Hallmarks of Mycolic Acid Biosynthesis: A Comparative Genomics Study

Karthik Raman
{karthik@rishi.serc.iisc.ernet.in}

Preethi Rajagopalan
{preethi.rajagopalan@gmail.com}

Nagasuma Chandra*
{nchandra@serc.iisc.ernet.in}

*Author for correspondence:
Dr. Nagasuma Chandra
Bioinformatics Centre and Supercomputer Education and Research Centre
Indian Institute of Science
Bangalore - 560012
INDIA

Phone:
+91-80-22932469
+91-80-23601409

Fax: +91-80-23600551
E-Mail: nchandra@serc.iisc.ernet.in

Short Title: Hallmarks of Mycolic Acid Biosynthesis

Keywords: Comparative genomics; mycobacteria; phylogenetic profiling; pathway profile; emergent properties.

Abstract

Purpose: Mycolic acids, which render unique qualities to mycobacteria, are known to be important for mycobacterial growth, survival and pathogenicity. It is of interest to understand the evolutionary origins of the mycolic acid pathway (MAP), as well as the common minimum principles critical for generating the capability of mycolic acid biosynthesis. The recent curation of a comprehensive model of the MAP in *Mycobacterium tuberculosis* and the availability of a large number of genome sequences make it feasible to carry out detailed sequence and phylogenetic analyses, to address these questions.

Methods: A comprehensive phylogenetic pathway profile analysis was carried out for 318 fully sequenced bacterial genomes, for each of the proteins present in the MAP. The organisms were clustered based on the co-occurrence of the MAP proteins in their proteome, while the proteins were clustered based on their phylogenetic profiles. The MAP proteins were also searched against the non-redundant sequence database, to identify similar proteins from other phyla.

Principal Findings: The pathway profiles indicate that four proteins and certain protein domains stand out as more characteristic to mycolate producing organisms. Further analysis leads to the identification of the desaturases DesA1 and DesA2 and certain domains of Fas and Pks13 as hallmarks of the pathway. The roles of these proteins in some other organisms, as well as the distribution of these proteins across all known genome sequences are also briefly discussed. The clustering of organisms, carried out to group organisms with similar profiles, provides a means of obtaining finer classification as compared to the standard taxonomic method. The results indicate that the MAP and hence the capacity of mycolic acid production in mycobacteria is an example of an emergent property that has come about by recruiting enzymes from unrelated pathways in plants, presumably through lateral gene transfer. The understanding of the hallmarks of mycolic acid biosynthesis will also find application in evaluating drug targets.

1 Introduction

Mycolic acids have been one of the defining features of bacteria such as *Mycobacterium tuberculosis* and *Mycobacterium leprae*, belonging to the Actinomycetales order [1]. Chemically, they are long chain high molecular weight α-alkyl-β-hydroxy fatty acids consisting of an α-meromycolate moiety, with carbon chain lengths of up to C_{56} and a long saturated α-branch of C_{20} to C_{24}. The mycolic acids of the pathogenic mycobacteria differ from those of other related genera such as *Corynebacterium*, *Nocardia* and *Rhodococcus*, in that they are the largest (C_{70} to C_{90}) [2]. These compounds, which form the principal components of mycobacterial envelopes, confer unique properties to these organisms, such as unusually low permeability and consequent resistance to common antibiotics [2]. They are also known to exhibit adjuvant immunological properties [1]. It comes as no surprise therefore, that mycolic acids have been a subject of several investigations [1,3,4], not only because of their unique chemical nature but also because the enzymes involved in their biosynthesis form excellent targets for the design of anti-tubercular drugs. In fact, it turns out that the existing front-line anti-tubercular drugs such as isoniazid [5] and ethionamide [6] owe their therapeutic properties to the inhibition of one or the other proteins in the mycolic acid biosynthesis pathway.

The components of the mycolic acid pathway (MAP) have been identified by a number of groups; the components have been curated into a comprehensive pathway model in [7], comprising 197 metabolites participating in 219 biochemical reactions catalysed by 28 proteins. Figure S1 shows a schematic representation of the MAP.

The MAP can be considered to be made of four sub-pathways: (A) production of malonyl CoA, (B) FAS-I pathway, (C) FAS-II pathway and (D) condensation of FAS-II and FAS-I products into α- (D1), methoxy- (D2) and keto-mycolic acids (D3), as illustrated in Figure S1. Much of the early part of the pathway utilises the general pathways of fatty acid biosynthesis seen in most organisms. Fatty acid synthases (FAS) are of two types: (a) FAS-I, present predominantly in eukaryotes, which contain the various catalytic activities as discrete functional domains, either on a single polypeptide chain or, in some cases, on two different multi-functional proteins of comparable size and (b) the FAS-II, seen in bacteria, which consists of independent proteins encoded by a series of separate genes [8]. Mycolic acid biosynthesis in mycobacteria requires both FAS-I, capable of *de novo* fatty acid synthesis and FAS-II, which only elongates the products of FAS-I [1]. The production of malonyl-CoA in (A) is crucial for FAS-I and FAS-II ((B) and (C) sub-pathways) as it adds on two carbons to the elongating acyl-chain. The products of (B) and (C) are then condensed and processed into different mycolic acids in (D).

During analysis of the sequences of the MAP proteins, close sequence homologues for most of them were observed in many other organisms, besides mycobacteria. This observation immediately begs a question as to what are the determinants of mycolic acid biosynthesis, given that very few organisms are known to be capable of it. Here, we seek to address this question and identify determining features of the mycolic acid biosynthesis, through a genome sequence analysis and a phylogenetic study. We also explore a possible molecular taxonomic classification of prokaryotes, based on the profiling of the MAP proteins.

2 Results

Sequence analyses were carried out to identify homologues for each of the 26 proteins of the MAP across all available genome sequences. Close homologues were observed for all the 26 proteins of *M. tuberculosis* H37Rv in all the other available tuberculosis related genomes such as *M. tuberculosis* CDC1551, *M. avium paratuberculosis* and *M. bovis*, thus leading to identical pathway landscapes in these organisms. Close homologues for some of the proteins were also observed in many other organisms, besides mycobacteria.

Two of the proteins in the pathway, Fas and Pks13 happen to be very large, having 3069 and 1733 amino acids respectively. Both are made up of several domains. In order to account for the presence of smaller proteins corresponding to individual domains across the genomes, we felt it necessary to analyse each of the domains in these two proteins separately, as if they were individual proteins themselves. The phylogenetic profiles discussed in the subsequent sections are all based on domain-wise analysis for these two proteins, along with the profiles of the entire proteins for the rest of the pathway. Table I shows the domains of Pks13 and Fas and their locations in the respective sequences.

2.1 Description of the pathway profile

Figure 1A illustrates the profile of the MAP, composed of the phylogenetic profiles of all the individual proteins/domains. The clustering here is based on the patterns of occurrence and the extents of similarities of individual proteins across all 318 genomes. It must be noted that close homologues of the MAP proteins have been removed and only a 'non-redundant' subset of the MAP proteins have been used in clustering. The clustering is more clearly seen in Figure 1B.

Two distinct clusters can be discerned from the Figure 1B. The second cluster contains those proteins which exhibit similar profiles, indicating that these proteins by and large coexist. The proteins in the first cluster on the other hand, are more ubiquitous as they are seen in most of the 318 organisms studied here. Thus, the second cluster is of direct interest to this study. Strong similarities in profiles are observed in the sets {DesA2, Pks13_1}, {Fas_6, Fas_7} and {Fas_2/3, DesA2/3} strongly suggestive of functional linkages with each other. Functional linkages in general could indicate the possibility of the corresponding proteins: (a) interacting with each other forming structural complexes, or (b) catalysing sequential reactions in a pathway, or at least (c) catalysing reactions that are dependent on one another directly or indirectly, through shared metabolites or regulatory linkages. Given our understanding of the components of the MAP and the sequence of the reactions therein, the linkages observed here appear to belong mainly to the third type. This further implies that if the products of the individual reactions are made in the same organism, they are strongly suggestive of the existence of a functional pathway.

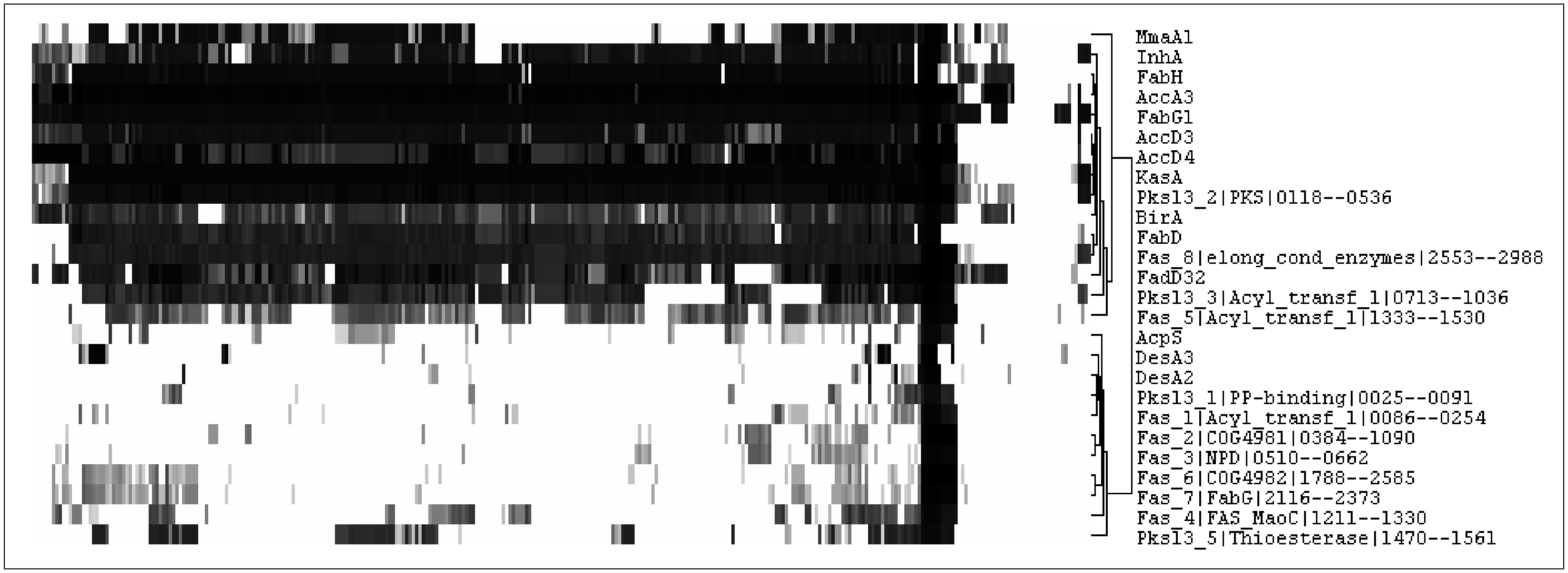

Figure 1A: **Phylogenetic profiles of MAP proteins: Profiles of all non-redundant proteins of the MAP but with Fas and Pks13 dissected into their component domains.** White indicates the absence of a match, while black indicates a match with full identity. Grey shading reflects the extent of similarity with the corresponding protein from *M. tuberculosis*. The figure was generated using jTreeView.

Although most of the individual domains of Fas have grouped together into the second cluster, two domains (Fas_5 and Fas_8) show very different profiles and are therefore a part of the first cluster. Similarly, Pks13_2 and Pks13_3 exhibit very different patterns from that of Pks13_1 and Pks13_5, the latter two, clustering into distinct sub-clusters. The profiles in Figure 1A indicate that proteins in Cluster 2 are much more sparsely distributed across bacterial genomes, as compared to those in Cluster 1. Figure 2, which illustrates a uniqueness plot of the MAP proteins/domains indicates that three proteins — AcpS and the desaturases DesA2 (also DesA1) and DesA3 — and many domains of Fas and the Phosphopantetheine attachment site of Pks13 stand out as absent in majority of the genomes. Further, domains of Fas, viz. Fas_1, Fas_2, Fas_3, Fas_6, Fas_7 and Pks13, viz. Pks13_1 are also found to be more characteristic to mycobacteria and closely related Actinomycetales, that are capable of making mycolic acids. The uniqueness plot thus provides a ready first list of candidate proteins that are hallmarks of the pathway.

2.2 Clustering of organisms based on pathway profiles

Information obtained from the profiling of the MAP proteins across 318 genomes also helps us to cluster the organisms, as illustrated in Figure 3. The organisms present in the same sub-cluster as *M. tuberculosis* (marked as •, coloured navy) have homologues for many (or all) of the proteins of the MAP. These organisms are inevitably the mycobacteria (*M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551, *M. avium paratuberculosis*, *M. bovis* and *M. leprae*), as well as the closely related *Streptomyces* and *Nocardia* species. *M. leprae* has most of the proteins, but lacks even a distant homologue for DesA3 and hence clusters in a different sub-cluster from the rest of the mycobacteria. Organisms from the *Corynebacterium* species also cluster close to mycobacteria (indicated by ◆, coloured maroon), but they do not contain homologues for FabH, FabD and DesA2. The organisms in the cluster marked by ▼ (coloured gray) do not have any of the MAP proteins — they are mainly the *Mycoplasma* and a few archaea. All organisms belonging to the Actinomycetales class have been indicated in bold and underlined; those occurring outside the coloured sub-clusters have also been marked with a ■. This tree provides a systematic functionally oriented basis for classification of organisms, in the phylogenetic space focussed around the mycobacteria, given that a pathway is a fundamental unit of the functional capabilities of an organism. This kind of molecular taxonomy of the organisms is also readily useful in understanding the evolutionary trace of the pathway being considered. For example, sub-clusters closest to those of the mycobacteria such as the one containing the Corynebacteria have most but not all of the proteins of MAP, indicating an intermediate step in the evolution of the pathway. Similarly, bacteria such as *Propionibacterium acnes*, which have clustered reasonably away from the mycobacteria have few but not most of the proteins of the MAP. A conventional taxonomic classification places many of these into a common class of Actinomycetales. On the other hand, *Bifidobacterium longum*, which is not an Actinomycetales, but belonging to Actinobacteridae, clusters

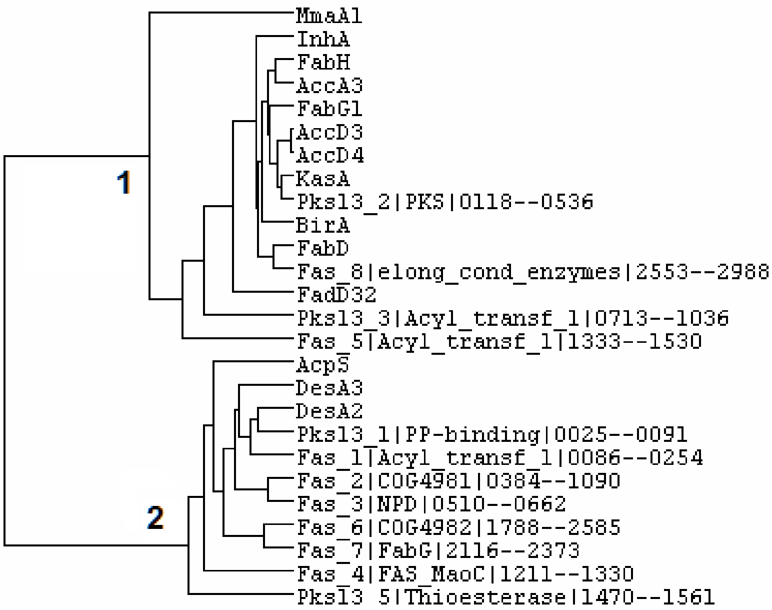
1
2
MmaA1
InhA
FabH
AccA3
FabG1
AccD3
AccD4
KasA
Pks13_2|PKS|0118--0536
BirA
FabD
Fas_8|elong_cond_enzymes|2553--2988
FadD32
Pks13_3|Acyl_transf_1|0713--1036
Fas_5|Acyl_transf_1|1333--1530
AcpS
DesA3
DesA2
Pks13_1|PP-binding|0025--0091
Fas_1|Acyl_transf_1|0086--0254
Fas_2|COG4981|0384--1090
Fas_3|NPD|0510--0662
Fas_6|COG4982|1788--2585
Fas_7|FabG|2116--2373
Fas_4|FAS_MaoC|1211--1330
Pks13_5|Thioesterase|1470--1561

Figure 1B: **Clustering of the MAP proteins/domains, based on the similarity of their profiles across the genomes.**

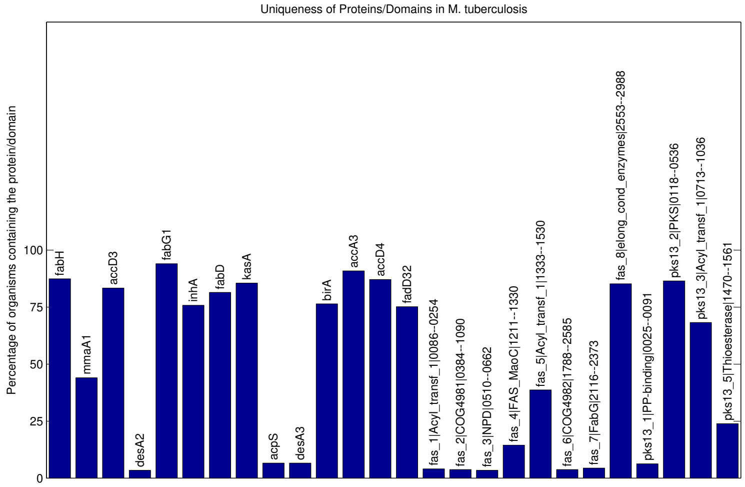

Figure 2: **A uniqueness plot of the MAP proteins/domains across different genomes.** The bars show the percentage of organisms in which a particular protein/domain of the MAP was present. Profile scores < 0.1 were taken as indicative of presence of a very good match.

close to the mycobacteria. While the two systems of classification agree by and large, significant differences in finer classification can be clearly discerned from this analysis.

2.3 Analysis of the putative determining features of MAP

The proteins determined from the analysis of profiles and uniqueness have been further analysed in terms of their individual functional roles to obtain a final list of determinant proteins. The functional annotations for all of the proteins of the MAP are shown in Table S2. The broad functional roles of AcpS and the desaturases, as well as some of the domains of Fas and Pks13, which were more characteristic of the mycolate producers are elaborated below:

DesA1 and DesA2: These are homologous to the plant stearoyl-ACP desaturase which introduce the first double bond in the saturated fatty acids, C_{16} and C_{18}, the products of fatty acid biosynthesis. These fatty acids are then incorporated in the membrane glycerolipids, cuticular lipids and oilseeds of plants [10]. Involvement of these proteins in mycolic acid synthesis has been suggested based on sequence annotations [1,7,11] and structural characterisation [12]. However, experimental evidence regarding their functional roles are not presently available.

DesA3: DesA3, annotated as a linoleoyl-CoA desaturase (Δ^6-desaturase), is found across different phyla and is essential for the synthesis of nutritionally important γ-linolenic acid and stearidonic acid. Δ^6-desaturases introduce Δ^6-double bonds, usually converting linoleic acid ($\Delta^{9,12}C_{18:2}$) and α-linolenic acid ($\Delta^{9,12,15}C_{18:3}$) into γ-linolenic acid ($\Delta^{6,9,12}C_{18:3}$) and stearidonic acid ($\Delta^{6,9,12,15}C_{18:4}$), respectively [13].

AcpS: Acyl-carrier-protein synthase (AcpS) is an essential enzyme in the biosynthesis of fatty acids in all bacteria. AcpS catalyzes the transfer of 4′-phosphopantetheine from coenzyme A (CoA) to apo-ACP, thus converting apo-ACP to holo-ACP that serves as an acyl carrier for the biosynthesis of fatty acids and lipids. However, the homology of other bacterial AcpS proteins with *M. tuberculosis* AcpS is quite poor.

Pks13_1/4 Phosphopantetheine attachment site: A 4′-phosphopantetheine prosthetic group is attached to the protein through a serine. This prosthetic group acts as a 'swinging arm' for the attachment of activated fatty acid and amino-acid groups. This domain forms a four helix bundle. *Fas_1 Acyl transferase domain I:* This is associated with the transferase activity of the enzyme. It is present in enzymes such as bacterial malonyl CoA-acyl carrier protein transacylase and eukaryotic fatty acid synthase that are involved in fatty acid biosynthesis [14]. *Fas_2 Enoyl reductase domain:* Enoyl reductase domain of yeast-type FAS-I, which is equivalent to InhA of the FAS-II Cycle, which reduces the β-enoyl intermediate, to yield a four carbon acyl substrate for further cyclic elongation. *Fas_3 2-Nitropropane dioxygenase domain:* Members of this family catalyse the denitrification of a number of nitroalkanes using either FAD or FMN as a cofactor. *Fas_6 3-oxoacyl-[acyl-carrier-protein]:* The reaction cycle of

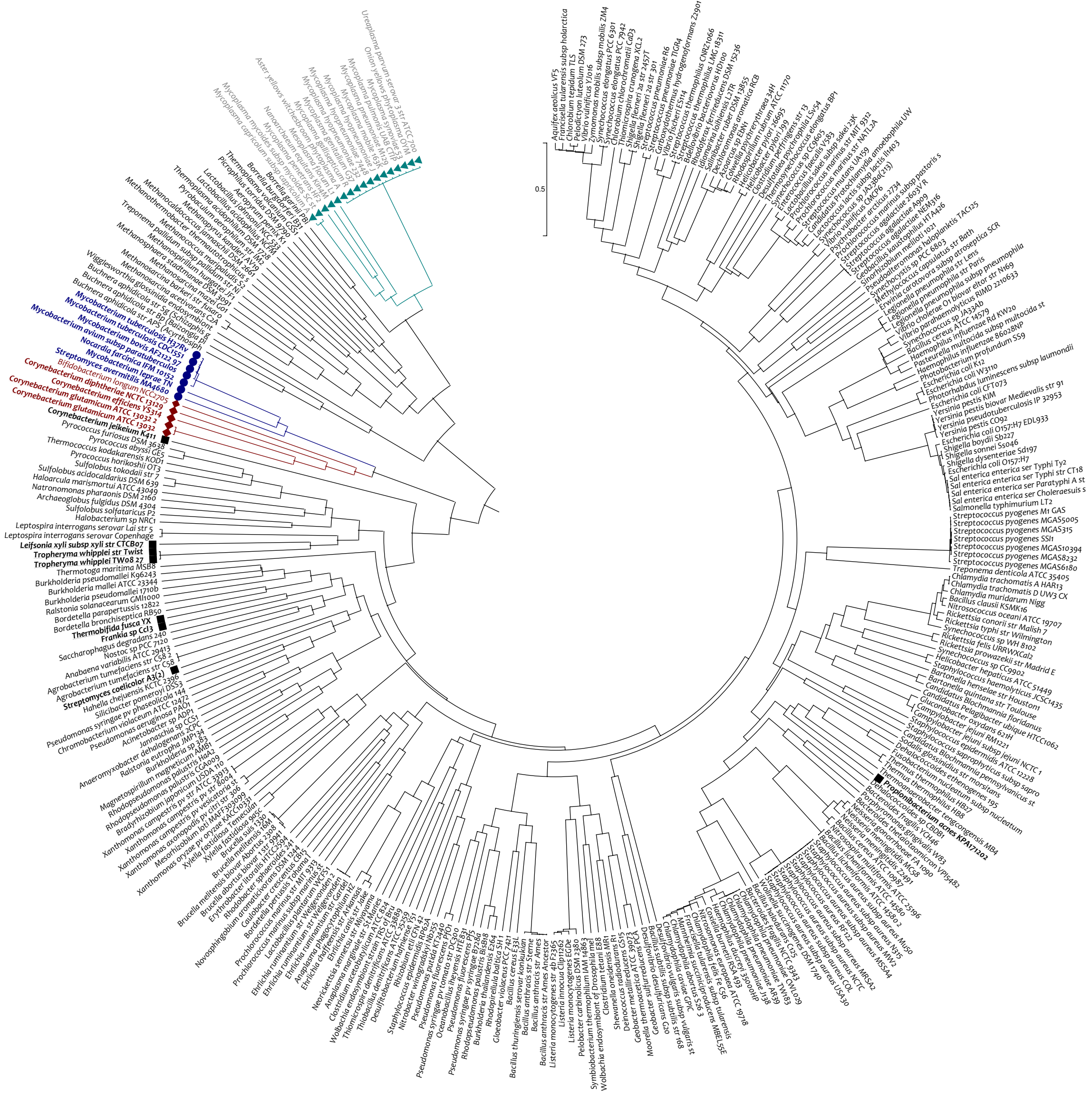
0.5
Aquifex aeolicus VF5
Francisella tularensis subsp holarctica
Chlorobium tepidum TLS
Zymomonas mobilis subsp mobilis ZM4
Synechococcus elongatus PCC 7942
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus agalactiae NEM316
Legionella pneumophila str Paris
Legionella pneumophila str Lens
Vibrio parahaemolyticus RIMD 2210633
Pasteurella multocida subsp multocida str
Haemophilus influenzae 86028NP
Photobacterium profundum SS9
Photorhabdus luminescens subsp laumondii
Escherichia coli W3110
Escherichia coli K12
Escherichia coli CFT073
Yersinia pestis KIM
Yersinia pestis biovar Medievalis str 91
Yersinia pseudotuberculosis IP 32953
Yersinia pestis CO92
Escherichia coli O157:H7 EDL933
Shigella boydii Sb227
Shigella sonnei Ss046
Shigella dysenteriae Sd197
Escherichia coli O157:H7
Sal enterica enterica ser Typhi Ty2
Sal enterica enterica ser Typhi str CT18
Sal enterica enterica ser Paratyphi A st
Sal enterica enterica ser Choleraesuis s
Salmonella typhimurium LT2
Streptococcus pyogenes M1 GAS
Streptococcus pyogenes MGAS5005
Streptococcus pyogenes MGAS315
Streptococcus pyogenes SSI1
Streptococcus pyogenes MGAS10394
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes MGAS6180
Treponema denticola ATCC 35405
Chlamydia trachomatis A HAR13
Chlamydia trachomatis D UW3 CX
Chlamydia muridarum Nigg
Bacillus clausii KSMK16
Nitrosococcus oceani ATCC 19707
Rickettsia conorii str Malish 7
Rickettsia typhi str Wilmington
Synechococcus sp WH 8102
Rickettsia felis URRWXCal2
Rickettsia prowazekii str Madrid E
Synechococcus sp CC9902
Helicobacter hepaticus ATCC 51449
Staphylococcus haemolyticus JCSC1435
Bartonella henselae str Houston
Bartonella quintana str Toulouse
Campylobacter jejuni RM1221
Propionibacterium acnes KPA171202
Thermus thermophilus HB8
Bacteroides fragilis YCH46
Bacteroides fragilis NCTC 9343
Listeria monocytogenes EGDe
Listeria monocytogenes str 4b F2365
Bacillus anthracis str Sterne
Bacillus anthracis str Ames
Bacillus anthracis str Ames Ancestor
Bacillus thuringiensis serovar konkukian
Gloeobacter violaceus PCC 7421
Pseudomonas fluorescens Pf5
Rhodopseudomonas palustris CGA009
Ralstonia eutropha JMP134
Acinetobacter sp ADP1
Pseudomonas aeruginosa PAO1
Chromobacterium violaceum ATCC 12472
Pseudomonas syringae pv phaseolicola 1448A
Silicibacter pomeroyi DSS3
Hahella chejuensis KCTC 2396
Streptomyces coelicolor A3(2)
Agrobacterium tumefaciens str C58
Agrobacterium tumefaciens str C58 2
Anabaena variabilis ATCC 29413
Nostoc sp PCC 7120
Saccharophagus degradans 240
Frankia sp CcI3
Thermobifida fusca YX
Bordetella bronchiseptica RB50
Bordetella parapertussis 12822
Ralstonia solanacearum GMI1000
Burkholderia pseudomallei 1710b
Burkholderia mallei ATCC 23344
Burkholderia pseudomallei K96243
Thermotoga maritima MSB8
Tropheryma whipplei TW08 27
Tropheryma whipplei str Twist
Leifsonia xyli subsp xyli str CTCB07
Leptospira interrogans serovar Copenhage
Leptospira interrogans serovar Lai str 5
Halobacterium sp NRC1
Sulfolobus solfataricus P2
Archaeoglobus fulgidus DSM 4304
Natronomonas pharaonis DSM 2160
Haloarcula marismortui ATCC 43049
Sulfolobus acidocaldarius DSM 639
Sulfolobus tokodaii str 7
Pyrococcus horikoshii OT3
Thermococcus kodakarensis KOD1
Pyrococcus abyssi GE5
Pyrococcus furiosus DSM 3638
Corynebacterium jeikeium K411
Corynebacterium glutamicum ATCC 13032
Corynebacterium glutamicum ATCC 13032 2
Corynebacterium efficiens YS314
Corynebacterium diphtheriae NCTC 13129
Bifidobacterium longum NCC2705
Streptomyces avermitilis MA4680
Mycobacterium leprae TN
Nocardia farcinica IFM 10152
Mycobacterium avium subsp paratuberculosis
Mycobacterium tuberculosis CDC1551
Mycobacterium bovis AF2122 97
Mycobacterium tuberculosis H37Rv
Buchnera aphidicola str APS (Acyrthosiph
Wigglesworthia glossinidia endosymbiont
Treponema pallidum subsp pallidum str Ni
Methanococcus maripaludis S2
Lactobacillus johnsonii NCC 533
Borrelia garinii PBi
Borrelia burgdorferi B31
Nanoarchaeum equitans Kin4 M
Aster yellows witches broom phytoplasma
Onion yellows phytoplasma OY M
Mycoplasma genitalium G37
Mycoplasma pneumoniae M129
Mycoplasma mobile 163K
Mycoplasma synoviae 53
Mycoplasma penetrans HF 2
Ureaplasma parvum serovar 3 str ATCC 700970

Figure 3: **A circular view of the phylogenetic tree of the organisms, based on the profiles of MAP proteins.** The significance of the highlighted sub-clusters are explained in the text. The figure was generated using MEGA3 [9].

FAS is initiated by the transfer of the acyl moiety of the starter substrate acetyl-CoA to the acyl carrier protein. *Fas_7 Dehydrogenases with different specificities:* This domain, related to short-chain alcohol dehydrogenases, is equivalent to FabG of the FAS-II cycle, involved in the production of β-hydroxyacyl-ACP from a ketoacyl intermediate.

3 Discussion

While single proteins have been regularly profiled across organisms [15], it has now become possible to profile an entire pathway across organisms, to yield insights into the interaction between the pathway proteins, as well as the critical components of a pathway in various organisms. Given that specific interactions between proteins ultimately give rise to specific phenotypes, it is important to consider the properties of all the pathway components simultaneously. Analysing the patterns of co-occurrence of the proteins as shown in Figure 1A also provide insights about possible functional linkages. Phylogenetic profiles have been used extensively to infer possible interactions among pairs/groups of proteins [15–17].

It must be noted, however, that a limitation of sequence based phylogenetic analysis is the inability to identify functional analogues whose sequences have diverged substantially beyond recognition by standard sequence analysis methods. The accuracy of the annotations of hypothetical proteins also influences the level of confidence in the interpretations. Nevertheless, given that most protein families exhibit characteristic sequence signatures, sequence-based analyses have proved to be invaluable in obtaining hypotheses that can be verified by biochemical and other experimental studies.

3.1 Hallmarks of mycolic acid synthesis and the emergence of mycolic acid synthesis capability

Mapping of the profiles of the proteins that stand out in Figure 1 to the sub-clusters containing the corresponding organisms (Figure 3) indicates the predominant absence of four proteins, viz. DesA2 (hence DesA1 also), DesA3 and AcpS in all except the sub-cluster belonging to the mycobacteria (coloured navy, marked • in Figure 3). The profiles of these four proteins appear to be characteristic of the mycobacteria that produce mycolic acids, leading to their consideration as putative determining features of mycolic acid pathway.

A closer examination of the individual sequence alignments indicates that close homologues of Fas and Pks13 over their entire lengths are present only in other mycolate-producing organisms such as *M. leprae*, *C. glutamicum* and *Nocardia farcinica*. This suggests that, possibly, the Fas system is also a characteristic feature of mycolic acid synthesis among bacteria. However, close homologues for some of the domains are present in many bacteria (Figure 1), whereas certain domains of Fas, viz. Acyl transferase domain 1 (Fas_1), Enoylreductase domain (Fas_2), 2-nitropropane dioxygenase (Fas_3), 3-oxoacyl-[acyl-carrier-protein] (Fas_6) and Dehydrogenases (Fas_7), as well as the Phosphopantetheine attachment sites (Pks13_1/4) of Pks13 are relatively unique to the mycobacteria.

To further analyse the validity of these proteins as the characteristic features of the pathway, the functional annotations for all of the proteins of the MAP were examined. AcpS appears to have functional counterparts, which are sequence homologues at lower levels of similarity in some organisms. This is not surprising given that AcpS is involved in the initial steps of fatty acid synthesis in most organisms and hence caters to a broader requirement than that of mycolic acid biosynthesis. AcpS cannot be classified as an ideal hallmark of mycolic acid synthesis. The desaturases, on the other hand, do not have identifiable homologues in most bacteria, even with relaxed similarity criteria. The involvement of DesA3 in mycolic acid synthesis is also not clear [3], but it has been shown to be involved in oleic acid synthesis, a component believed to be an essential constituent of mycobacterial membrane phospholipids and triglycerides [18, 19]. Moreover, *M. leprae*, which is known to make mycolic acids, does not have a close sequence homologue for DesA3, again questioning the role of DesA3 in the pathway. This leaves DesA1 and DesA2 and domains of Fas and the Pks13_1 domain, which can be thus considered as hallmarks of the pathway, or in other words, indicative of mycolic acid production capability.

It is perhaps logical for desaturases to be regarded as hallmarks of mycolic acid biosynthesis, considering the nature of the biochemical reactions that they catalyse. The requirement for the introduction of a double bond, catalysed by DesA1/2 is not as widespread as compared to the other biochemical reactions such as acylation, elongation, biotinylation and condensation, that are required in many other metabolic pathways. Besides the desaturases, it appears that fully functional Pks13 and Fas proteins are required for the production of mycolates. Although proteins moderately homologous to the domains of Fas are present in the FAS-II cycle (MAP (C), Figure S1), it is possible that the divergence in their finer structures and hence their specificities, are significant enough to account for the capabilities for mycolate production.

Further, it may be noted that the genes encoding DesA1/2, Pks13 and Fas have been classified as essential genes for mycobacterial survival [20], whereas those encoding AcpS and DesA3 are not. Given that mycolic acid synthesis is characteristic of mycobacteria and essential for its survival, the presence of the desaturases and AcpS can be readily correlated with such a biochemical capability. A recent study of the pathway by flux balance analysis [7] by us also indicated the proteins DesA1/2, Fas, Pks13 to be among the 16 proteins identified as essential for mycolic acid production. These four proteins/domain components also happen to be characteristic features or determinants of the pathway as discussed in this report.

The distinction between essentiality and the determinant/hallmark of a pathway must be highlighted here. Essentiality of a protein for a pathway does not necessarily imply that it is a determinant of the pathway. This is because homologues of many proteins may be present in the same organism or other organisms, as part of different biochemical pathways. For example, AccA3 is an essential protein for the pathway, but not a determinant.

3.1.1 Functional roles of MAP components in different organisms

Classification of desaturases as determining features is further strengthened when the pathway profile in *Corynebacterium* and *Nocardia* species are considered. Corynebacteria produce smaller mycolic acids, which are short chain (C_{22} – C_{36}) fatty acids. However, they lack the FAS-II enzymes, desaturases and MmaAs and hence are not capable of production of larger mycolates such as α-mycolate, a principal component of *M. tuberculosis*. The corynomycolic acid pathway would correspond to reactions in subpathways (A), (B) and the final steps catalysed by Pks13, Rv2509 (Figure S1), thus bypassing the FAS-II cycle and several steps in the final modifications. *N. farcinica*, on the other hand, make other mycolates, which are longer than the corynomycolates, perhaps more comparable to α-mycolates, though slightly shorter [21], and has identifiable desaturases.

Figure 4 shows a brief overview of the various pathways existing in the organisms, in comparison to *M. tuberculosis*. It indicates the routes of normal fatty acid synthesis as well as mycolic acid synthesis in the various organisms. Further modifications that happen to the fatty acids in the cell are also indicated. The (A) part of the MAP (see Figure S1) appears to be universal and is common to all bacteria. Most bacteria, including *E. coli* and *S. coelicolor* have the (A) and (C) portions of the MAP, as it exists in *M. tuberculosis*. Possibly because the fatty acid intermediates are not routed through (B), the final products do not have very long chains. In *C. glutamicum* as well, the chain extension is not significant. Also, there is no desaturation or cyclopropanation. It is only in *M. tuberculosis* and *N. farcinica* where all the steps — of desaturatation, cyclopropanation and condensation — take place, to yield long chain mycolic acids. It can also be clearly seen that the fatty acid intermediates are recruited to various other pathways in the cell.

Other interesting profiles pertain to *Streptomyces* species, which have almost all the components, except a few domains of Fas. Although these species have very close lineage to *M. tuberculosis*, there is no explicit evidence in literature about their ability to make mycolic acids. These organisms have genes encoding for polyketide synthases, which participate in entirely different pathways, synthesising a range of antibiotics. Thus, the role of the polyketide synthase enzymes in these genomes appears to be in the production of secondary metabolites and not in the modification of fatty acids to produce mycolic acids. A close synteny for the entire *M. tuberculosis* genome with the core of the *Streptomyces* genome has been reported and it is suggested that the two organisms may have had a common ancestor [22, 23]. In fact, it is possible that an important outcome/driving factor for the divergence of the two species from a common ancestor may have been the mycolic acid production and the special features it imparts to *M. tuberculosis*. A detailed sequence and structural analysis of all the components may provide insights into possible evolution of new ligand specificities that may have led to such an altered function.

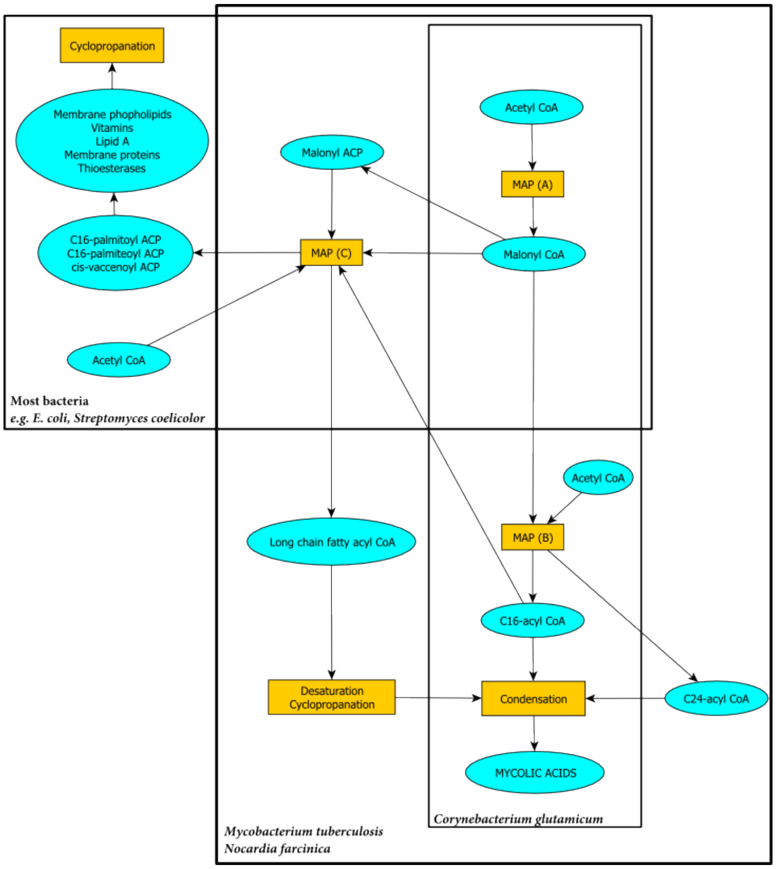

Figure 4: **An overview of the larger roles of fatty acid intermediates and hence of the proteins that synthesise them, in various organisms.** The end products of the intermediates in different organisms are indicated in appropriately labelled boxes. While *M. tuberculosis* and *N. farcinica* have mycolic acids as end products of the pathway, other organisms can be seen to utilise the intermediates for other purposes.

The profiles suggest that the presence of elongation machine followed by condensation and desaturation (e.g. Fas-II, Pks13) is sufficient to produce simple mycolic acids and hence can be regarded as the minimum principles of the MAP. Other enzymes such as Fas and the desaturases appear to be important for the production of longer mycolates such as α-, methoxy- and keto-mycolates. The roles of each of these variants is becoming clearer and it appears that a combination of these as in *M. tuberculosis* is nature's insurance for the bacteria against attack on any one component of the pathway. In fact, this is corroborated by experimental data, which show that the mutant bacilli can grow and survive for limited periods by producing a significant excess of keto-mycolate, to compensate for the absence of α-mycolate [24].

A BLAST [25] search against the non-redundant sequence database indicated the presence of desaturases primarily in plants. The functional roles for many of these are also known, which includes a diverse range such as from the synthesis of plant cell membrane components to linoleic acid and other secondary metabolites, again all unique requirements in plants. It appears that the desaturases have been recruited into the system from the plant genomes through horizontal gene transfer. The suggestions of mycobacteria originating as soil organisms are in line with that observation [26].

It has been shown several times before that most parasitic lineages derive from non-parasitic ancestors with diverse mechanisms employed during the course of evolution [27]. This is particularly striking in parasites, which have to evolve to resist host defences, through mechanisms such as antigenic variation or hiding in 'safe' compartments in the host cell. Many examples illustrate how intracellular parasites have mastered this game of manipulating the structure and pathways of the host cell in order to create a more hospitable environment for themselves. The biosynthesis of mycolic acids is one such clever strategy by the mycobacteria, giving rise to a thick waxy coat, impenetrable to most drugs and natural defence proteins, allowing absorption of required nutrients from the host cell. This has been achieved by utilising the near-ubiquitous fatty acid synthesis pathway to start with and rendered unique by the addition of condensation, elongation, desaturation and cyclopropanation steps. The pathway is a beautiful example illustrating the emergence of new properties by pooling together components from different reaction systems.

3.2 Finer taxonomic insights from pathway profiling

The clustering of organisms also provides a means for achieving a molecular taxonomic classification. In this study, the clustering has been based on the profiles of the MAP proteins. It is observed that most of the Actinomycetales have grouped together, which correlates in general with the conventional taxonomic classification. However, differences in finer details, such as the grouping of *S. avermitilis* and *S. coelicolor* into separate clusters, explained by the extent of conservation of the MAP proteins, serves to demonstrate the utility of using pathways for obtaining finer taxonomic insights in the phylogenetic space focussed around the organism of interest.

Taxonomic insights may also be obtained by profiling entire genomes. However, the information in such profiles could be a combination of ubiquitous proteins as well as those of specialised pathways. Given the lack of mechanisms to provide weighting schemes for different sets of proteins and hence leading to difficulties in identifying the signal from noise, analysing pathway profiles confer a specific advantage for comparing organisms. A pathway also has an advantage, compared to individual proteins because the actual constraint during evolution is a functional biochemical process, which is generally achieved by the pathway as a whole. Hence, a properly identified pathway captures the required amount of detail that will help us discriminate between organisms with different functional capabilities.

4 Conclusions

This paper presents an extensive phylogenetic analysis of the MAP in *M. tuberculosis*. Simultaneous phylogenetic profiling of the entire pathway across organisms has led to the identification of proteins that are heavily linked to mycolic acid synthesis and hence classified as hallmarks of the pathway.

Knowledge of these factors help not only in understanding what renders mycolic acid synthesis capabilities in a cell, but also helps enormously in a variety of applications. For instance, desaturases could be specific drug targets for *M. tuberculosis*. Further, desaturases, given their uniqueness to mycobacteria could also be used for the identification/definite diagnosis of mycobacteria. Phylogenetic profiling and clustering can help in deriving molecular taxonomic relationships and provide an insight into evolution of the species. The study has also led to the derivation of the minimal components required to synthesise mycolates. Knowledge of the hallmarks of the pathway will also be highly useful (a) in rendering the drug target validation process more comprehensive, (b) understanding the evolutionary features leading to the genesis of the pathway and perhaps even for (c) the design of a laboratory system for obtaining mycolic acids, from a synthetic biology perspective.

5 Methods

5.1 Genome information and Sequence analysis

318 complete genomes, listed at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi (list of organisms given in Table S3) were downloaded from the NCBI website. Starting from our recent landscape for mycolic acid synthesis in *M. tuberculosis* (see Figure S1), sequence comparisons from the available genome sequences were carried out to obtain the required information, for other organisms, as appropriate. Figure 5 indicates a flowchart illustrating the methodology of the study.

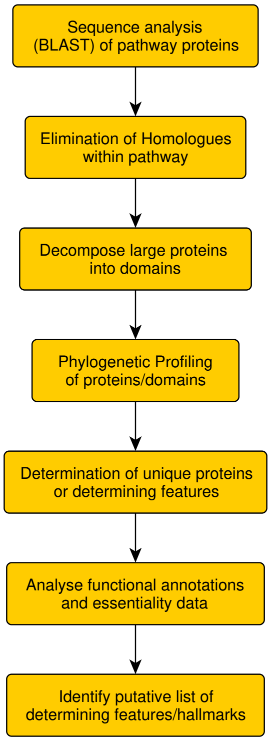
Sequence analysis
(BLAST) of pathway proteins
Elimination of Homologues
within pathway
Decompose large proteins
into domains
Phylogenetic Profiling
of proteins/domains
Determination of unique proteins
or determining features
Analyse functional annotations
and essentiality data
Identify putative list of
determining features/hallmarks

Figure 5: **Progression of various experiments in this study.**

Of the 28 proteins identified to constitute the pathway, only 26 of them could be annotated from the genome. The other two are known to be involved in the pathway through biochemical studies for catalysing the dehydration of hydroxy-acyl-ACP [1] and the formation of the *cis* and *trans* forms of keto-meroacyl-ACP [3], respectively. However, they have not been linked to any particular sequence in the mycobacterial proteome and hence remain as unidentified proteins in terms of their sequences. Therefore, only 26 proteins of this pathway could be used for all subsequent analysis. Further, few of these 26 proteins have homologues within the pathway itself. From each set of homologues, the protein with least length was taken for profiling and subsequent clustering (Table II). Two of the large proteins in the pathway, Fas and Pks13, are made up of several domains. The domains in Fas and Pks13 were taken from the NCBI Sequence Viewer entry for the corresponding protein. Regions are indicated in the sequence, along with a cross reference to the Conserved Domains Database [28]. Table I lists the domains of Pks13 and Fas. It can be seen that some of the domains listed have similar descriptions/CDD IDs. However, no significant similarity was detected between any pair of Fas_1, Pks13_3, Fas_5. We have therefore included all three of them in the study. However, Pks13_1 and Pks13_4 share sufficient similarity and hence, only Pks13_1 (the smaller of the two) has been considered for profiling. Thus, it has been ensured that there is no redundancy in the set of proteins/domains chosen for profiling.

The standalone version of the BLAST program (version 2.2.13) was used for sequence comparisons and the results parsed to obtain the required parameters using home-grown scripts. Stringent criteria were used in detecting homologues. Hits exhibiting greater than 50% similarity with BLOSUM62 matrix at E-values less than 0.1, for lengths greater than 70% of that of the query protein were considered as homologues of that query protein.

Functional annotations were obtained from databases such as the SwissPROT (http://www.expasy.ch/), NCBI Protein database, QuickGO and InterPro at EBI. A literature study to confirm the role of some of the proteins was also carried out. Functional annotations for the domains were obtained from the Conserved Domains Database/PFam [29] databases.

5.2 Phylogenetic Analysis

For phylogenetic analysis, Python scripts, based on BioPython (http://www.biopython.org/) were written to parse the BLAST outputs to obtain the E-value of the best hit for each sequence. The E-values thus obtained were converted to scores between 0 and 1, with 0 representing a strong match and 1 representing a weak match. The score was calculated is $-1/\log(E)$. Hits with $E > e^{-4}$ were all neglected and given a score of 1.0. This is identical to the scoring scheme of the on-line service, Protein Link EXplorer (PLEX) [30], which however currently considers only 89 genomes. For all proteins in the mycolic acid pathway, BLAST searches conducted against each of the 318 genomes were used to generate profile strings comprising scores for the hits of MAP proteins. Each profile string thus, encodes the presence or absence of each of the 16 proteins and where present, the extent of similarity to the corresponding one in *M. tuberculosis* H37Rv. A pathway profile matrix was correspondingly constructed, based on these profiles, with each organism corresponding to a row and each protein, to a column. The profile matrix thus obtained was of size 318 ×16. For the domain analysis, the profile matrix was of size 318 ×26. The profile matrix is given in Supplementary File S4.

5.3 Clustering

Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm) was used for clustering the profile data. A hierarchical clustering was done, using the Average Link algorithm provided by Cluster. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method was used for computing a hierarchical cluster tree, from the profile matrix. The clustering of both organisms and the proteins (including domains) was done in this fashion.

Pairwise Euclidean distances were also computed for the profiles and the organisms were clustered based on the pairwise distances between them, using

MEGA [9]. The cluster figures were generated using the open-source software jTreeView (http://jtreeview.sourceforge.net/).

6 Acknowledgments

Financial support from the computational genomics initiative of the Department of Biotechnology is gratefully acknowledged. Use of facilities at the Interactive Graphics Based Molecular Modeling Facility and Distributed Information Centre (both supported by the Department of Biotechnology) and the facilities at the Supercomputer Education and Research Centre are also gratefully acknowledged.

References

[1] Barry III CE, Lee RE, Mdluli K, Simpson AE, Schroeder BG, Slayden RA, Yuan Y. Mycolic acids: Structure, biosynthesis and physiological functions. Prog Lipid Res 1998;37:143–179.

[2] Draper P, Daffé M. Tuberculosis and the tubercle bacillus. American Society of Microbiology Press. 2005; 261–273.

[3] Takayama K, Wang C, Besra GS. Pathway to synthesis and processing of mycolic acids in *Mycobacterium tuberculosis*. Clin Microbiol Rev 2005; 18:81–101.

[4] Watanabe M., Aoyagi Y., Ridell M., Minnikin DE.. Separation and characterization of individual mycolic acids in representative mycobacteria. Microbiology 2001;147:1825–1837.

[5] Lei B, Wei CJ, Tu SC. Action mechanism of antitubercular isoniazid. Activation by *Mycobacterium tuberculosis* KatG, isolation, and characterization of *inha* inhibitor. J Biol Chem 2000;275:2520–2526.

[6] Banerjee A, Dubnau E, Quemard A, Balasubramanian V, Um KS, Wilson T, Collins D, de Lisle G, Jacobs Jr. WR. *inhA*, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*. Science 1994; 263:227–230.

[7] Raman K, Rajagopalan P, Chandra N. Flux balance analysis of mycolic acid pathway: Targets for anti-tubercular drugs. PLoS Comput Biol 2005; 1:e46.

[8] Schweizer E, Hofmann J. Microbial Type I Fatty Acid Synthases (FAS): Major players in a network of cellular FAS systems. Microbiol and Mol Biol Rev 2004;68:501–517.

[9] Kumar S., Tamura K., Nei M.. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings in Bioinformatics 2004;5:150–163.

[10] Ohlrogge J, Browse J. Lipid biosynthesis. Plant Cell 1995;7:957–970.

[11] Cole S T, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S V, Eiglmeier K, Gas S, Barry C E, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail M A, Rajandream M A, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston J E, Taylor K, Whitehead S, Barrell B G. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998;393:537–544.

[12] Dyer D H, Lyle K S, Rayment I, Fox B G. X-ray structure of putative acyl-ACP desaturase DesA2 from *Mycobacterium tuberculosis* H37Rv. Protein Sci 2005;14:1508–1517.

[13] Domergue F, Abbadi A, Zähringer U, Moreau H, Heinz E. *In vivo* characterization of the first acyl-CoA Delta6-desaturase from a member of the plant kingdom, the microalga *Ostreococcus tauri*. Biochem J 2005;389:483–490.

[14] Maier T, Jenni S, Ban N. Architecture of mammalian fatty acid synthase at 4.5 *Å* resolution. Science 2006;311:1258–1262.

[15] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 1999;96:4285–4288.

[16] Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285:751–753.

[17] Huynen M A, Bork P. Measuring genome evolution. Proc Natl Acad Sci U S A 1998;95:5849–5856.

[18] Phetsuksiri B, Jackson M, Scherman H, McNeil M, Besra G S, Baulard A R, Slayden R A, DeBarber A E, Barry C E, Baird M S, Crick D C, Brennan P J. Unique mechanism of action of the thiourea drug isoxyl on *Mycobacterium tuberculosis*. J Biol Chem 2003;278:53123–53130.

[19] Chang Y, Fox B G. Identification of Rv3230c as the NADPH oxidoreductase of a two-protein DesA3 acyl-coa desaturase in *Mycobacterium tuberculosis* H37Rv. Biochemistry 2006;45:13476–13486.

[20] Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Molecular Microbiology 2003;48:77–84.

[21] Ishikawa J, Yamashita A, Mikami Y, Hoshino Y, Kurita H, Hotta K, Shiba T, Hattori M. The complete genomic sequence of *Nocardia farcinica* IFM 10152. Proc Natl Acad Sci U S A 2004;101:14925–14930.

[22] Bentley SD., Chater KF., Cerdeño-Tárraga AM., Challis GL., Thomson NR., James KD., Harris DE., Quail MA., Kieser H., Harper D., Bateman A., Brown S., Chandra G., Chen CW., Collins M., Cronin A., Fraser A., Goble A., Hidalgo J., Hornsby T., Howarth S., Huang CH., Kieser T., Larke L., Murphy L., Oliver K., O'Neil S., Rabbinowitsch E., Rajandream MA., Rutherford K., Rutter S., Seeger K., Saunders D., Sharp S., Squares R., Squares S., Taylor K., Warren T., Wietzorrek A., Woodward J., Barrell BG., Parkhill J., Hopwood DA.. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 2002;417:141–147.

[23] Hopwood D A. The *Streptomyces* genome — be prepared! Nat Biotechnol 2003;21:505–506.

[24] Glickman MS., Cox JS., Jacobs WR.. A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. Mol Cell 2000;5:717–727.

[25] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.

[26] Cole ST. Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. Microbiology 2002;148:2919–2928.

[27] Beverley S M. Hijacking the cell: parasites in the driver's seat. Cell 1996; 87:787–789.

[28] Marchler-Bauer A, Anderson J B, Cherukuri P F, DeWeese-Scott C, Geer L Y, Gwadz M, He S, Hurwitz D I, Jackson J D, Ke Z, Lanczycki C J, Liebert C A, Liu C, F Lu, Marchler G H, Mullokandov M, Shoemaker B A, Simonyan V, Song J S, Thiessen P A, Yamashita R A, Yin J J, Zhang D, Bryant S H. CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 2005;33:D192–D196.

[29] Bateman A, Coin L, Durbin R, Finn R D, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E L, Studholme D J, Yeats C, Eddy S R. The Pfam protein families database. Nucleic Acids Res 2004; 32:138–141.

[30] Date SV, Marcotte EM. Protein function prediction using the protein link explorer (PLEX). Bioinformatics 2005;21:2558–2559.

Tables

Table I – Domains of Pks13 and Fas

Pks13 Domains			
Location	**Name**	**Description**	**CDD ID**
25..91	Pks13_1	Phosphopantetheine attachment site	CDD:40636
118..536	Pks13_2	Polyketide synthases (PKSs)	CDD:29420
713..1036	Pks13_3	Acyl transferase domain	CDD:40779
1236..1304	Pks13_4	Phosphopantetheine attachment site	CDD:40636
1470..1561	Pks13_5	Thioesterase domain	CDD:41048
Fas Domains			
Location	**Name**	**Description**	**CDD ID**
86..254	Fas_1	Acyl transferase domain	CDD:40779
384..1090	Fas_2	Enoyl reductase domain of yeast-type FAS1	CDD:34586
510..662	Fas_3	2-nitropropane dioxygenase	CDD:42994
1199..1321	Fas_4	MaoC dehydratase like domain	CDD:48042
1333..1530	Fas_5	Acyl transferase domain	CDD:40779
1788..2585	Fas_6	3-oxoacyl-[acyl-carrier protein]	CDD:34587
2116..2373	Fas_7	Dehydrogenases with different specificities	CDD:31231
2553..2988	Fas_8	Elongating condensing enzymes	CDD:29415

Table II – Homologues of proteins within the MAP

Rv Identifier	Protein	Homologues	Length
Rv0533c	FabH	—	335
Rv0645c	MmaA1	CmaA2, MmaA2, MmaA3, MmaA4, PcaA	286
Rv0904c	AccD3	—	495
Rv1094	DesA2	DesA1	275
Rv1483	FabG1	FabG2, FabG4	247
Rv1484	InhA	—	269
Rv2243	FabD	—	302
Rv2245	KasA	KasB	416
Rv2523c	AcpS	—	130
Rv2524c	Fas	—	3069
Rv3229c	DesA3	—	427
Rv3279c	BirA	—	266
Rv3285	AccA3	—	600
Rv3799c	AccD4	AccD5	518
Rv3800c	Pks13	—	1733
Rv3801c	FadD32	—	637

Supporting Information

Figure S1 — Mycolic Acid Pathway in *M. tuberculosis*

Schematic diagram of the Mycolic acid pathway (MAP) in *M. tuberculosis*. This is a revised version of Figure 1 from Raman K, Rajagopalan P, Chandra N (2005) Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-Tubercular Drugs. *PLoS Comput Biol* 1(5): e46

Table S2 — Known functional roles of MAP proteins

The known functional roles of the MAP proteins in different organisms are indicated in this table.

Table S3 — List of 318 genomes used in this analysis

The list of 318 genomes used for the analysis is indicated here.

Supplementary File S4 — Data used for clustering

This excel file provides the names of the organisms and the proteins, as well as the score matrix used for clustering.