

# Analysis of the proteome of *Mycobacterium tuberculosis* in silico

F. Tekaia\*, S. V. Gordon†, T. Garnier†, R. Brosch†, B. G. Barrell‡, S. T. Cole†

\*Unité de Génétique Moléculaire des Levures, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France

†Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

‡Sanger Centre, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

**Summary** Novel bioinformatics routines have been used to provide a more detailed definition of the proteome of *Mycobacterium tuberculosis* H37Rv. Over half of the current proteins result from gene duplication or domain shuffling events while one-sixth show no similarity to polypeptides described in other organisms. Prominent among the genes that appear to have been duplicated on numerous occasions are those involved in fatty acid metabolism, regulation of gene expression, and the unusually glycine-rich PE and PPE proteins. Protein similarity analysis, coupled with inspection of the genetic neighbourhood, was used to explore possible functional relatedness. This uncovered four large *mce* operons whose proteins may mediate initial interactions between the tubercle bacillus and host cells, together with a cluster of genes that might encode components of a structure required for secretion of ESAT-6 like proteins. Close linkage of the *mmpL* genes, encoding large membrane proteins, with those required for fatty acid metabolism suggests involvement in lipid transport. Compared to free-living bacteria, *M. tuberculosis* has a significantly smaller transport protein repertoire and this may reflect its intracellular lifestyle. © 1999 Harcourt Publishers Ltd

## INTRODUCTION

Proteins are the workhorses of the cell since most biological reactions are catalysed by enzymes. These assure the vast majority of the biosynthetic reactions as well as the catabolic ones that generate energy. Gene expression and responses to environmental stimuli are mediated mainly by proteins as are the faithful replication of the chromosome and its segregation into daughter cells. Proteins also play major structural roles and are key parts of subcellular components such as membranes, division septa, and organelles ranging from ribosomes to secretory complexes.

Traditionally, knowledge and understanding of proteins has been generated by biochemists then refined with the help of structural biologists. Nowadays, the widespread application of genomics is providing us with complete and unbiased information about the nature of proteins present in many different lifeforms and it is becoming

increasingly apparent that many new protein families exist.<sup>1</sup> Intriguingly, the number of proteins that bear no sequence resemblance to known proteins continues to grow as more genome sequences become available. This raises the possibility that these 'orphan' proteins could confer highly specific biological functions on their host organism. There is, therefore, a growing need to provide better documentation and even role prediction for such polypeptides so that experiments can be designed to test their functions.

With the completion of the genome sequence of the H37Rv strain of *Mycobacterium tuberculosis*<sup>2</sup> a vast body of information about the bacterium's protein content, or proteome, became available. This provided us with new insight into the biochemical and physiological processes that govern the life of the tubercle bacillus, highlighting the importance of lipid metabolism as well as indicating the existence of two novel protein families, PE and PPE, with unusual amino acid sequences.<sup>2,3</sup> The aim of this study was to use bioinformatics to further characterize the proteome, to assess the level of gene duplication undergone by *M. tuberculosis*, then to probe the chromosomal context of multigene families in an attempt to detect possible functional and regulatory themes.

Correspondence to: S. T. Cole, Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 Rue du Docteur Roux, 75724 Paris Cedex 15, France.  
Tel.: 3-1-45 68 84 46; Fax: 33-1-40 61 35 83; E-mail: stcole@pasteur.fr

Received: 28 January 1999; Revised: 22 April 1999; Accepted: 30 April 1999

## MATERIALS AND METHODS

### Intraproteome comparisons

Comparisons were restricted in our analyses to amino acid sequences of predicted proteins present in our inhouse *Mycobacterium tuberculosis* database. Intraproteome comparisons were performed using BLASTP version 1.4.8.<sup>4</sup> Each predicted ORF product served as a query sequence against the entire database of *M. tuberculosis*. The BLASTP comparisons were performed using *pam* 250 as substitution matrix in order to favour recognition of distantly related proteins and with the *seg* filter,<sup>5</sup> to mask compositionally biased regions in the query sequences.

Thresholds for similarity significance were determined by simulation.<sup>6</sup> We determined the significance of BLASTP probability scores using 4000 random sequences (i.e. one proteome equivalent), which were generated with sizes and compositions equal to the average size and amino acid composition of the entire predicted proteome. Each random sequence was compared against the entire database, and the best probability scores recorded. This revealed that over 95% of the random sequences showed no statistically significant relatedness to *M. tuberculosis* proteins when a BLASTP score of  $P < 10^{-5}$  was obtained. Consequently, this value was used as a cut-off to avoid losing meaningful information despite the risk of encountering some residual 'noise'. Only sequences satisfying this criterion were retained for partition analysis.

### Definition of a partition

A set of ORF products in a given organism defines a 'partition' if, and only if, the following three properties are satisfied: a) each member of the set has at least one highly significant match with another member of the set; b) no member of the set has highly significant matches with members not included in the set; and c) the set cannot be partitioned into subsets verifying a) and b) (i.e. the set is minimal). Thus, a partition includes all ORF products with common ancestry in a given organism. Note that an ORF product that has no significant match in its own proteome fulfills these properties and is therefore considered as a single member partition.

### Correspondence analysis

Correspondence analysis<sup>7,8</sup> is a multivariate method applicable to the analysis of large positive numerical datasets. Lines of such tables are the 'observations' and columns the 'variables'. In genome investigation it is of interest to identify genes or proteins that are characterized by some preferred codons or amino-acids. These investigations generally involve large and complex datasets

and correspondence analysis is an appropriate method for handling such data as a whole. It constructs an orthogonal system of axes (called factors and denoted F1, F2, etc...) where observations and variables can be simultaneously displayed. The factors are constructed according to the information they represent and are presented, therefore, in decreasing order of importance. A maximum of  $n-1$  such factors can be determined, where  $n$  is the lowest of the two numbers of observations and variables. The information included in the factorial subspace constructed from the first  $p$  factors equals the sum of information they include. The average proportion of the total information represented by one factor is  $100/(n-1)$ . This value serves as a guide in determining the relative importance of a given factor. In this system, proximity between observations and/or variables is indicative of strong similarity or relationship. The ability to display simultaneously observations and variables in the same factorial space makes it easy to discover the salient information in a given data table. This method was used to analyse the codon composition of each gene and the amino-acid composition of the *M. tuberculosis* proteome.

### Identifying partitions with common motifs using MEME and MAST

The set of ORF products having at least one highly significant match was partitioned (see definition). Each partition was successively analysed using the MEME<sup>9</sup> and MAST programs.<sup>10</sup> The Multiple Expectation-maximization for Motif Elicitation (MEME) identifies non-gapped motifs of highly informative sequence, present in some, or all, members of the partition. These are then analysed and displayed by the Motif Alignment and Search Tool (MAST).

A motif is a sequence pattern that occurs repeatedly in a group of related protein sequences. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. The 'tcm' (two-component mixture) and 'nmotifs=15' (maximum number of motifs) options were generally used. Tcm is useful for detecting motif repeats and assumes that each sequence contains any number of non-overlapping occurrences of each motif. In order to reduce computing time, the 'zoops' (zero or one occurrence per sequence) option was used instead of tcm for very large partitions.

For each motif MEME computes a position-dependent scoring matrix for use by MAST, a database searching program for sequences that contain one or more of a group of known motifs. Among the most informative MAST outputs are the motif diagrams showing the order and spacing of the motifs within each matching sequence. To facilitate the classification, particularly in large partitions,

of the proteins according to their common motifs, the distribution of each motif in each protein was scored by indicating its presence (denoted 1) or absence (denoted 0), and total in the proteins of the partition. The table of motif distribution was inserted before the motif diagrams in the MAST output. In the case of the largest partitions, attempts were made to divide them into smaller units for operational reasons.

### Other bioinformatic techniques and phylogenetic analysis

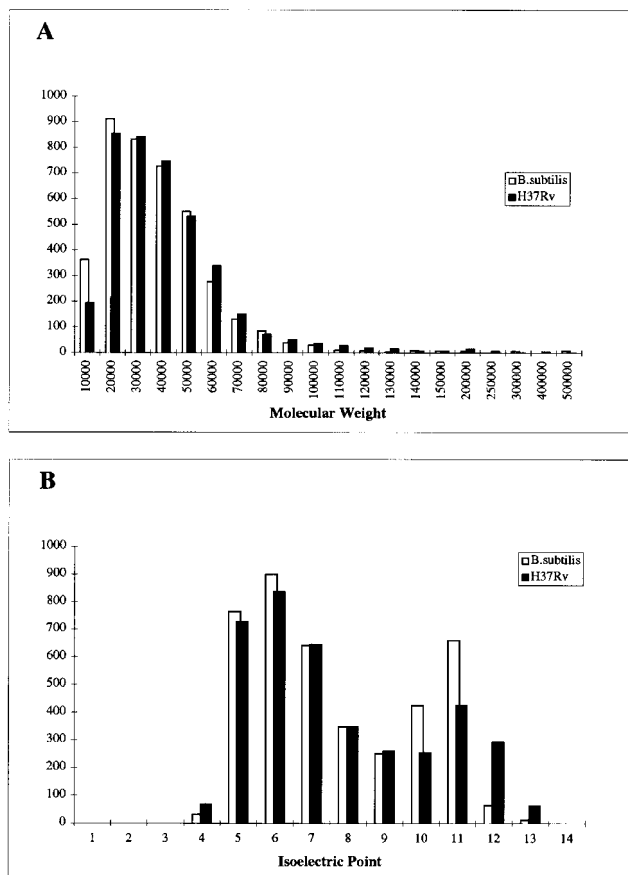
Molecular weights and isoelectric points were predicted using DIANA and the ISOELECTRIC routine in the GCG package.<sup>11</sup> Protein similarities were generated using FASTA<sup>12</sup> and proteins grouped into families on the basis of their percentage identities. These families were essentially the same as those assembled by BLASTP and partition analysis, described above. To identify potential signal peptides and transmembrane domains sequences were analysed using SignalP<sup>13</sup> (<http://www.cbs.dtu.dk/services/SignalP/index.html>) and TMHMM<sup>14</sup> (<http://www.cbs.dtu.dk/services/TMHMM-1.0/>), respectively. Dedicated relational databases for *M. tuberculosis* (<http://bioweb.pasteur.fr/GenoList/TubercuList/>) and *Bacillus subtilis* (<http://bioweb.pasteur.fr/GenoList/SubtiList/>) were consulted interactively via Netscape.

Alignments for phylogenetic trees with sequences from the partition P24.1 and Rv0176, considered as outgroup, were constructed using ClustalW<sup>15</sup> and the trees drawn by PHYLIP.<sup>16</sup> All gaps introduced by Rv0176 were removed. The statistical confidence of each node of the tree was estimated by bootstrapping, which involved the generation of 100 multiple random subsets of the alignment using the SEQBOOT program. DNAdist was used to calculate the Kimura 2-parameter distance matrices corresponding to the 100 random multiple alignments. The resultant distance matrices were used to draw neighbour-joining trees, using the NEIGHBOUR and CONSENSE programs from the PHYLIP package<sup>16</sup> to draw the final tree.

## RESULTS

### General properties of the proteome

During analysis of the genome sequence, over 3920 genes were uncovered that encode proteins larger than 80 amino acids in length. It is conceivable that some smaller genes may have been missed as these are difficult to detect using current methods. The complete proteins ranged in size from the 38 amino acid long, 50S ribosomal protein RpmJ, to the putative polyketide synthase, PKS12, comprising 4151 residues. The mean length was 339



**Fig. 1** The proteins comprising the proteomes of *M. tuberculosis* H37Rv (right columns) and *B. subtilis* (left columns) were classed according to their predicted molecular weights. (B) Comparative distribution of proteins in the proteomes of *M. tuberculosis* H37Rv and *B. subtilis* according to their predicted isoelectric points calculated using the ISOELECTRIC program in the GCG package.

( $\pm 284$ ) amino acid residues, fairly typical of other prokaryotes such as *Bacillus subtilis* (Fig. 1A).

To gain insight into the possible subcellular location of the proteins their pI was predicted and the distribution is shown in Figure 1B. Two peaks were apparent at pI 6 and 11 with over 60% of the proteins displaying an acidic isoelectric point ( $pI < 7$ ). The distribution of the proteins throughout the pI range was essentially similar to that seen with *B. subtilis* except for the existence in the tubercle bacillus of about 4-fold more proteins with extremely basic  $pI^{12-13}$ . Many of these correspond to ribosomal, DNA-binding, cell envelope or transmembrane (TM) proteins. Proteins with acidic pI were for the most part cytoplasmic enzymes involved in various metabolic functions. Also prominent among the acidic group were the bulk of the PE and PPE proteins (see below) suggesting that these might be localized in the cytoplasm.

The amino acid content of the proteome was calculated and compared with those of other bacteria for which

**Table 1** Comparison of amino acid composition (%) of the proteomes of *M. tuberculosis* and *B. subtilis*

Amino acid	<i>B. subtilis</i>	<i>M. tuberculosis</i>
A	7.7	13.2
C	0.8	0.9
G	6.9	10.0
P	3.7	5.8
S	6.3	5.5
T	5.4	5.9
H	2.3	2.2
K	7.1	2.0
R	4.1	7.3
D	5.2	5.8
E	7.2	4.7
N	3.9	2.5
Q	3.8	3.1
I	7.4	4.3
L	9.6	9.8
M	2.8	1.8
V	6.7	8.6
F	4.5	3.0
W	1.0	1.5
Y	3.5	2.1

complete genome sequence data were available (<http://www-alt.pasteur.fr/~tekaia/aafreq.html>), but for discussion purposes we will restrict ourselves to *B. subtilis* (Table 1). The most common amino acids in *M. tuberculosis* are alanine and glycine, and the difference in abundance with respect to *B. subtilis* (Table 1), and other bacteria (data shown at <http://www-alt.pasteur.fr/~tekaia/aafreq.html>), is statistically significant. It is also interesting to note the rarity of lysine which is offset by an increased content of arginine in the proteome. Glutamic acid, asparagine, isoleucine, phenylalanine and tyrosine are all encoded by A/T-rich codons and are used significantly less often in *M. tuberculosis*, while alanine, arginine, tryptophan and proline are much more frequent.<sup>2</sup> It is conceivable that the biased amino acid composition, which is a reflection of the skewed codon usage imposed by the high G/C content of the genome, may have immunological consequences as host proteins will be of markedly different composition.

### Correspondence analysis of the proteome

The most powerful means of scrutinizing large datasets such as the amino acid composition of a bacterial proteome is by correspondence analysis<sup>7</sup> since plotting information about proteins and their amino acids in factorial space allows proteins of atypical composition to be readily identified.<sup>8</sup> When the constituents of the *M. tuberculosis* proteome were evaluated in this way, it was immediately apparent that three distinct groups of proteins could be defined as a function of their amino acid composition (Fig. 2). The bulk of the proteins (situated near

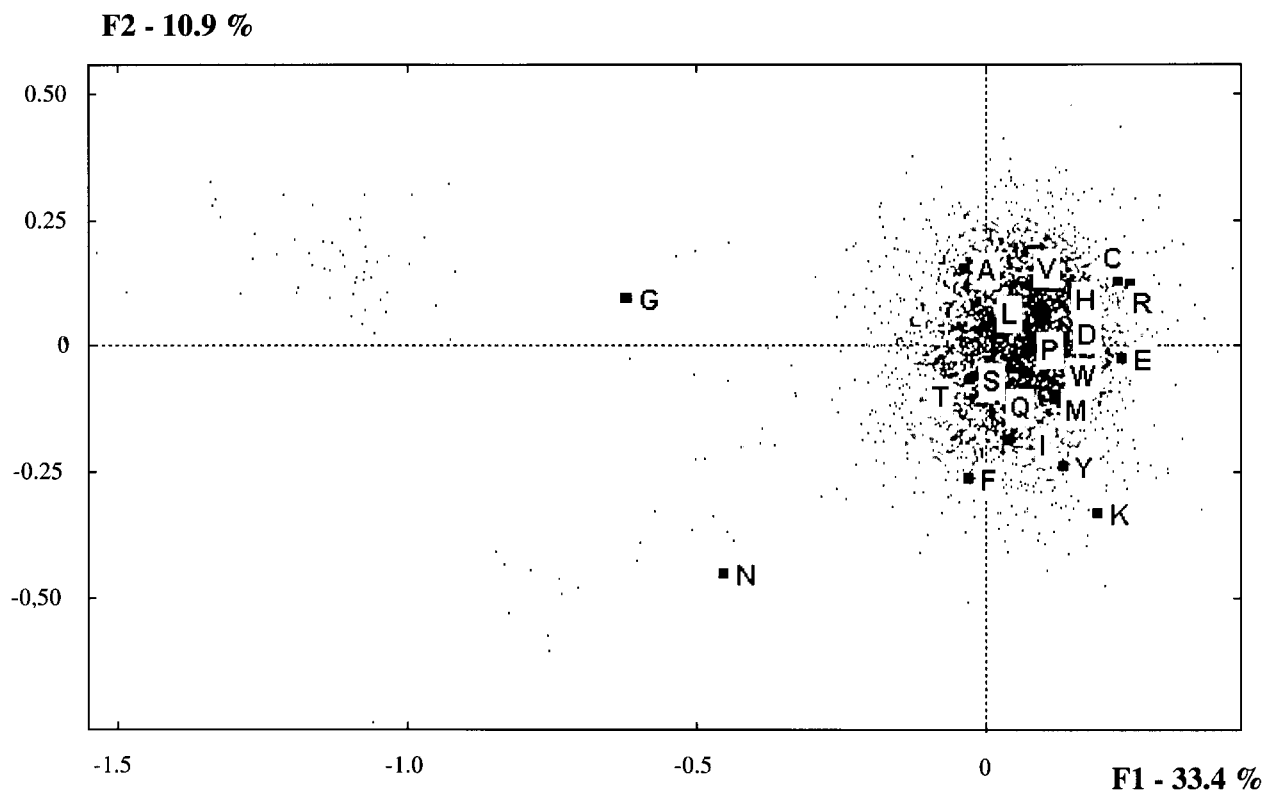
the origin of the axes) have normal composition and are strongly associated with 18 amino acids whereas many outlying proteins show a strong preference for glycine and asparagine. These appear as spurs from the central cluster (Fig. 2) and, on examination, were found to correspond to the PE and PPE protein families that have been described previously in some detail.<sup>2,3</sup>

Briefly, the 99 proteins comprising the PE family all have an N-terminal segment of ~110 amino acids with the motif Pro-Glu (PE) occurring frequently at positions 8–9. In the majority of cases, this is followed by a domain that is variable in size but very rich in glycine and in some proteins this small amino acid accounts for over 50% of the residues. Revised analysis of the PE family members and motif definition can be found at (<http://bioweb.pasteur.fr/GenoList/TubercuList/PE.html>) and it is clear that the most commonly occurring motif is GlyAsnGlyGly AlaGlyGly. On backtranslation, part of this motif gives rise to the DNA sequence CCGCGCAA that is characteristic of the polymorphic GC-rich sequences, PGRS,<sup>17</sup> and, as its name implies this sequence is associated with genomic polymorphisms.<sup>18,19</sup> About 60 members of the PE family contain PGRS motifs and some of these have been shown to undergo sequence variation leading to the suggestion that they might correspond to variable protein antigens.<sup>2,3</sup> Twenty-one PE proteins contain the core sequence only and this may indicate that their genes have undergone retraction or are awaiting expansion.

The PPE proteins are less numerous than the PE proteins but present many similarities notably, the existence of a conserved N-terminal segment of ~180 residues followed by COOH-terminal domains of variable length and sequence (<http://bioweb.pasteur.fr/GenoList/TubercuList/PPE.html>). Although the PPE proteins can be classified into three broad groups the subclass whose genes contain the major polymorphic tandem repeat sequence, MPTR, is particularly striking.<sup>2,3</sup> These polypeptides can contain over 3700 residues and are rich in asparagine and glycine which occur frequently in the MPTR-encoded signature AsnThrGlyPheGlyAsnMetGly (P208.1). The abundance of asparagine in the PPE proteins is intriguing given its relative paucity in the proteome as a whole. This suggests that asparagine storage could be a possible function for the PPE proteins as this amino acid is one of the preferred nitrogen sources of the tubercle bacilli.<sup>20</sup>

### Gene duplication and protein families

Intra-proteome comparisons were performed to determine what percentage of the proteins present in *M. tuberculosis* result from gene duplication events or domain shuffling.<sup>6</sup> The sequences of 1939 proteins were found to be unique whereas 2003 showed highly significant levels of relatedness. Thus 52% of the proteome can be estimated as being



**Fig. 2** Correspondence analysis of the proteins present in the *M. tuberculosis* proteome expressed as a function of amino acid composition. The first and second factorial axes, F1 and F2, correspond to 30.4% and 4.9, respectively of the total information. The origin of these axes represents the average amino acid content for typical proteins and every protein is represented by a dot; since the bulk of the proteome consists of proteins with unbiased amino acid content these proteins cluster strongly around the origin. Note the existence of two groups of proteins with biased contents of Gly (G) and Asn (N) corresponding to the PE and PPE family members, respectively.

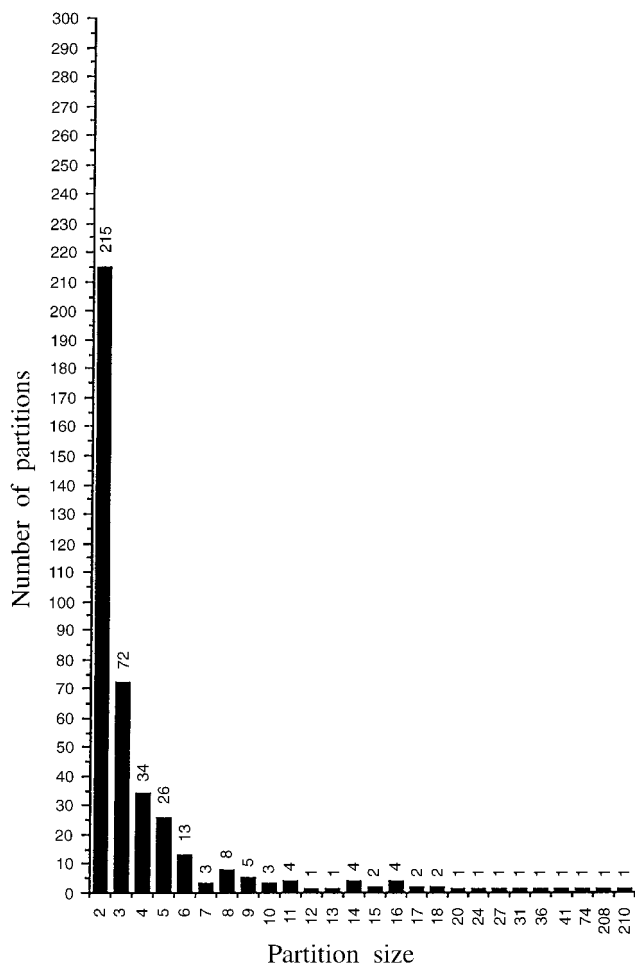
derived from gene duplication in one form or another and, using the criteria defined above, fell into 408 partitions containing from two to 210 members (Fig. 3). These partitions were then analysed further using the programs MEME<sup>9</sup> and MAST<sup>10</sup> to identify statistically significant conserved motifs (see Materials and Methods) and to refine the alignments. After inspection of the annotation for the corresponding CDS a complete list of putative functions was established and this can be found at [http://bioweb.pasteur.fr/GenoList/TubercuList/P\\_functions](http://bioweb.pasteur.fr/GenoList/TubercuList/P_functions).

A subset of this information is presented in Table 2 which contains information about partitions with more than 9 members. Close inspection reveals that, in addition to the PE and PPE proteins, three broad biological activities are particularly prominent among the larger protein families, namely, the regulation of gene expression, fatty acid metabolism, and transport of metabolites. The properties of another major component of the proteome, IS-encoded functions, have been described elsewhere.<sup>21</sup> At present, it is unclear if all or any of these duplicated genes are transcribed and classical approaches to studying gene expression and function are required to

establish their relative contributions to the physiology of *M. tuberculosis* under appropriate growth conditions.

### Regulatory proteins

Transcription of genes is controlled at the level of the promoter in two main ways. Initiation of mRNA synthesis requires RNA polymerase containing a sigma factor capable of recognizing its cognate promoter and the 13 sigma factors responsible for promoter recognition fall into two classes containing 10 (P10.3) and three (P3.15) members, respectively.<sup>2</sup> The smaller group contains the house-keeping sigma factors, SigA and SigB (plus SigF), while the larger group comprises the so-called extra-cytoplasmic function (ECF) sigma factors.<sup>22,23</sup> These subunits are often only produced in response to a given physiological signal. Promoter recognition can be prevented by repressor proteins and in *M. tuberculosis* the two most prominent repressor families comprise proteins resembling the TetR and ArsR repressors (Table 2). Examples of other kinds of repressors (MerR, IclR, etc.) may be found at [http://bioweb.pasteur.fr/GenoList/TubercuList/P\\_functions](http://bioweb.pasteur.fr/GenoList/TubercuList/P_functions). At least one



**Fig. 3** Partition distribution of duplicated proteins in *M. tuberculosis*. The size of the partition is shown along the horizontal axis and the number of partitions of this size indicated above the columns, e.g. there are 215 partitions consisting of two members.

member is present of each of the principal transcriptional activator protein families AraC, Crp/Fnr, Lrp, LysR, LuxR with the MoxR (P4.1) and GntR (P5.22) families being most numerous. Gene expression is often activated in response to an external stimulus and this can be transmitted via a phosphorelay mediated by either the two component systems comprising sensor histidine kinases (P11.1) and response regulators (P31.1) or, possibly, by means of the serine-threonine protein kinases (P210n.PKN).

### Fatty acid metabolism

Given the complex nature of the mycobacterial cell wall, lipid metabolism is likely to be one of the major metabolic activities. It was no great surprise, therefore, to find that the genome encodes multiple copies of many of the enzymes involved in either the synthesis of mycolic acids (P10.1), and other complex lipids or polyketides (P208.A), and this has been discussed at some length already.<sup>2,24</sup>

**Table 2** Partitions with >9 members

P10.1	Cyclopropane mycolic acid synthase 1, etc.
P10.2	Misc. semialdehyde dehydrogenases
P10.3	ECF subfamily sigma subunits
P11.1	Sensor histidine kinases
P11.2	Acetyl-CoA C-acetyltransferases, lipid carrier proteins
P11.3	Probable esterases, penicillin binding proteins
P11.4	UNK
P12.1	Cation-transporting ATPase
P13.1	UNK
P14.1	Esat-6 family
P14.2	UNK
P14.3	UNK
P14.4	Unknown dehydrogenases
P15.1	UNK
P15.2	UNK
P16.1	MmpL family
P16.2	13E12 repeat family
P16.3	Sugar transporters
P16.4	UNK
P17.1	N-terminal part of IS6110 transposase
P17.2	Transcriptional regulator (mostly ArsR family)
P18.1	Putative oxidoreductases, AAEQMGFDAMWVAEH family
P18.2	C-terminal part of IS6110 transposase
P20.1	Cytochrome P450 family
P24.1	Mce family
P27.1	Enoyl-CoA hydratases
P31.1	Transcriptional regulator (LuxR), 2-comp. response regs
P36.1	Acyl-CoA dehydrogenases
P41.1	Transcriptional regulators (mainly TetR family)
P74.1	ABC-transporters, MFS etc.
P208.1	Various proteins involved in lipid metabolism, etc.
P210.1	PPE, PE, etc.

UNK – unknown. Further details of otler partitions can be found at [http://bioweb.pasteur.fr/GenoList/TubercuList/P\\_functions](http://bioweb.pasteur.fr/GenoList/TubercuList/P_functions)

As outlined previously, *M. tuberculosis* appears to contain numerous enzymes which could interact to constitute alternative beta-oxidation cycles to the multiactivity FadA/FadB enzymes such as the 36 acyl-CoA synthases, FadD1-36 (P208.A), 36 acyl-CoA dehydrogenases, FadE1-36 (P36.1), the >21 enoyl-CoA hydratases (P27.1) and a limited number of FadB enzymes that produce the corresponding 3-keto acids. There is a very large family of short-chain alcohol dehydrogenases with over 40 members that are similar in sequence to the 3-oxoacyl-[ACP] reductases FabG (P208.B). It is unclear at present what role these dehydrogenases play but it seems likely that they will catalyse oxidoreduction steps involving alcohols and aldehydes derived from lipid degradation.

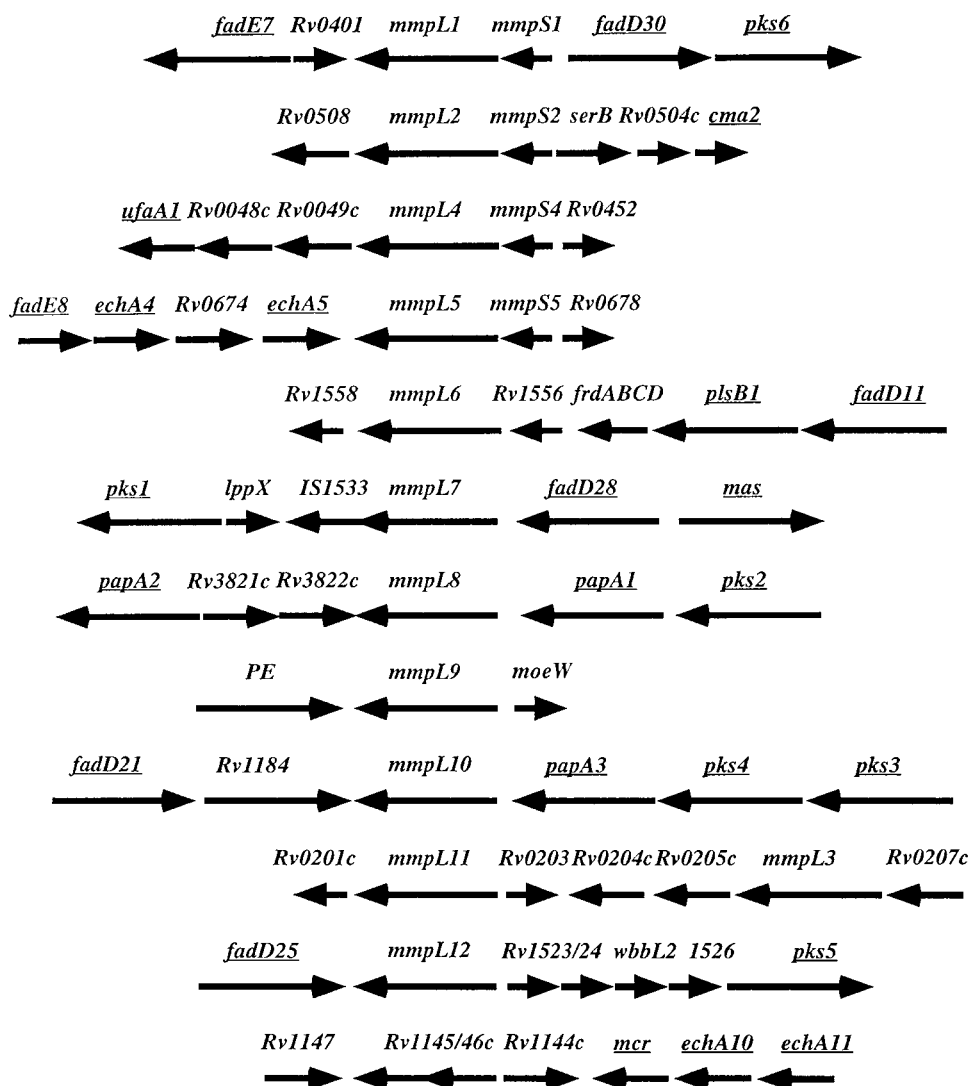
Two other large groups of proteins that might interact with fatty acids, or even sterols, were detected by partition analysis and the more homogeneous of these includes members of the cytochrome P-450 superfamily (P20.1; Table 2). In a variety of bacteria, especially Actinomycetes, these enzymes have been shown to participate in the degradation of xenobiotics, fatty acids and cholesterol.<sup>25</sup> In many fungi, the important role of cytochrome P-450 type enzymes in the biosynthesis of

sterols has made them attractive targets for novel azole fungicides. Although epoxide hydrolases are generally considered to be involved in detoxification reactions, often in conjunction with cytochrome P-450 enzymes, following oxidative damage to lipids, there is evidence from eukaryotes that the primary function of these enzymes may be the transformation of fatty acid epoxides.<sup>26</sup> The tubercle bacillus appears to have at least five epoxide hydrolases (EphA – EphF) of around 300 amino acids in length, and a sixth which is predicted to be the N-terminal part of a fusion protein, EphD, that has a short chain alcohol dehydrogenase domain at its C-terminus. These epoxide hydrolases may be involved in epoxide

catabolism and are highly related to non-haem haloperoxidases and various esterases that belong to the  $\alpha/\beta$  hydrolase fold family of enzymes which have a catalytic triad at their active site.<sup>27–29</sup> There are at least 25 members of this family in *M. tuberculosis* and they all contain the highly-conserved catalytic motifs (P208.D). Clearly much work will be required to elucidate their substrate specificities.

#### Fatty acid transport?

Linked to the genes for many of these lipid-metabolizing enzymes (Fig. 4) are the coding sequences for another



**Fig. 4** Organization of genomic loci containing *mmpL* genes. Arrows represent genes and their direction of transcription with gene names as described previously.<sup>2</sup> The figure is centred around the *mmpL* genes and the neighbouring genes are shown. Those genes known or believed to be involved in lipid metabolism are underlined. Proteins Rv1146c and Rv1145c belong to the same partition (P16.1) as the MmpL proteins and may represent distant examples encoded by two separate genes rather than one. The figure is not to scale and operon structures are not respected.

large group of proteins, the hydrophobic MmpL family, and its relatives (Table 2). The Mmp proteins appear to be confined to mycobacteria hence the designation mycobacterial membrane protein or Mmp. The MmpL proteins are related to each other at both the sequence and structural levels. They comprise ~950 amino acid residues and are predicted to contain 12 membrane-spanning alpha helices organised in three clusters. The first TM segment occurs after some 20 residues, and is separated by a ~140 residue stretch, that may be extracytoplasmic, from a large hydrophobic domain comprising a cluster of six TM alpha helices that precedes a hydrophilic segment of 300–400 amino acid residues that is likely to face the exterior. This is then followed by a group of five TM alpha helices and a putative cytoplasmic domain of about 250 residues. This arrangement is reminiscent of that seen in the proton-dependent efflux proteins belonging to the RND family (Resistance, Nodulation, Division), which are also similar in size and topology.<sup>30,31</sup> Both the RND and the MmpL proteins appear to be derived from a gene duplication/fusion event.

In four cases the *mmpL* genes are preceded in the operon by the coding sequences for the MmpS proteins, small mycobacterial membrane proteins that have a hydrophobic alpha helix close to the NH<sub>2</sub>-terminus while their COOH-terminal domains of ~120 residues are predicted to be exposed to the exterior. The hydrophobic nature of the Mmp proteins and the close association of their genes with those involved in lipid metabolism (Fig. 4) suggests that they may be involved in the transport of fatty acids. Indirect support for this proposal is provided by knowledge that certain RND proteins from *Rhizobium* are required for the export of sulfated (lipo)oligosaccharides.<sup>30</sup>

### Transport proteins

In addition to the Mmp proteins, there are several other families of known transport proteins. One of the largest partitions contains members of the ATP-binding cassette (ABC) and the major facilitator superfamilies, MFS (Fig. 3; Table 2). One subdivision (P74.1ABC), ~50 ABC-proteins were found in contrast to the 157 present in the proteome of the free-living *B. subtilis*<sup>32</sup> which has a genome slightly smaller than that of H37Rv. There are 15 membrane proteins with six TM segments that clearly comprise part of binding-protein-dependent uptake systems (P16.3) and these appear to be selective for anions especially phosphate, *sn*-glycerol-3-phosphate, sulphate, and molybdate. The genes for these proteins occur in operons that include those for probable ABC-type transporters and the cognate solute-binding protein which is usually produced as a lipoprotein precursor.

In bacteria, at least six subfamilies comprise the MFS, and these differ in their substrate specificity and the num-

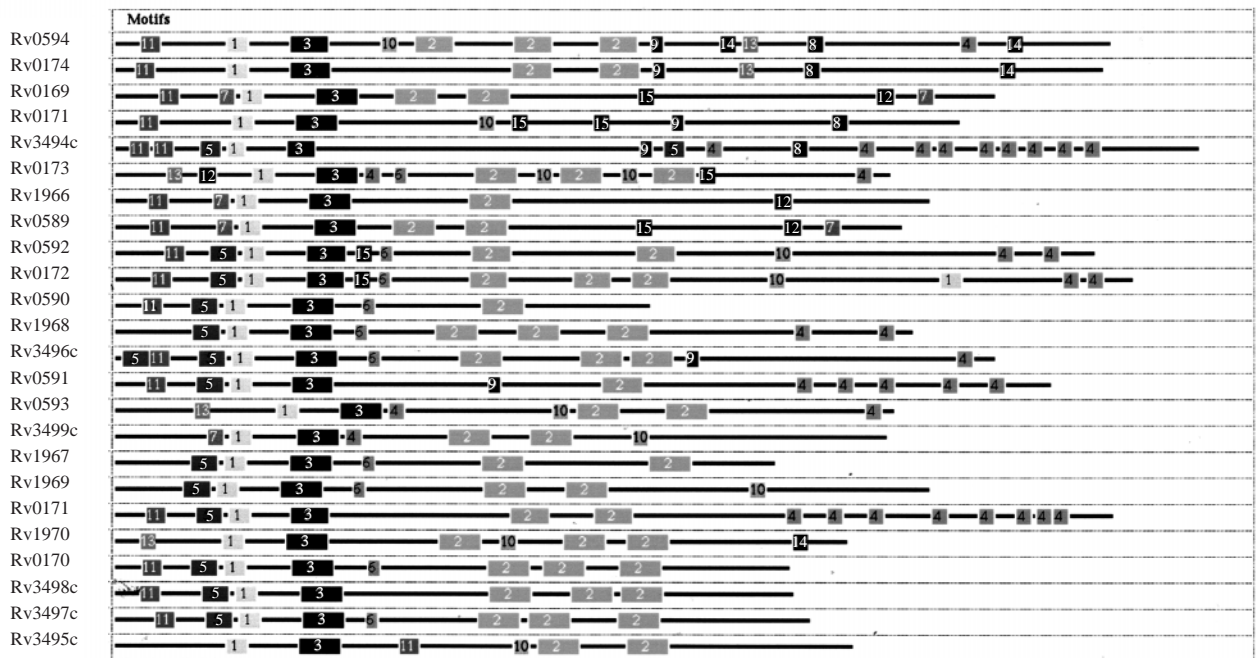
ber of TM segments, 12 or 14.<sup>33,34</sup> While *M. tuberculosis* appears to have 14 MFS proteins with 14 TM segments that could serve as PMF-dependent drug pumps (P74.1MFS), *B. subtilis* has about 40 of these proteins. It has already been demonstrated that one such protein from *M. tuberculosis* Rv1258c probably acts in drug efflux.<sup>35</sup> The members of the MFS with 12 TM segments are involved in the transport of drugs, sugars, amino acids, TCA cycle intermediates, phosphate esters and oligosaccharides. Although *M. tuberculosis* has examples of most of these classes,<sup>33,34</sup> there is only one sizeable partition (P8.2) and this contains mainly putative amino acid permeases belonging to the APC or amino acid-polyamine choline family.<sup>34</sup> Yet again, *B. subtilis* appears to possess considerably more of these transporters (~20).

In contrast, the tubercle bacillus contains 12 cation transporting P-type ATPases, with an average length of 750 amino acids (Ctp; P12. 1) (Table 2) whereas *B. subtilis* has only four.<sup>34</sup> The Ctp proteins probably mediate cation homeostasis and fall into two groups, the larger of these contains ten members with ~8 TM segments.<sup>36</sup> CtpH and CtpI comprise the smaller group and these proteins resemble each other strongly. They have an additional N-terminal domain of ~800 residues that shows no significant similarity to any known proteins. Taken together these observations suggest that *M. tuberculosis* has considerably fewer transport functions than the saprophyte *B. subtilis*, which encounters a much broader variety of substrates in the soil. Presumably, this limited transport repertoire is a reflection of the intracellular lifestyle of the tubercle bacillus and its heavy dependence on lipolysis. A more detailed analysis of the predicted transport functions will be presented elsewhere.

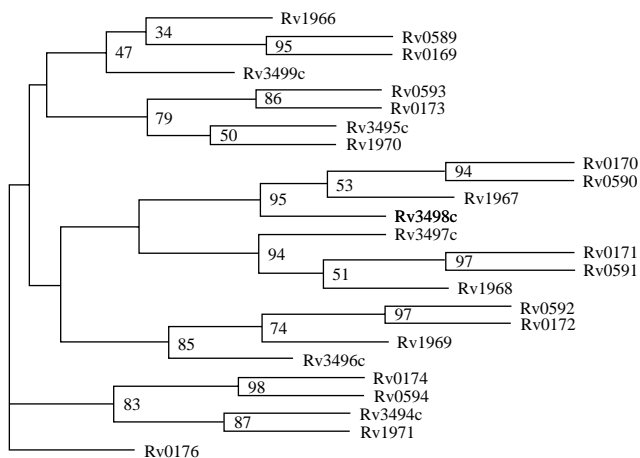
### Unidentified protein families

From Table 2 it can be seen that the proteome contains several large families about which no functional information could be gleaned. Close to 16% of the polypeptides predicted in *M. tuberculosis* bear no resemblance to known proteins,<sup>2</sup> and roughly half of these belong to families containing from two to 16 members. This is an exciting finding as these protein families undoubtedly contain clear candidates for mycobacterial specific functions. Amplification of the corresponding genes may indicate functional redundancy or, alternatively, could highlight the importance of the corresponding biological activities for the tubercle bacillus. Some of these proteins have counterparts of unknown function in *Mycobacterium leprae* or *Mycobacterium avium* whereas others are confined, at present, to *M. tuberculosis*. The latter group may contain enzymes that play specific roles in tuberculosis and thus represent attractive candidates for future research.





**Fig. 6** Motif organization of the Mce proteins. The sequences of the 24 Mce proteins (P24.1) were analysed with MEME and MAST to detect the presence of conserved motifs. The probability of the combined motifs occurring by chance was calculated and the proteins then ranked accordingly. The motifs, their identifier, length and consensus sequence were: 1, 11, VKVKGVKVGKV; 2, 22, LANKGNKLNKMLNKNLNMKMMK; 3, 22, IPANATAKIKAQTLGNKYVEL; 4, 8, PGGPPPGP; 5, 14, YYAYFTNAAGLYAG; 6, 6, IPLERT; 7, 8, AGLVMDTG; 8, 8, RGARNIPC; 9, 6, IEQLLV; 10, 8, AIDAANRL; 11, 10, VVLAVLAVVV; 12, 9, GPGGRPGC; 13, 7, LPACEWR; 14, 6, QQPNPC; 15, 6, RGGPYL. Note that the consensus sequence can differ extensively from the real sequence.



**Fig. 7** Bootstrapped phylogenetic tree of the Mce family. The sequences of the 24 Mce proteins (P24.1 partition) and an unrelated protein, Rv0276 for outgroup purposes, were aligned using CLUSTALW and gaps eliminated. The phylogenetic tree was generated by PHYLIP using the neighbour joining method with the Kimura-2 parameter distance and a bootstrap value of 100.<sup>16</sup> Note the relative distributions of the corresponding members of each operon. An essentially similar tree was obtained when motifs 1–3 from the MEME output were used instead of the complete sequences. Numbers show the percent occurrence of the nodes in 100 bootstrap replications.

operons as, with the exception of *mce2*, the transcriptional units conclude with two or four genes (Fig. 5) that encode proteins belonging to the same partition (Table 2; P13.1). It is clear that these also arose from an ancestral gene duplication event and subsequently diverged as they occur in tandem with each gene being more closely related to its counterpart in the other operons than to its neighbour. Functional information about these proteins, which contain 160–240 amino acids, is scarce although they all contain a single hydrophobic segment located about 130 residues from the COOH-terminus that implies a possible interaction with the cytoplasmic membrane.

### The ESAT6 loci

There has been considerable interest in ESAT-6 because this small protein has been shown to be present in the culture filtrates of tubercle bacilli, very early in the growth cycle, despite lacking obvious secretion signals. It is a highly potent T-cell antigen that offers great diagnostic potential<sup>38–40</sup> as its gene, *esx*, together with extensive flanking sequences, has been lost from the genome of the vaccine strain *M. bovis* BCG as part of deletion region RD1.<sup>41</sup> On analysis of the genome sequence, several proteins similar to ESAT-6 were uncovered and none of these



stretch some 40 residues from the N-terminus (P5.6). By contrast, the following gene(s) encodes a large membrane protein comprising some 1300 amino acids organised into at least two domains (Table 1; P8.1). The N-terminal part contains two clear TM segments whereas the larger COOH-terminal segment has three ATP-binding sites that are separated by ~350 and 230 residues. In two instances (Rv1783, 1784 and Rv3870, 3871), these distinct domains are encoded by two adjacent genes and the proteins total around 1300 amino acid residues.

Downstream of the *esx* module another gene coding for a TM protein can be found on five occasions (P5.24; Fig. 8). Its product comprises ~500 amino acid residues and is predicted to have 11 alpha helices that cross the lipid bilayer. This is closely followed by the gene for a putative serine protease that has a cleavable N-terminal signal sequence and a hydrophobic anchor at the COOH-terminal end. In three cases, the protease genes precede coding sequences for weakly related proteins (Rv3885c, Rv3882c, Rv1797) that all have two potential membrane spanning alpha helices near their N-termini.

## DISCUSSION

The aim of this work was to provide a more detailed description of the *M. tuberculosis* proteome than was possible previously<sup>2</sup> and to use advanced bioinformatic procedures to glean structural and functional information about some of the components whose biological roles were unknown or obscure. This procedure has led to improved understanding of the organisation of three groups of proteins MmpL/MmpS, Mce, and ESAT-6 and provided some insight into possible functions as well as a better definition of the chemical properties of the proteome.

The genomes of *M. tuberculosis* and *M. africanum* contain four copies of the Mce operon whereas those of *M. bovis*, *M. bovis* BCG, and *M. microti* only contain three.<sup>42</sup> While the initial Mce1 clone was described as rendering *E. coli* invasive for HeLa cells the molecular basis of this invasion remains obscure. It is now clear that tubercle bacilli contain either 18 or 24 genes encoding Mce proteins and this is surely a compelling argument for their playing an important biological role. The operon structure together with the bioinformatic analysis presented here indicate that the *mce* operons resulted from a succession of gene duplication events involving a pair of linked ancestral genes. These encoded two conserved hypothetical proteins, YrbE and YrbD, that probably localize to the cytoplasmic membrane in *E. coli* or *H. influenzae*. It is apparent that the Mce protein corresponds to the complete ~150 amino acid residue YrbD protein but that it has acquired an additional domain of 125–350 residues at the COOH-terminal end.

Two lines of direct evidence are available that indicate that the Mce proteins are exposed on the cell surface. Firstly, the information contained within the first 50 amino acid residues of the Mce protein Rv0590 is sufficient to promote the translocation of a  $\beta$ -lactamase derivative lacking its signal peptide across the inner membrane of *E. coli*.<sup>43</sup> Secondly, antibodies directed against Mce1 react with the cell surface of intact *M. tuberculosis* cells in immuno-electron microscopy localization studies (L. Riley, personal communication). The available data suggest, therefore, that the Mce proteins should be anchored to the cytoplasmic membrane via their N-terminus where they could associate with the TM YrbE proteins. The COOH-terminal domain is thus probably exposed to the exterior, and it is striking that in all the Mce proteins this segment contains several copies of an extended motif that is predicted to adopt an amphiphilic alpha helical structure. It is possible that if such amphiphilic helices were to associate, their hydrophobic surfaces would interact with the mycolic acid layer of the mycobacterial envelope and that the charged interior could then line a channel. Clearly, localization of the Mce proteins in the envelope is consistent with a role in adhesion to host cell surfaces, or invasion, as was originally proposed,<sup>37</sup> but a large body of experimental work is required to consolidate this suggestion and to clarify the role of the other components of the *mce* operons.

While many of the extracellular proteins of *M. tuberculosis* are exported via the signal peptide-dependent general secretory pathway,<sup>44</sup> and others like the chaperones are believed to be released after cell lysis,<sup>45</sup> the early culture filtrate also contains some small polypeptides that do not appear to contain typical secretion signals. Foremost among these is ESAT-6, a 96 amino acid residue polypeptide that is a major T-cell antigen.<sup>38,39</sup> It is shown here that ESAT-6 is part of a 14-membered family and that the corresponding genes are situated in genomic loci showing strongly conserved structure and organisation. Recently, Berthet et al.<sup>46</sup> reported similar but less extensive observations and demonstrated that *esx* is part of a transcriptional unit with another family member, termed *lhp*. Importantly, the Lhp protein (also referred to as CYT10) was found in the culture supernatant suggesting that all the ESAT-6 proteins might be secreted.<sup>46</sup>

Inspection of the chromosomal context of the *esx* genes revealed two basic arrangements. In the simpler of these, genes for PE, PPE, QILSS and ESAT-6 proteins are linked whereas in the more complex situation there are five highly conserved genes flanking this core. Remarkably, these genes are only found on the chromosome of *M. tuberculosis* adjacent to ESAT-6 coding sequences and this strongly suggests that they may constitute a functional unit. Of particular significance is the presence of the genes comprising the P8.1 family as these code for large

membrane proteins with the potential to hydrolyse ATP (Fig. 8). Although there is no sequence similarity, these resemble certain ABC transporters such as the multidrug resistance (MDR)-proteins in terms of size, the presence of separate ATP-binding sites and their being encoded by one large or two contiguous medium-sized genes. MDR-proteins, like the P-glycoprotein, have evolved by gene duplication and fusion events since they consist of two related parts comprising six membrane-spanning alpha helices and a domain that hydrolyses ATP.<sup>31</sup> The P8.1 family members are less hydrophobic and appear to have followed a different evolutionary route. Other conserved, integral membrane proteins, the P5.24 family, also figure in this setting and their genes are closely linked to those encoding putative envelope-bound serine proteases (Fig. 8) with the CbxX/CfqX gene representing another conserved component. The precise function of the CbxX/CfqX proteins is not known but they are required for the synthesis of Calvin cycle enzymes in autotrophic bacteria and plants so might act as chaperones.

Unlike their Gram negative counterparts, Gram positive bacteria do not appear to possess a specialised secretion apparatus to effect the translocation of extracellular proteins such as hydrolytic enzymes or virulence factors,<sup>47</sup> and these probably pass via the general secretory pathway.<sup>44</sup> There are, however, a few reports of lantibiotics and peptides of up to 50 amino acids long being secreted from *Lactobacillus plantarum* and related bacteria, by proteolytic ABC-transporters.<sup>48,49</sup> As outlined above, the more complex gene cluster encompassing the *esx* genes, which has apparently been duplicated on at least four occasions, has several features which suggest it might encode secretion machinery that could enable the ESAT-6 proteins to traverse the cytoplasmic membrane or even the arabinogalactan-mycolic acid layers of the envelope. This hypothesis generated *in silico* can now be tested experimentally.

The existence of a dedicated secretion apparatus would also be consistent with the observation that ESAT-6 can accumulate in the cytoplasm as intracellular stockpiling of proteins for secretion has been observed in *Shigella dysenteriae*, and other pathogens, that possess type III secretion mechanisms.<sup>50</sup> Recent studies with other extracellular or surface-exposed mycobacterial proteins that lack N-terminal signal sequences, such as the heparin-binding hemagglutinin,<sup>51,52</sup> also suggest the existence of alternative export pathways and this might explain why *M. tuberculosis* has two distinct copies of the SecA protein.<sup>2</sup> Finally, it is also striking that at least three loci encoding ESAT-6 proteins, RD1, RD5 and RD8, have been deleted from the chromosomes of tubercle bacilli following their divergence from a common ancestor.<sup>42</sup> Since ESAT-6 is a potent T-cell antigen<sup>38-40</sup> it is possible that strains harbouring deletions were selected as a conse-

quence of pressure exerted by the immune system of their respective animal and human hosts.

## ACKNOWLEDGEMENTS

We are grateful to B. Dujon, L. Riley and M.-L. Baribal for advice, unpublished information and support. Special thanks to A. Krogh for making the TMHMM software available before publication. This study was financed in part by the BIOMED program of the European Union (grants BMH4-CT96-1241; BMH4-CT97-2277), the Association Française Raoul Follereau, and the Institut Pasteur. S.V.G. received a Wellcome Trust Travelling Research Fellowship.

## REFERENCES

- Goffeau A. Genes in search of functions. *Nature* (London) 1994; 369: 101-102.
- Cole S T, Brosch R, Parkhill J et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* (London) 1998; 393: 537-544.
- Cole S T, Barrel B G. Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. In: Chadwick D J, Cardew G (eds.), *Genetics and tuberculosis* (Novartis Foundation Symposium 217) J Wiley, Chichester, 1998: 160-172.
- Altschul S, Gish W, Miller W, Myers E, Lipman D. A basic local alignment search tool. *J Mol Biol* 1990; 215: 403-410.
- Wootton J C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput & Chem* 1994; 18: 269-285.
- Tekaia F and Dujon B. Pervasiveness of gene conservation and persistence of duplicatges in cullular genomes. *J Mol Evol* 1999; 49: in press:
- Greenacre M. *Theory and application of correspondence analysis*. Academic Press, London, 1984:
- Benzecri J-P. *L'Analyse des Données*. 2 Dunod, Paris, 1973:
- Bailey T L, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994; 28-36.
- Bailey T L, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 1998; 14: 48-45.
- Devereux J, Haerberli P, Smithies O. A comprehensive set of sequence analysis programs for the VAX. *Nucl Acids Res* 1984; 12: 387-395.
- Pearson W, Lipman D. Improved tools for biological sequence comparisons. *Proc Natl Acad USA* 1988; 85: 2444-2448.
- Nielsen H, Engelbrecht J, Drunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Prot Eng* 1997; 10: 1-6.
- Sonnhammer E L L, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, 1998: 175-182.
- Thompson J D, Higgin D G, Gibson T J, Clustal W. Improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994; 22: 4673-4680.
- Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989; 5: 164-166.
- Poulet S, Cole S T. Characterisation of the polymorphic GC-rich repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Arch Microbiol* 1995; 163: 87-95.

18. Ross B C, Raios K, Jackson K, Dwyer B. Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. *J Clin Microbiol* 1992; 30: 942–946.
19. van Sooling D, de Haas P E W, Hermans P W M, Groenen P M A, van Embden J D A. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. *J Clin Microbiol* 1993; 31: 1987–1995.
20. Grosset J. Bacteriology of tuberculosis. In: Reichmann L B, Hershfield E S, eds. *Tuberculosis: A comprehensive international approach*. Marcel Dekker, NY, 1993: 49–74.
21. Gordon S V, Heym B, Parkhill J, Barrell B, Cole S T. New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 1999; 145: 881–892.
22. Lonetto M, Gribskov M, Gross C A. The  $\sigma^{70}$  family; Sequence conservation and evolutionary relationships. *J Bacteriol* 1992; 174: 3843–3849.
23. Gomez J E, Chen J-M, Bishai W R. The sigma factors of *Mycobacterium tuberculosis*. *Tubercle Lung Dis* 1998; 78: 175–183.
24. Barry C E, III, Lee R E, Mdulki K, Sampson A E, Schroder B G, Slayden R A, Yuan Y. Mycolic acids: Structure, biosynthesis, and physiological functions. *Prog Lipid Res* 1998; 37: 143–157.
25. Munro A W, Lindsay J G. Bacterial cytochromes P-450. *Mol Microbiol* 1996; 20: 1115–1125.
26. Arand M, Grant D F, Beetham J K, Friedberg T, Hammock B D O F. Sequence similarity of mammalian epoxide hydrolases to the bacterial haloalkane dehalogenase and other related proteins. *FEBS Lett* 1994; 338: 251–256.
27. Hofmann B, Tolzer S, Pelletier I, Altenbuchner J, van P K, Hecht H J. Structural investigation of the cofactor-free chloroperoxidases. *J Mol Biol* 1998; 279: 889–900.
28. Pelletier I, Altenbuchner J, Mattes R. A catalytic triad is required by the non-heme haloperoxidases to perform halogenation. *Biochim Biophys Acta* 1995; 1250: 149–157.
29. Pelletier I, Altenbuchner J. A bacterial esterase is homologous with non-haem haloperoxidases and displays brominating activity. *Microbiology* 1995;
30. Saier J, M H, Tam R, Reizer A, Reizer J. Two novel families of bacterial membrane proteins concerned with nodulation, cell division and transport. *Mol Microbiol* 1994; 11: 841–847.
31. Paulsen I T, Brown M H, Skurray R A. Proton-dependent multidrug efflux systems. *Microbiol Rev* 1996; 60: 575–608.
32. Kunst F, Ogasawara N, Moszer I et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature (London)* 1997; 390: 249–256.
33. Pao S S, Paulsen I T, Saier M J. Major facilitator superfamily. *Microbiology and Molecular Biology Reviews* 1998; 62: 1–34.
34. Paulsen I T, Sliwinski M K, Saier M J. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J Mol Biol* 1998; 277: 573–592.
35. Ainsa J A, Blokpoel M C J, Otal I, Young D B, de Smet K A L, Martin C. Molecular cloning and characterization of Tap, a putative multidrug efflux pump present in *Mycobacterium fortuitum* and *Mycobacterium tuberculosis*. *J Bacteriol* 1998; 180: 5836–5843.
36. Kanamaru K, Kashiwagi S, Mizuno T. A copper-transporting P-type ATPase found in the thylakoid membrane of the cyanobacterium *Synechococcus* species PCC7942. *Mol Microbiol* 1994; 13: 369–377.
37. Arruda S, Bomfim G, Knight R, Huima-Byron T, Riley L W. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. *Science* 1993; 261: 1454–1457.
38. Elhay M J, Oettinger T, Andersen P. Delayed-type hypersensitivity responses to ESAT-6 and MPT64 from *Mycobacterium tuberculosis* in the guinea pig. *Infect Immun* 1998; 66: 3454–3456.
39. Lalvani A, Brookes R, Wilkinson R J et al. Human cytolytic and interferon gamma-secreting CD8+ T lymphocytes specific for *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America* 1998; 95: 270–275.
40. Harboe M, Oettinger T, Wiker H G, Rosenkrands I, Andersen P. Evidence for occurrence of the ESAT-6 protein in *Mycobacterium tuberculosis* and virulent *Mycobacterium bovis* and for its absence in *Mycobacterium bovis* BCG. *Infect Immun* 1996; 64: 16–22.
41. Mahairas G G, Sabo P J, Hickey M J, Singh D C, Stover C K. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. *J Bacteriol* 1996; 178: 1274–1282.
42. Gordon S V, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole S T. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol* 1999; 32: 643–656.
43. Chubb A J, Woodman Z L, da Silva Tatley F M P R, Hoffman H J, Scholle R R, Ehlers M R W. Identification of *Mycobacterium tuberculosis* signal sequences that direct the export of a leaderless beta-lactamase gene product in *Escherichia coli*. *Microbiology* 1998; 144: 1619–1629.
44. Pugsley A P. The complete general secretory pathway in gram-negative bacteria. *Microbiol Rev* 1993; 57: 50–108.
45. Brennan P J. Proteins and antigens of *Mycobacterium tuberculosis*. In: Bloom B R (ed.), *Tuberculosis: pathogenesis, protection, and control*. American Society for Microbiology, Washington DC 20005, 1994: 307–336.
46. Berthet F X, Rasmussen P B, Rosenkrand I, Andersen P, Gicquel B. A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low-molecular-mass culture filtrate protein (CFP-10). *Microbiology* 1998; 144: 3195–3203.
47. Finlay B B, Falkow S. Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev* 1997; 61: 136–169.
48. Nissen M J, Nes I F. Ribosomally synthesized antimicrobial peptides: their function, structure, biogenesis, and mechanism of action. *Arch Microbiol* 1997; 167: 67–77.
49. Havarstein L S, Diep D B, Nes I F. A family of bacteriocin ABC transporters carry out proteolytic processing of their substrates concomitant with export. *Mol Microbiol* 1995; 16: 229–240.
50. Menard R, Sansonetti P, Parsot C. The secretion of the *Shigella flexneri* Ipa invasins is activated by epithelial cells and controlled by IpaB and IpaD. *Embo Journal* 1994; 13: 5293–5302.
51. Menozzi F D, Bischoff R, Fort E, Brennan M J, Locha C. Molecular characterization of the mycobacterial heparin-binding hemagglutinin, a mycobacterial adhesin. *Proc Natl Acad Sci USA* 1998; 95: 12625–12630.
52. Weldingh K, Rosenkrands I, Jacobsen S, Rasmussen P B, Elhay M J, Andersen P. Two-dimensional electrophoresis for analysis of *Mycobacterium tuberculosis* culture filtrate and purification and characterization of six novel proteins. *Infect Immun* 1998; 66: 349–500.