the binding site. This suggests that the ancient proteins form a distinct cluster and are different from proteins that evolved later.

We previously reported that the corynemyltransferase Ncgl2777 gene in *C. glutamicum* (protein with ~300 amino acid residue C-terminal extension) is associated with a 55 amino acid residue ‘LGFP’ tandem repeat that is likely to be associated with maintaining cell-wall integrity (Adindla et al 2004b). Our hypothesis was based on the work of Brand et al., 2003 who have demonstrated that the deletion of Ncgl2777 gene in *C. glutamicum* resulted in a 10-fold increase in cell volume of the organism thereby suggesting its involvement in cell

shape formation. In this work, we observed that the ‘LGFP’ tandem repeats are also present in the C-terminal region of *Nocardia* (NfaA1840) and *C. diphtheria* (Dip2193) proteins and accordingly may be involved in maintaining cell wall integrity.



**Figure 3(b).** Structural overlay corresponding to the proteins with PDB code: 1F0P and gene identities: Nfa1810, Ncgl0336 and Ncgl0987 indicating the position of the two loops; loop 1 (D192 – E230 loop) and loop 2 (W82-W97 loop).

## Conclusions

The comparative analysis of mycolyltransferase proteins from different genomes suggested that the gene cluster corresponding to the ten gene families located between Rv3799 and Rv3807 in *Mycobacterium tuberculosis* genome represents the 'Ancient Conserved Region' in CMN genera. According to the evolutionary trace analysis twelve amino acid residues are 'absolutely conserved' in all CMN proteins analyzed. These CMN proteins fall into two distinct clusters in the phylogeny that correlates with the presence or absence of insertion loop close to the active site. Some *Corynebacterium* and *Nocardia* proteins with extra C-terminal 300 amino acid residues are associated with the LGFP tandem repeats.

## Acknowledgments

HGR thanks UGC, New Delhi for a JRF fellowship. SA thanks CSIR New Delhi for a SRF fellowship. LGP thanks DBT, New Delhi for research funding.

## References

- Abou-Zeid C, Ratliff TL, Wiker HG, et al. 1988. Characterisation of fibronectin-binding antigens released by *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG. *Infect Immun*, 56:3046–3051.
- Adindla S, Guruprasad K, Guruprasad L. 2004a. Three-dimensional models and structure analysis of corynemycolyltransferases in *Corynebacterium glutamicum* and *Corynebacterium efficiens*. *Int J Biol Macromol*, 34:181–189.
- Adindla S, Inampudi KK, Guruprasad K, et al. 2004b. Identification and analysis of novel repeats in the cell surface proteins of archaeal and bacterial genomes using computational tools. *Comp Funct Genom*, 5:2–16.
- Altschul SF, Gish W, Miller W, et al. 1990. Basic local alignment search tool. *J Mol Biol*, 215:403–410.
- Altschul SF, Madden TL, Schäffer AA, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402.
- Anderson DH, Harth G, Horwitz MA, et al. 2001. An interfacial mechanism and a class of inhibitors inferred from two crystal structures of the *Mycobacterium tuberculosis* 30 kDa Major secretory protein (antigen 85B), a mycolyl transferase. *J. Mol. Biol*, 307:671–681.
- Belisle JT, Vissa VD, Sievert T, et al. 1997. Role of the major antigen of *Mycobacterium tuberculosis* in the cell wall biogenesis. *Science*, 276:1420–1422.
- Brand S, Niehaus K, Puhler A, et al. 2003. Identification and functional analysis of six mycolyltransferase genes of *Corynebacterium glutamicum* ATCC 13032: the genes cop1, cmt1, and cmt2 can replace each other in the synthesis of trehalose dicorynomycolate, a component of the mycolic acid layer of the cell envelope. *Arch. Microbiol*, 180:33–44.
- Cocito A, Delville J. 1985. Biological, chemical, immunological and staining properties of bacteria isolated from tissues of leprosy patients. *Eur J. Epidemiol*, 1:202–231.
- Cole ST, Brosch R, Parkhill J, et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685):537–544.
- Cerdeno-Tarraga AM, Efstratiou A, Dover LG, et al. 2003. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Res*, 31(22):6516–6523.
- Collins MD, Goodfellow M, Minnikin DE. 1982. A survey of the structures of mycolic acids in *Corynebacterium* and related taxa. *J Gen. Microbiol*, 128:129–149.
- Daffé M, Draper P. 1998. The envelope layers of mycobacteria with reference to their pathogenicity. *Adv Microb Phys*, 39:131–203.
- Damien P, Célia de Sousa-D'Auria, Christine H, et al. 2004. A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proc Natl Acad Sci USA*, 101:314–319.
- Gande R, Gibson KJC, Brown AK, et al. 2004. Acyl-CoA Carboxylases (*accD2* and *accD3*), Together with a Unique Polyketide Synthase (*Cg-pks*), Are Key to Mycolic Acid Biosynthesis in *Corynebacteriaceae* Such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*. *J. Biol. Chem*, 279:44847–44857.
- Innis CA, Shi J, Blundell TL. 2000. Evolutionary trace analysis of TGF- $\beta$  and related growth factors: implications for site-directed mutagenesis. *Prot. Engin*. 13:839–847.
- Ishikawa J, Yamashita A, Mikami Y, et al. 2004. The complete genomic sequence of *Nocardia farcinica* IFM 10152. *Proc Natl Acad Sci USA*, 101:14925–14930.
- Kalinowski J, Bathe B, Bartels D, et al. 2003. The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J Biotechnol*, 104(1–3):5–25.
- Kawarabayasi Y, Yamazaki J, Hino Y, et al. 2002. The entire genomic sequence of *Corynebacterium efficiens* YS-314. Unpublished.
- Minnikin DE. 1982. The Biology of the Mycobacteria. In: Ratledge C, Stanford JL, eds. London: Academic Press, p 95–184.
- Ratliff TL, McGarr JA, Abou-Zeid C, et al. 1988. Attachment of mycobacteria to fibronectin-coated surfaces. *J. Gen. Microbiol*, 134:1307–1313.
- Ronning DR, Klabunde T, Besra GS, et al. 2000. Crystal structure of the secreted form of antigen 85C reveals potential targets mycobacterial drugs and vaccines. *Nat. Struc. Biol*, 7(2):141–146.
- Ronning DR, Vissa V, Gurdyal B, et al. 2004. *Mycobacterium tuberculosis* antigen 85A and 85C structures confirm binding orientation and conserved substrate specificity. *J. Biol. Chem*, 279:36771–36777.
- Wiker HG, Harboe M. 1992. The antigen 85 complex: a major secretion product of *Mycobacterium tuberculosis*. *Microbiol. Rev*, 56:648–661.

## Appendix-A

### Multiple sequence alignment corresponding to carboxylesterase domain in CMN mycolyltransferases.

cmytD	LNGLRLEKWSVASPSMQRNVDVQIMKSAEADSPAPMLYMLDGI	GGNKNSSGWINGGEGPKVFADENVTVV
Cel488	MDGLRLERWTVASPSMQRNVDVQIMRSDVAGAPAPMLYMLDGI	GGNRNSSGWINHGQPKVFGDENVTVV
cmytE	VDGDRIRQINAYSPSMGRITPLVWVVPEDNTVPGPTVYALGGG	GGGQGNWVTRTDLDELTSENNINLI
cmytF	VDGDRIRQINAYSPSMGRITPLVWVVPEDNTVPGPTVYALGGG	GGGQGNWVTRTDLLEELTSDNNINLI
Ce0984	VDGERIRQINAYSPSMERWIPLVWVVPEDTSEPRPTLYALGGG	GGGQGSANWITKTDMPELMSNNVHVI
cmytC	AADERVKEMWAYSPSMDRNVPLVVIITADESAGPRPVIYLLNGG	DGGEAANWVMTDVLDFYLEKNVNVV
Ce0356	ASGERVKEMWAYSPSMDRDVPLVVIITADESAGPRPVIYLLNGG	DGGEANANWIMQTDVDFYLEKNVNVV
Dip0365	ATGDRVVEMWAHSPSMNRNPLVVLKAAANPG--RPTIYLLNGG	DGGEGSANWVMTKALDFYRDKDNVNV
Nfa38260	VVDARTVRLRVYSAAMGRVIDIDVQRPADTGAPRPTLYLLAGAGG	GEDSASWAKQTSVLEFLADKKNVNV
Nfa7210	IKDDRNLRLYVYSAAMDENVIIDVQRPADASVPRPTLYLLNGAGG	EDDASWVAKSDALKFLSDKNVNV
Nfa32420	AKEGRTWHLTVYSAAMDTEIAVDVQRPADDSVPAPNLYMLNGLD	GGEGTASWAAATHALDNLADKPNVNI
Nfa23770	GTPARLVDLAVYSPAMQRSIAVKVLRPADTTRPAPTLYLLNGAGG	GEDAANWFGQTDAVEFFADKHNVVV
Nfa30260	PRSDREVEVIVHSAAMAAEIPIRLLRAADPDRPAPTLYLLNGIT	GGGDDGNWFDRTGVAAFFAGEQVNVA
Nfa43800	PENDRLLDLEIHSPAMDSTTRVLLLRAPDPRPAPTLYLLNGASGH	VDG-SWHDRTDYQRFADKQVNVV
Nfa45560	PLGGRQLEVVVHSAAMNRPIITLWMS---HPGPGAPALYLLNAVD	GGEDGGPWNRTDVAFFADKKNVNI
Nfa25110	PLAPRVQVQVYSPSMDAVVSTVIR---ADGPAPTLYLLAGAGGGT	DGISWVHHTDVRQFFADKKNVNVV
Nfa5610	ELSPTRSAVFDVSPAMGRVIQVQLHP-AGGAARPSYLLDGLDPG	VQGSTWTNATDABEAFFRGKNVNVV
Dip2339	DERFVDVRLFIESPAMRRIVQVQVQHPKDRTPAPMLYLLDGVTA	P-SQSGWLKRGVQVGAMANEHVTVI
cmytA	--ILTIQSAAMPERPIKVQLLLPRDWYSSPNREFPEI WALDGLRA	IEEQSGWTIETNIEQYYADKNAIVV
Ce2709	--VLSIQSAAMPERPIKVQLLLPRDWYSSPDRDFPEI WALDGLRA	IEKQSGWTIETNIEQFFADKNAIVV
Dip2193	--AVYVNTPSMGQVQVQIILLARDWFQDPNRSFSPVWALDGLRAT	DVENGWTIGTNI EQFYSKKNVNI
Nfa1840	ALWVNSPSMG-APVQVQLLLARDWNA--KPEARFPLLIMLDGLRAT	DDESGWTKDAGABEFFADKNVTVV
Ce2710	--GVGYWVQRCDVYSPAMGRNITAVQIQPAQRGGNAGLYLLDGM	RATWNSAWLVDTNAAALYAPHNITLV
cmytB	--AVGFVWQRCDVWSPAMGRNIPVQIQPAGRGGNAGLYLLDGM	RATEYSNAWLVDTNAAALYAPHNITLV
Dip2194	--GVAHWVQRCDVFS PAMGRNITVQIQPAQRGGNAALYLLDGA	RANETANAWTDAHVQDLFVDHNITLV
Nfa1820	--AFDPAAFDVWVDSMGPIKSRI LRA-ADGNTNRVYVLDGMRAP	ETLNGWEIETDVPALLASWNINVV
Nfa1810	--AFNPDGDFWVDSMGPIKSRI FRA-ADGNTNRVYVLDGMRAR	NDLSGWEIDTEVARELTKWNINVV
Nfa1830	--APAGGYEELMVSPVMGPIKVQVQWA-SRG-GDAALYLLDGLR	ARRDRNWSFETNAMEQFKNDNITLV
Rv3804c	--RPGLPVEYLQVPSPSMGRDIKVQFQ-SGGANS PALYLLDGLRA	QDDFSGWINDINTPAFEWYDQSGLSVV
1f0pA	--RPGLPVEYLQVPSPSMGRDIKVQFQ-SGGNSP AVYLLDGLRA	QDDYNGWDINTPAFEWYDQSGLSIV
Rv0129c	--RPGLPVEYLQVPSASMGRDIKVQFQ-GGG--PHAVYLLDGLRA	QDDYNGWDINTPAFEWYDQSGLSVI
Rv3803c	--AKAAPYENLMVPSPSMGRDIPVAFI-AGG--PHAVYLLDAFN	AGPDVSNWVTAGNAMNTLAGKGISVV

| 10                      | 20                      | 30                      | 40                      | 50                      | 60

cmytD	MPLGAASSMYSDWLEEDP -----	ALGRIKWETFIVEELA-PLLEA
Cel488	MPLGAAASMYSDWVEEDP -----	ALGRIMWETFIVEELA-PLLEA
cmytE	MPMLGSFSFYADWAGESES -----	MGGAQQWETFMLHELP---EPL
cmytF	MPMLGSFSFYADWAGESES -----	MGGAQQWETFMLHELP---EPL
Ce0984	MPMLGSHSFYADWVEENDS -----	LGGKQQWETFFLTHELP---EPL
cmytC	IPMEGKFSYYTDWVEENAS -----	LGGKQMWETFVLKELP---GPL
Ce0356	IPMEGKFSYYTDWVQENAA -----	LGGKQMWETFVLKELP---GPL
Dip0365	IPMAGKFSYYTDWVSEAPS -----	LGGKQNWETFFLTKELP---GPI
Nfa38260	QPIGGAWTYTDRAPDP -----	ALGVNKWTFFLTEELP---PVI
Nfa7210	QPIGGKWSYTDWIKDDP -----	TLGRNKWTFFFTEELP---PLV
Nfa32420	QPIGGRGSYYTDWLRDRP -----	ELGMNKWRTFFTEELP---PLL
Nfa23770	IPMEGAFSYYTDWERADEGLAETL -----	GNNGRNMWTFFLTEELP---PVI
Nfa30260	MPIGGAGSYFADWRARDP -----	VLGLQRWASFLTREL P---PLL
Nfa43800	IPLGGAGSYFTDWR AEDP -----	VLGRQRWATFLTEELP---PLL
Nfa45560	VPMGGRASYTDWVADDP -----	VLGRNKWSTFLTAELP---PLL
Nfa25110	MPIGGRFSLYTDWQADDP -----	VLGRNRWSTFLTREL P---AAM
Nfa5610	LPVGGQASYYTDWQTDDP -----	KFGRYKWETFFLTREL P---PII
Dip2339	MPTEAGGTNYTDWNETDP -----	YLGRAKWETFLLIKELPGVLVQP
cmytA	LPVGGESSFYSDWEGPNNGK -----	NYQWETFLLTQELA--PILD
Ce2709	LPVGGESSFYTDWNEPNNGK -----	NYQWETFLLTEELA--PILD
Dip2193	LPVGGQSSFYSDWQPNNGK -----	HYKWETFLLTNELV--PVLK
Nfa1840	LPVGGQSSFYADWMQPNNGR -----	NYKWETFLLTKELP--PLLE
Ce2710	MPVGGAGSFYADWNHPATLSSA -----	EPVVYMWETFLLTAELP--AYLE



```

cmytB      MPVGGAGSFYADWNSQASLSSS ----- DPVIYMWETFLTQELP--AYLE
Dip2194    MPVGGAGSFYTDWVGPAGPQN ----- AIYRWETFLTQELP--GYLA
Nfa1820    MPVGGMSSFYADWNAPSEFFGIPAGS ---- GSSSGSGALNAFTGGPGKSYRYQWETFLTNELR--WALR
Nfa1810    MPVGGMSSFYADWNAPSTILGIGGGSSGSASGSSSGSGALQMFAGGPGKSTRYTWTWETFLTNNLR--WALR
Nfa1830    MPVGGQSSFYTDWYAPSNTN ----- GQKTTYKWETFLTQELP--NFLA
Rv3804c    MPVGGQSSFYSDWYQPACGK ----- AGCQTYKWETFLTSELQ--GWLQ
1f0pA      MPVGGQSSFYSDWYSPACGK ----- AGCQTYKWETFLTSELQ--QWLS
Rv0129c    MPVGGQSSFYTDWYQPSQSN ----- GQNYTYKWETFLTREMP--AWLQ
Rv3803c    APAGGAYSMYTNWEQDG ----- SKQWDTFLSAELP--DWLA
          *           *
          |           |
          70          80
          |           |
          90          100         110

cmytD      EEELNFNHGHRIGGLSMGATGAVHLANSNPDLFDGVIIGISGCYSTLDPIGQTTVSLIVNSR--GGNVENMW
Ce1488     EEELNFNHGHRIGGLSMGATGAVHLANANPDDFAVIGISGCYSTLDPIGQATVSLIVKSR--GGDVENMW
cmytE      EAAIGADGQRSIVGMSMSGGSVLFNFATHDPNFYSSVGSFSGCAETNSWMGRRGIAATAYNG--NVVPEQIF
cmytF      EAAIGADGQRSIVGMSMSGGSVLFNFATHDPNFYSSVGSFSGCAETNSWMGRRGIAATAYNG--NVVPEQIF
Ce0984     EAAIGGDGQRSIIIGMSMSGGSVVNIASHQPNFYSSVASLSGCAETNSWMGRRGVAATVYSG--NATPTQIF
cmytC      EEKLNLDGQRAIAGMSMATSLLFPQHFPGFYDAAASFSGCAATSSLLPWEYKLTLDLRG--NATPEQMW
Ce0356     EEELNADGQRAIAGMSMATSLLFPQHYPGFYDAAASFSGCASTSQPLPWEYIRLTLDRG--NATPEQMW
Dip0365    ERHLGASNKRAIAGLSMATSALVLAEHAQGFYDAAGSFSGCAATSSPLTYHFLRLTLERG--GATPEQMW
Nfa38260   DAALGTNGVNALAGLSMSGTSALQLPIAAPGLYRAVAAYSGCAQISDPVGHFVATVVAAG--HGDVVNMY
Nfa7210    DGALGTNGINAIAGLSTSGTTVLALPIAKPGLYKAAAAYSGCAQTSDPVGSFVKLTVETWGGGDTENMW
Nfa32420   DATLRSTGRNALTGLSTSGTSLQLAEAKPGLWRSVAAYSGCAQIADPTGRQFVKLAVETWAGGDTENMY
Nfa23770   DATFGATGANALAGISMGSSVLDLTIQAPTRYRSVAAYSGCAMSDDLGRMFTVVISLG--GGDPENMW
Nfa30260   DNAFRGQSFYNAVIGVSMAGTSVFQLALHAPGVYRAIGSFSGCVPTSDARGRAVVNTVVRAY--GGDPVNLW
Nfa43800   DEHFHSGGANAVAGVSMAGTSVFQLALAAPGLYRAIGSFSGCVRTSDPQQQVMVNAVVAASH--RGNPVMNW
Nfa45560   EQRFGMTGRNAVAGLSMATSALNLDLADPGRYQAVGAYSGCARTSDPAGRALIYAELAVF--GANATNMW
Nfa25110   TPWLGATGRDAIAGVMSAASALDLAIQAGDRYRAVAAYSGCPWRADPPGMTLVAAQVLRG--GNNPVMNW
Nfa5610    DAQFAGNGVNGIGGLSMGGNAAYILAARNPHLYTAVAGYSACPDGTGLATG AVMFSIANR--GGNPLNMW
Dip2339    ETKIAYNGKSYIGGLSMGSSAAVRLANLYPEKVFVGTFGVSGCYSPVNTSGRELFNLAARVI--GGNPDLMW
cmytA      KGFRSN--TDRAITGISMGGTAAVNIATHHPDMFKFVGSFSGYLDTTSGMPIAI SAALADAGGYDANAMW
Ce2709     KGFRSN--GERAITGISMGGTAAVNIATHNPEMFNFVGSFSGYLDTTSNMGMPAAIGAALADAGGYVNVAMW
Dip2193    NGFRTN--DDRRAVGLSMGGTAAINLAERRPDLKFKFVGSFSGYLDTTSIGMPAAIRAAQRDAGGYDSTAMW
Nfa1840    SQWRAT--DVRGMQGLSMGGTAAMFLAGRNPGFVRYAASYSGFLTTTTLGMPQAIQFAMRDAGGFDSAAMW
Ce2710     QHFGVARNNNSVAGLSMGGTAAINLAAKHPGQFRQAMSYSGYLTTTAPGMQTMRLRLAMLDTGGFNVNAMY
cmytB      QNFVGARNNNSIGGLSMGGTAAINLAAKHPDQFRQAMSWSGYLNTTAPGMQTLRLRLAMLDTGGFNVNAMY
Dip2194    ANFGVSPNTNSIAGLSMGATAAMNLAALHPDQFRQVLSYSGYLSMSVPGTYLMMTLALQEVGGFNINMY
Nfa1820    DRLGFNPNRNGVFLSMGSSAALT LAAYHPDQFSFAGSFSGYLNISAPGMREAIRVAML DAGGYNV DAMA
Nfa1810    DRLGFNPNRNGVFLSMGSSAALT LAAYHPDQFSYAGSYSGYLNVSAPGMREALRVAML DAGGYNIDAMA
Nfa1830    G--YGVSKTNNNAVAGLSMGGSAALALAAAYHRDQFKYAAASYGYLNISAPGMREAIRIAMLDAGFRNVDSMA
Rv3804c    ANRHVKPTGSAAVGLSMAASSALTLAIYHPQQFVYAGAMSGLLDPSQAMGPTLIGLAMGDAGGYKASDMW
1f0pA      ANRAVKPTGSAAIGLSMAGSSAMI LAAYHPQQFIYAGLSALLDPSQMGPSLIGLAMGDAGGYKAADMW
Rv0129c    ANKGVSPGTGNAAVGLSMGSSALILAAAYPPQFPYAAASLSGFLNPSEGWWPTLIGLAMNDSGGYNANSMW
Rv3803c    ANRGLAPGGHAAVGAQQGGYGAMALAAFHDPDRFGFAGSMGFLYPSNTTTNGAIAAGMQQFGGVDTNMGW
          *           *
          |           |
          120          130          140          150          160          170          180

cmytD      GPTGSETWKAHDVTSNPEGLRD--MAVYLSAANGVVDIDLAD ----- SEKEPFYNLLAGVVLE
Ce1488     GPVGSRTWQEHVVSNSPEGLRN--MAVYLSAANGVVDDEIDREE ----- YADEFYNLLAGTVLE
cmytE      GEVDSYSDYRYNDPLLNAAKLEEQ--DNLYIFAGSGVFSELDVI ----- GDNAPIDEDAFKNRVLVGFIE
cmytF      GEVDSYSDYRYNDPLLNAAKLEEQ--DNLYIFAGSGVFSELDVI ----- GDNAPIDEDAFKNRVLVGFIE
Ce0984     GEVDSYARYNDPVINAHRLEKQ--DNLYVFAASGVWSEVDVE ----- GENAPEDEKGLKNRITVGFRIE
cmytC      GPRGGEYNIYNDALINSDKLRG--TELYVSNASGLAGEWESVDSPRFEGLNQQVQSIAMAETVVTGGIE
Ce0356     GPRGGEVNIYNDALINSDKLRG--TDLYISNAGSLAGHWESANSPRFNGLDQAYLSLAMTETIVTGGIE
Dip0365    GPQGSEVNRYNDALINAERLRG--TEVYVSNNSGAVGKYDLPPSPRLAGKDPVTIFATNLITATEGGIE
Nfa38260   GPPDDPMWAANDPYVQAERLRG--LELFLSTGTGFLPGKWDTLN ---- GPHAMPGSDGLTNQVLVGGILE
Nfa7210    GPPGSEEWKNDPYVNAEGLRG--LELYISTGNGIPGPDYDTLN ---- GPYALPGSYGLANQILIGGVIE
Nfa32420   GPDDSPWLWRENDPVVNAEKLGRG--TQLYISTGSGIPVLEDVQY ---- YLNAAPGPMGAVN--LGLGVIE
Nfa23770   GPTGGDGWREHPYLQAHRLPP--IPMYISSGSLPGPHDTLA ---- NPLRHNDRLNQLNLVGGAE
Nfa30260   GPPEDPAWAANDPSLRAAEALRD--TAVYVYTAGTGRPGALDSLQ ---- GPGIDADPLALADQLLIGGALE
Nfa43800   GPPTDPTWRANDPYLHADRLRG--TAYIISGSGLPGLDNP ----- AAVGGDPMQLRQLVLFAPLE
Nfa45560   GGPDSPLWAAHPVLRRAEELRG--LAIYVSAGDGRPRHETLT ---- APGIDGNPLDLVERTVVGGLME
Nfa25110   GPPGDPGWQSHDAFRNAGALAG--KTVYLSAASGIPGPI DRG ----- GLPAPTLE
Nfa5610    GPPGSPAWAEHDPARLAGNLRG--KTYLSTGTGIPGPHEAEL-----KP ----- QLAENIFLGGPVE
Dip2339    GRDITEQRRRNDVVANPSGIAS--MDTYIYVANGVATPSDVNG ----- PKEDGFPTLFGNIVLE

```



