Sequence analysis corresponding to the PPE and PE proteins in *Mycobacterium tuberculosis* and other genomes

SWATHI ADINDLA and LALITHA GURUPRASAD*

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

**Corresponding author (Fax, 91-40-23012460; Email, lgpsc@uohyd.ernet.in)*

Amino acid sequence analysis corresponding to the PPE proteins in H37Rv and CDC1551 strains of the *Mycobacterium tuberculosis* genomes resulted in the identification of a previously uncharacterized 225 amino acid-residue common region in 22 proteins. The pairwise sequence identities were as low as 18%. Conservation of amino acid residues was observed at fifteen positions that were distributed over the whole length of the region. The secondary structure corresponding to this region is predicted to be a mixture of α-helices and β-strands. Although the function is not known, proteins with this region specific to mycobacterial species may be associated with a common function. We further observed another group of 20 PPE proteins corresponding to the conserved C-terminal region comprising 44 amino acid residues with GFxGT and PxxPxxW sequence motifs. This region is preceded by a hydrophobic region, comprising 40–100 amino acid residues, that is flanked by charged amino acid residues. Identification of conserved regions described above may be useful to detect related proteins from other genomes and assist the design of suitable experiments to test their corresponding functions. Amino acid sequence analysis corresponding to the PE proteins resulted in the identification of tandem repeats comprising 41–43 amino acid residues in the C-terminal variable regions in two PE proteins (Rv0978 and Rv0980). These correspond to the AB repeats that were first identified in some proteins of the *Methanosarcina mazei* genome, and were demonstrated as surface antigens. We observed the AB repeats also in several other proteins of hitherto uncharacterized function in *Archaea* and *Bacteria* genomes. Some of these proteins are also associated with another repeat called the C-repeat or the PKD-domain comprising 85 amino acid residues. The secondary structure corresponding to the AB repeat is predicted mainly as 4 β-strands. We suggest that proteins with AB repeats in *Mycobacterium tuberculosis* and other genomes may be associated as surface antigens. The *M. leprae* genome, however, does not contain either the AB or C-repeats and different proteins may therefore be recruited as surface antigens in the *M. leprae* genome compared to the *M. tuberculosis* genome.

[Adindla S and Guruprasad L 2003 Sequence analysis corresponding to the PPE and PE proteins in *Mycobacterium tuberculosis* and other genomes; *J. Biosci*. **28** 169–179]

1. Introduction

The complete genome sequences of *Mycobacterium tuberculosis* (H37Rv strain) consists of 4,411,529 base pairs and approximately 3,986 proteins (Cole *et al* 1998) and the CDC1551 strain consists of 4,403,836 base pairs and approximately 4,187 proteins (Fleischmann *et al* 2001). Nearly 10% of the proteins in each of these genomes encode the PPE and PE protein families (Cole *et al* 1998). The names PE and PPE derive from the amino acid sequence motifs Pro-Glu (PE) and Pro-Pro-Glu (PPE) located near the N-terminus in a majority of these proteins. Members of the PE and PPE protein families are characterized by highly conserved N-terminal domains with approximately 110 and 180 amino acid residues, respectively. The C-terminus, however, varies considerably

Keywords. AB repeats; *Mycobacterium tuberculosis* genome; PE-PPE domain; PPE, PE proteins; sequence analysis; surface antigens

in sequence and the number of amino acid residues. These family of proteins were also identified in the *Mycobacterium leprae* genome (Cole *et al* 2001).

According to Cole *et al* (1998), the amino acid sequences corresponding to the PPE protein family have been classified into three categories: (i) proteins with the NxGxGNxG, major polymorphic tandem repeats (MPTR) sequence motif, (ii) proteins with a conserved GxxSV PxxW motif around position 350 along the sequence and (iii) other 'unrelated' proteins. As of date, there is neither structure nor function data available for any member of the PPE or PE protein family. However, it has been speculated that these proteins may have an immunological role (Cole *et al* 1998). For instance, few PPE proteins that do not consist the NxGxGNxG sequence motif, have similarity to one of the major serine-rich antigens recognized by leprosy patients (Vega Lopez *et al* 1993). It has therefore been indicated that members of the PE and PPE protein families might play a role in antigenic variation or interfere with immune responses by inhibiting antigen processing.

Further, the C-terminus corresponding to most PE proteins varies in length ranging from 100–1,400 amino acid residues. Cole *et al* (1998) classified PE proteins into several subfamilies based on phylogeny analysis. The largest of these is the polymorphic repetitive sequence class (PGRS) characterized by a high glycine content (up to 50%) as a result of multiple tandem repeats of the Gly-Gly-X type. Most members of the PGRS subfamily comprise the PE domain followed by the PGRS region described above. However, members of other PE subfamilies are known to share very little C-terminal sequence similarity.

In the present work, we therefore intended to: (i) characterize 'unrelated' PPE proteins corresponding to the third category according to the classification by Cole *et al* (1998); and (ii) to explore sequence similarities corresponding to the variable C-terminal region in PE protein family, in light of new sequence data accumulating in databanks as a result of several completed genome projects.

2. Methods

We selected the PPE and PE proteins in *M. tuberculosis* genome using SRS (Schaftenaar *et al* 1996) available at the website www.ebi.ac.uk. The results were manually inspected in order to exclude entries corresponding only to sequence fragments. In the case of PPE proteins, we selected only those proteins classified as 'unrelated' according to Cole *et al* (1998) and in case of the PE proteins we selected sequences corresponding only to the C-terminal variable regions (Cole *et al* 1998). We excluded the N-terminal 180 amino acid residue region characteristic of the PPE protein family from the individual query sequences and searched the SWALL database (Henning *et al* 2003) using the WU-Blast2 program (Altshul *et al* 1990) available at the website www.ebi.ac.uk and against the non-redundant Genbank database using the PSI-BLAST program (Altshul *et al* 1997) available at the website www.ncbi.nlm.nih.gov/BLAST/. Also we carried out similar searches against the unfinished microbial genomes using the BLASTP program available at the website www.ncbi.nlm.nih.gov/BLAST/. BLASTP is a reliable and rapid program to identify all known homologs/analogs to a query protein sequence in a given database. The blosum62 matrices were used and hits were sorted on *P*-value in the WU-Blast2 program. The results of searches were used for database searches once again in order to be able to retrieve the original query sequence (reciprocal searches). Sequences identified from above searches were aligned using the multiple sequence alignment program CLUSTALW (Thompson *et al* 1994) available at the website www.ebi.ac.uk. The default parameters corresponding to a penalty of 10 for opening a gap, 0·05 for gap extension and penalty 8 for gap separation was assigned for the alignment. The secondary structure predictions corresponding to individual amino acid sequences were carried out using the PHD program that uses the method of neural network (Rost *et al* 1994). The PHD method provides more than 70% secondary structure-prediction accuracy.

3. Results and discussion

3.1 *PPE protein family*

Three PPE proteins (Rv1800, Rv3539 and Rv2608) showed sequence similarity corresponding to their C-terminal regions. In order to detect other proteins that possibly shared this similarity, we searched the SWALL database in SRS with the amino acid sequence corresponding to the region 245-469 in Rv1800 using the WU-Blast2 program. This identified PE proteins (Rv1430, Rv0151, Rv0152, Rv0159, Rv0160) in *M. tuberculosis* genome. Reciprocal searches described earlier with the newly detected protein sequences retrieved the original query sequence and additional proteins with significant *P*-values (e^{-20}). Similarities were detected that corresponded to hypothetical proteins in *M. tuberculosis* (Rv3822, Rv1184) and *M. leprae* (ML1232, q49633) genomes. In all twenty-two proteins were retrieved as a result of these searches. We term this conserved region that is common to some members of the PE and PPE protein families as the 'PE-PPE domain'. The PE-PPE domain comprises approximately 225 amino acid residues. The schematic representation of the PE-PPE domain observed in different proteins is shown in figure 1. The multiple sequence

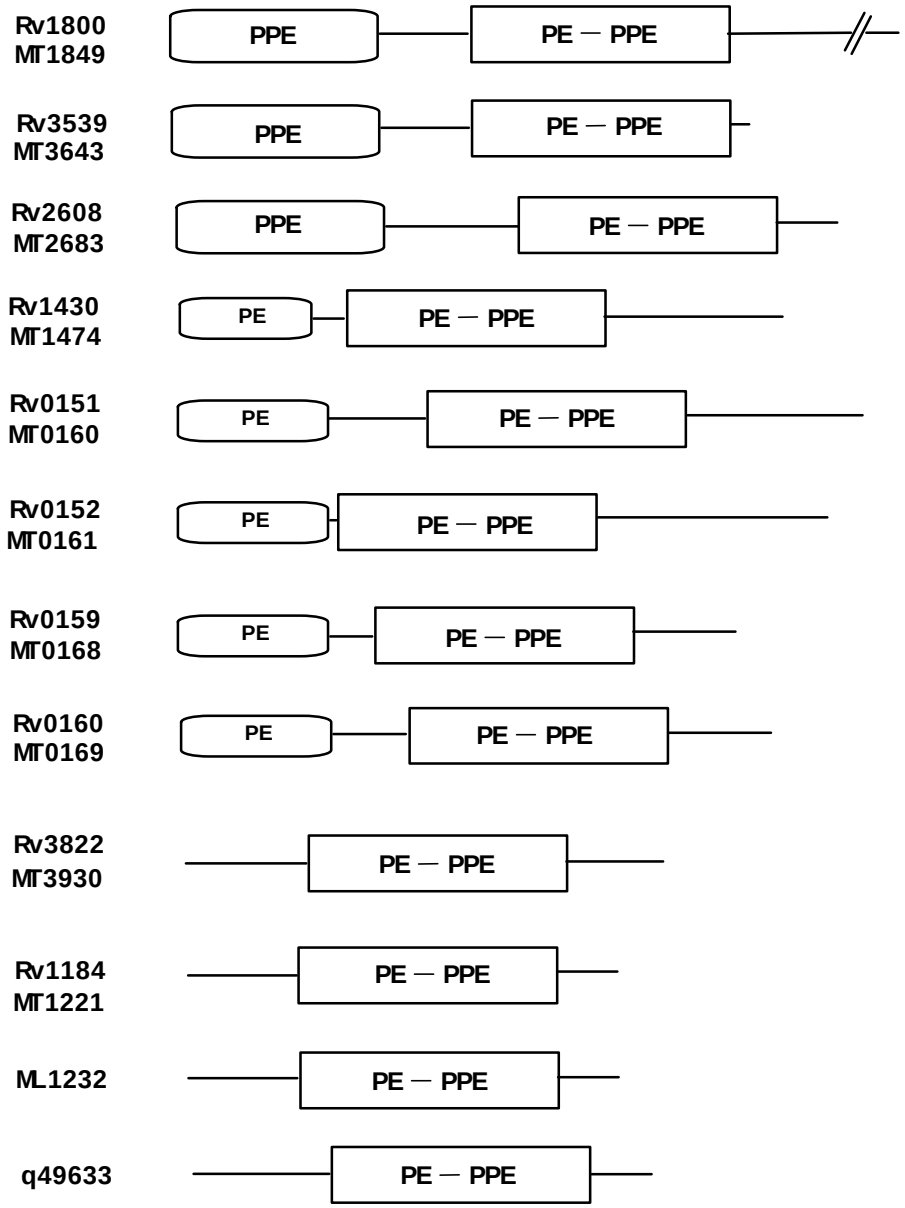

Figure 1. Schematic representation of proteins containing the PE-PPE domain. Abbreviations: PPE, conserved 180 amino acid-residue domain; PE, conserved 110 amino acid-residue domain; PE-PPE, conserved 225 amino acid-residue domain.

alignment corresponding to the PE-PPE domain is shown in figure 2a. Some of these protein sequences share as low as 18% sequence identity. There is a conservation of amino acid residues, at fifteen positions that is distributed over whole length of the domain. The secondary structure predicted for individual sequences corresponding to this domain suggests a mixture of *a*-helices and *b*-strands arranged in the order H1-E1-H2-E2-E3-E4-E5-H3-E6-E7-H4 that is likely to have a compact fold. The lengths of *a*-helices and *b*-strands may vary in the individual proteins. No significant hits were observed that corresponded to sequences of known three-dimensional structure. It may be noted that the PE-PPE domain may not have a fixed location in the host protein. The dendrogram showing relatedness of the sequences based on sequence identity is shown in figure 2b. It is clear that the PPE proteins, PE proteins and hypothetical proteins fall into distinct clusters with the PE and PPE proteins more closely related to each other than to the hypothetical proteins. The identification of this domain only in *M. tuberculosis* and *M. leprae* genomes suggests that it may be specific to the mycobacterial species. Further, we believe

that although the PE and PPE proteins are grouped into separate families, some members constituting this domain may have common ancestors and be associated with a common evolution and function. This domain is not overlapping with the conserved N-terminal PE or PPE region already described by Cole *et al* (1998).

Further, we observed another set of 20 PPE proteins in the 'unrelated' category of proteins according to Cole *et al* (1998). These are characterized by a highly conserved 44 amino acid residue region in the C-terminus. These regions were observed only in the PPE protein family in *M. tuberculosis* and *M. leprae* genomes. The schematic representation indicating the length and location of these regions in the different PPE proteins is shown in figure 3. One protein (q9ans8) in the *Mycobacterium avium* genome is not indicated in figure 3 although it contains the highly conserved 44 amino acid region as the protein sequence available in the SWALL database corresponds only to the fragment sequence. This 44 amino acid residue region comprises highly conserved

```
     Secondary structure                                  hhhhhhhhhhhhhhhhh      eeeeeee
1.   Rv2608(300-525)      -PGYTATFLETPSQFFPFTG-LNSLTYDVSVAQGVTNLHTAIMAQLA-AGNEVVVFGTSQ
2.   MT2683(300-525)      -PGYTATFLETPSQFFPFTG-LNSLTYDVSVAQGVTNLHTAIMAQLA-AGNEVVVFGTSQ
3.   Rv1430(144-368)      -LGYAFSGLYTPAQFQPWTG-IPSLTYDQSVAEGAGYLHTAIMQQVA-AGNDVVVLGFSQ
4.   MT1474(144-368)      -LGYAFSGLYTPAQFQPWTG-IPSLTYDQSVAEGAGYLHTAIMQQVA-AGNDVVVLGFSQ
5.   Rv1800(245-469)      -PGVIAQALFTPQGLYPVVV-IKNLTFDSSVAQGAVILESAIRQQIA-AGNNVTVFGYSQ
6.   MT1849(245-469)      -PGVIAQALFTPQGLYPVVV-IKNLTFDSSVAQGAVILESAIRQQIA-AGNNVTVFGYSQ
7.   Rv0151(193-415)      ---PVVKALVTPEELYPITG-VKSLPFQTSVQLGLQILDGAIWEQIN-AGNHVTVFGYSQ
8.   MT0160(193-415)      ---PVVKALVTPEELYPITG-VKSLPFQTSVQLGLQILDGAIWEQIN-AGNHVTVFGYSQ
9.   Rv0152(109-335)      ---RALQAVFTPEELYPLTG-VRSLVLNTSVEEGLTILHDAIMVELATTGNAVTVFGWSQ
10.  MT0161(117-343)      ---RALQAVFTPEELYPLTG-VRSLVLNTSVEEGLTILHDAIMVELATTGNAVTVFGWSQ
11.  Rv0159(151-375)      PNNPVAQ--YTPEQWWPF---IGNLSLDQSIAQGVTLLNNGINAELQ-NGHDVVVFGYSQ
12.  MT0168(151-375)      PNNPVAQ--YTPEQWWPF---IGNLSLDQSIAQGVTLLNNGINAELQ-NGHDVVVFGYSQ
13.  Rv0160(172-398)      PG-AVSQGLFTPEQFWPVTPDLGNLTFNQSVTEGVALLNTAVNNQLA-LDNKVVAFGYSQ
14.  MT0169(172-398)      PG-AVSQGLFTPEQFWPVTPDLGNLTFNQSVTEGVALLNTAVNNQLA-LDNKVVAFGYSQ
15.  Rv3539(240-461)      -PGTTPEVVSYPATIGVLSGSLGAVDANQSIAIGQQMLHNEILAATASGQ-PVTVAGLSM
16.  MT3643(240-461)      -PGTTPEVVSYPATIGVLSGSLGAVDANQSIAIGQQMLHNEILAATASGQ-PVTVAGLSM
17.  Rv3822(104-324)      -PTATRHVVSYPGSFWP-VTGLNSPTVGSSVSAGTNNLDAAIRSTDG----PIFVAGLSQ
18.  MT3930(104-324)      -PTATRHVVSYPGSFWP-VTGLNSPTVGNSVSAGTNNLDAAIRSTDG----PIFVAGLSQ
19.  ML1232(86-309)       --PARAPFRYVPTFLVP--GPRDEVTIGEAIAVATKNLNQAIHRGTE----PAAVVGLSQ
20.  q49633(119-342)      --PARAPFRYVPTFLVP--GPRDEVTIGEAIAVATKNLNQAIHRGTE----PAAVVGLSQ
21.  Rv1184(89-316)       --PAGAAFSWWPTMLLPPGSHQDNMTVGVAVKDGTNSLDNAIHHGTD----PAAAVGLSQ
22.  MT1121(101-328)      --PAGAAFSWWPTMLLPPGSHQDNMTVGVAVKDGTNSLDNAIHHGTD----PAAAVGLSQ

     consensus/80%        ...hshphh.hPtthhP.ss.ltslshspSlt.Ght.LpsAIhtths.....lsVhGhSQ
                                     *                       *                 * *
```

```
     Secondary structure          hhhhhhhhhhhh            eeeeeeee         eeeeeee    eeeeeee
     Rv2608(300-525)      SATIATFEMRYLQSLPAHLRPGLDELSFTLTGNPNR-----PDGGILTRF-GFSIPQLGF
     MT2683(300-525)      SATIATFEMRYLQSLPAHLRPGLDELSFTLTGNPNR-----PDGGILTRF-GFSIPQLGF
     Rv1430(144-368)      GASVATLEMRHLASLPAGVAPSPDQLSFVLLGNPNN-----PNGGILARFPGLYLQSLGL
     MT1474(144-368)      GASVATLEMRHLASLPAGVAPSPDQLSFVLLGNPNN-----PNGGILARFPGLYLQSLGL
     Rv1800(245-469)      SATISSLVMANLAAS--ADPPSPDELSFTLIGNPNN-----PNGGVATRFPGISFPSLGV
     MT1849(245-469)      SATISSLVMANLAAS--ADPPSPDELSFTLIGNPNN-----PNGGVATRFPGISFPSLGV
     Rv0151(193-415)      SAVIASLEMQHLISLG-PNAPSPSQLNFILIGNEMN-----PNGGILARIPGLNVTTLGL
     MT0160(193-415)      SAVIASLEMQHLISLG-PNAPSPSQLNFILIGNEMN-----PNGGILARIPGLNVTTLGL
     Rv0152(109-335)      SAIIASLEMQRFTAMG-GAAPSASDLNFVLVGNEMN-----PNGGMLARFPDLTLPTLDL
     MT0161(117-343)      SAIIASLEMQRFTAMG-GAAPSASDLNFVLVGNEMN-----PNGGMLARFPDLTLPTLDL
     Rv0159(151-375)      SAAVATNEIRALMALPPGQAPDPSRLAFTLIGNINN-----PNGGVLERYVGLYLPFLDM
     MT0168(151-375)      SAAVATNEIRALMALPPGQAPDPSRLAFTLIGNINN-----PNGGVLERYVGLYLPFLDM
     Rv0160(172-398)      SATIINNYINSLMAMG---SPNPDDISFVMIGSGNN-----PVGGLLARFPGFYIPFLDV
     MT0169(172-398)      SATIINNYINSLMAMG---SPNPDDISFVMIGSGNN-----PVGGLLARFPGFYIPFLDV
     Rv3539(240-461)      GSMVIDRELAYLAIDP--NAPPSSALTFVELAGPE------RGLAQTYLPVGTTIPI--A
     MT3643(240-461)      GSMVIDRELAYLAIDP--NAPPSSALTFVELAGPE------RGLAQTYLPVGTTIPI--A
     Rv3822(104-324)      GTLVLDREQARLANDP--TAPPPGQLTFIKAGDPN-------NLLWRAFRPGTHVPI--I
     MT3930(104-324)      GTLVLDREQARLANDP--TAPPPGQLTFIKAGDPN-------NLLWRAFRPGTHVPI--I
     ML1232(86-309)       GSLALDTEQEQLATDP--TAPPPDQLTFNTFGDPSGYHGFGKSVLASIFRPGDYIPL--I
     q49633(119-342)      GSLALDTEQEQLATDP--TAPPPDQLTFNTFGDPSGYHGFGKSVLASIFRPGDYIPL--I
     Rv1184(89-316)       GSLVLDQEQARLANDP--TAPAPDKLQFTTFGDPTGRHAFGASFLARIFPPGSHIPIPFI
     MT1121(101-328)      GSLVLDQEQARLANDP--TAPAPDKLQFTTFGDPTGRHAFGASFLARIFPPGSHIPIPFI

     consensus/80%        uuhlhs.E.ttLhs.s...uPssspLsFshhGs.pt......shhhhhhhsGh.lP...h
                                           *      *
```

Figure 2a. (*Continued*).

```
Secondary structure                 eeeeeeeee         hhhhhhhhhhhhh    eeeeee
Rv2608(300-525)       TLSGATP-ADAYPTVDYAFQYDGVNDFPKYPLNVFATANAIAG-ILFLHSGLIALPPD-L
MT2683(300-525)       TLSGATP-ADAYPTVDYAFQYDGVNDFPKYPLNVFATANAIAG-ILFLHSGLIALPPD-L
Rv1430(144-368)       TFNGATP-DTDYATTIYTTQYDGFADFPKYPLNILADVNALLG-IYYSHSLYYGLTPEQV
MT1474(144-368)       TFNGATP-DTDYATTIYTTQYDGFADFPKYPLNILADVNALLG-IYYSHSLYYGLTPEQV
Rv1800(245-469)       TATGATP-HNLYPTKIYTIEYDGVADFPRYPLNFVSTLNAIAG-TYYVHSNYFILTPEQI
MT1849(245-469)       TATGATP-HNLYPTKIYTIEYDGVADFPRYPLNFVSTLNAIAG-TYYVHSNYFILTPEQI
Rv0151(193-415)       PFYGATP-DNPYPTTTYTLEYDGFADFPRYPLNVLSDINAVFG-ILTVHTTYADLTPAQI
MT0160(193-415)       PFYGATP-DNPYPTTTYTLEYDGFADFPRYPLNVLSDINAVFG-ILTVHTTYADLTPAQI
Rv0152(109-335)       TFYGATPSDTIYPTAIYTLEYDGFADFSRYPLNFISDLNAVAG-ITFVHTKYLDLTPAQV
MT0161(117-343)       TFYGATPSDTIYPTAIYTLEYDGFADFSRYPLNFISDLNAVAG-ITFVHTKYLDLTPAQV
Rv0159(151-375)       SFNGATPPDSPYQTYMYTGQYDGYAHNPQYPLNILSDLNAFMG-IRWVHNAYP-FTAAEV
MT0168(151-375)       SFNGATPPDSPYQTYMYTGQYDGYAHNPQYPLNILSDLNAFMG-IRWVHNAYP-FTAAEV
Rv0160(172-398)       PFNGATPANSPYPTHIYTAQYDGIAHAPQFPLRILSDINAFMG-YFYVHNTYPELMATQV
MT0169(172-398)       PFNGATPANSPYPTHIYTAQYDGIAHAPQFPLRILSDINAFMG-YFYVHNTYPELMATQV
Rv3539(240-461)       GYTVGNAPESQYNTSVVYSQYDIWADPPDRPWNLLAGANALMG-AAYFHDLTAYAAPQ--
MT3643(240-461)       GYTVGNAPESQYNTSVVYSQYDIWADPPDRPWNLLAGANALMG-AAYFHDLTAYAAPQ--
Rv3822(104-324)       DYTVPAPAESQYDTINIVGQYDIFSDPPNRPGNLLADLNAIAAGGYYGHSATAFSDPARV
MT3930(104-324)       DYTVPAPAESQYDTINIVGQYDIFSDPPNRPGNLLADLNAIAAGGYYGHSATAFSDPARV
ML1232(86-309)        DYTMPQRMDSQYDSNRVVAAYDGLSEFPDRADNLLAQLNCFAG-GAISHTPSGFFNPEDV
q49633(119-342)       DYTMPQRMDSQYDSNRVVAAYDGLSEFPDRADNLLAQLNCFAG-GAISHTPSGFFNPEDV
Rv1184(89-316)        EYTMPQQVDSQYDTNHVVTAYDGFSDFPDRPDNLLAVANAAIG-AAIAHTPIGFTGPGDV
MT1121(101-328)       EYTMPQQVDSQYDTNHVVTAYDGFSDFPDRPDNLLAVANAAIG-AAIAHTPIGFTGPGDV

consensus/80%         sasssss.cs.YsT.hhshpYDGhu-.PphPhNllushNAhhG.hhhsHs.hh.hsPtpl
                                *        **               *        *

Secondary structure                       eeeeeee                  hhhhhhhhhhhhhhhh
Rv2608(300-525)       ASGVVQPVSSP-DVLTTYILLPSQDLPLLVPLRAIPLLGNPLADLIQPDLRVLVELGY
MT2683(300-525)       ASGVVQPVSSP-DVLTTYILLPSQDLPLLVPLRAIPLLGNPLADLIQPDLRVLVELGY
Rv1430(144-368)       ASGIVLPVSSP-DTNTTYILLPNEDLPLLQPLRGI--VPEPLLDLIEPDLRAIIELGY
MT1474(144-368)       ASGIVLPVSSP-DTNTTYILLPNEDLPLLQPLRGI--VPEPLLDLIEPDLRAIIELGY
Rv1800(245-469)       DAAVPLTNTVG-PTMTQYYIIRTENLPLLEPLRSVPIVGNPLANLVQPNLKVIVNLGY
MT1849(245-469)       DAAVPLTNTVG-PTMTQYYIIRTENLPLLEPLRSVPIVGNPLANLVQPNLKVIVNLGY
Rv0151(193-415)       ASATQLPTQG--TTSNTYYIIETEHLPLLAPLRAIPVIGPPLAALVEPNLEVIVNLGY
MT0160(193-415)       ASATQLPTQG--TTSNTYYIIETEHLPLLAPLRAIPVIGPPLAALVEPNLEVIVNLGY
Rv0152(109-335)       EGATKLPTSPGYTGVTDYYIIRTENRPLLQPLRAVPVIGDPLADLIQPNLKVIVNLGY
MT0161(117-343)       EGATKLPTSPGYTGVTDYYIIRTENRPLLQPLRAVPVIGDPLADLIQPNLKVIVNLGY
Rv0159(151-375)       ANAVPLPTSPGYTGNTHYYMFLTQDLPLLQPIRAIPFVGTPIAELIQPDLRVLVDLGY
MT0168(151-375)       ANAVPLPTSPGYTGNTHYYMFLTQDLPLLQPIRAIPFVGTPIAELIQPDLWVLVDLGY
Rv0160(172-398)       DNAVPLPTSPGYTGNTQYYMFLTQDLPLLQPIRDIPYAGPPIADLFQPQLRVLVDLGY
MT0169(172-398)       DNAVPLPTSPGYTGNTQYYMFLTQDLPLLQPIRDIPYAGPPIADLFQPQLRVLVDLGY
Rv3539(240-461)       QGIEIAAVTSSLGGTTTTYMIPSPTLPLLLPLKQIG-VPDWIVGGLNNVLKPLVDAGY
MT3643(240-461)       QGIEIAAVTSSLGGTTTTYMIPSPTLPLLLPLKQIG-VPDWIVGGLNNVLKPLVDAGY
Rv3822(104-324)       APRDITTTTNSLGATTTTYFIRTDQLPLVRALVDMAGLPPQAAGTVDAALRPIIDRAY
MT3930(104-324)       APRDITTTTNSLGATTTTYFIRTDQLPLVRALVDMAGLPPQAAGTVDAALRPIIDRAY
ML1232(86-309)        PPQNIRTTVNSRGAKTTTYLIPVNHLPLTLPLRYLG-WSDALVDQIDAVLQPKIDAAY
q49633(119-342)       PPQNIRTTVNSRGAKTTTYLIPVNHLPLTLPLRYLG-WSDALVDQIDAVLQPKIDAAY
Rv1184(89-316)        PPQNIRTTVNSRGATTTTYLVPVNHLPLTLPLRYLG-MSDAEVDQIDSVLQPQIDAAY
MT1121(101-328)       PPQNIRTTVNSRGATTTTYLVPVNHLPLTLPLRYLG-MSDAEVDQIDSVLQPQIDAAY

consensus/80%         ssts.hssssss.ss.TphYhl.sppLPLl.PLRtls.hssslsshlpssLpsllshuY
                                           **                  *       *
```

Figure 2a. Multiple sequence alignment corresponding to the conserved 225 amino acid-residue PE-PPE domain produced using the CLUSTALW program. Numbers indicated within brackets correspond to the start and end amino acid-residue number within the domain. Amino acid residues are coded according to the 80% consensus as derived from the website www.bork.embl-heidelberg.de/Alignment/consensus.html. polar (p: CDEHKNQRST); hydrophobic (h: ACFGHIKLMRTVWY); small (s: ACDGNPSTV); tiny (u: AGS); aliphatic (l: ILV); turn-like (t: ACDEGHKNQRST); aromatic (a: FHWY). A capital letter indicates ~ 80% conservation of amino acid residue and (∗) indicates 100% conservation. The secondary structure is derived from PHD method. Amino acid residues predicted to be in the helical conformation are represented as 'h' and in strand conformation as 'e'.

sequence motifs GFxGT and PxxPxxW shown in figure 4. We further observed that this 44 amino acid region is preceded by a hydrophobic region varying in length between 40–100 amino acid residues that is flanked by regions of varying length rich in charged amino acid residues as indicated in figure 3. The secondary structure predicted for the hydrophobic regions flanked by charged amino acid residues suggests a mixture of ***a***-helices and

b-strands and a loop conformation for the conserved 44 amino acid-residue region. We do not understand the significance of the sequence motifs in the highly conserved 44 amino acid-residue C-terminal region in PPE proteins, although it is likely that these mycobacterial proteins may constitute a new sub-family that may be associated with a common function possibly through the predicted conserved loop.

Further, one 'unrelated' PPE protein; Rv3873 (corresponding to MT3987 in CDC1551 strain) in the *M. tuberculosis* genome is observed to be homologous to a protein (ML0051) from *M. leprae* genome and two other 'unrelated' proteins in *M. tuberculosis* genome; Rv3892 and Rv3144 (corresponding to MT4007 and MT3231, respectively, in the CDC1551 strain) do not show significant similarity to any proteins in the SWALL database suggesting that these may be unique to the *M. tuberculosis* genome.

3.2 *PE protein family*

Two PE proteins (Rv0978 and Rv0980) in *M. tuberculosis* genome contained tandem repeats comprising 41–43 amino acid residues corresponding to the variable C-terminal region. Further, database searches with each of these sequences detected other proteins containing this tandem repeat from *Methanosarcina mazei* (q50245, q50246, q977v5), hypothetical proteins from *Methylobacterium extorquens* (o30796), *Rhizobium meliloti* RB0179, *Sulfolobus tokodaii* ST1748 and two non-PE proteins from *M. tuberculosis*; Rv1057 (H37Rv strain) and MT1087 (CDC1551 strain). Reciprocal searches with sequences corresponding to the tandem repeat from these proteins once again identified the seed query and several proteins from other species belonging to *Bacteria* and *Archaea* genomes and to four new proteins corresponding to unfinished microbial genomes; a 382 amino acid-residue protein from *Rhodobacter sphaeroides* genome (doe_1063), a 544 amino acid-residue protein from *Rhodopseudomonas palustris* genome (doe_1076), a 912 amino acid-residue protein from *Nitrosomonas europaea* genome (doe_915) and a 827 amino acid residue protein from *Cytophaga hutchinsonii* genome (doe_985). These correspond to proteins of hitherto unknown function. The list of proteins containing the tandem repeat is shown in table 1. Literature searches indicated that the tandem AB repeat was first observed in the *M. mazei* genome (Mayerhofer *et al* 1995). It was recognized as a surface antigen based on experiments designed to elucidate the antigenic mosaic of *M. mazei* with antibodies. Three genes in *M. mazei* genome; *orf*492, *orf*375 and *orf*783 (corresponding to SWALL identities; q50245, q50246 and q977v5, respectively) were reported to code for proteins recognized by antibodies against the cell surface antigens. The multiple sequence alignment corresponding to AB repeat and the conservation pattern of individual amino acid residues is shown in figure 5. The secondary structure corresponding to the AB repeat is predicted to comprise four *b*-strands. The individual *b*-strands may, however, vary in length. We noted that two consecutive AB repeats is referred as NHL repeat (accession number: PF01436) according to the Pfam classification in the protein family domain database (Bateman *et al* 2000) (see figure 6). The NHL repeat is associated with RING finger and B-box motifs (Slack and Ruvkun 1998). However, searching the sequence database with AB-repeat regions in Rv0978 and Rv0980 in *M. tuberculosis* as query sequence does not identify the NHL repeat or vice-versa. We also further noted that in addition to tandem AB repeats, some proteins in *M. mazei* genome are associated with another repeat called the C-repeat comprising 85 amino acid residues. The C-repeat is referred as the PKD domain (accession number: PF00801) according to the Pfam protein family classification (Bateman *et al* 2000). The PKD domain was first identified in the Polycystic kidney disease protein present mostly in the extracellular parts of proteins, and is involved in adhesive protein–protein and

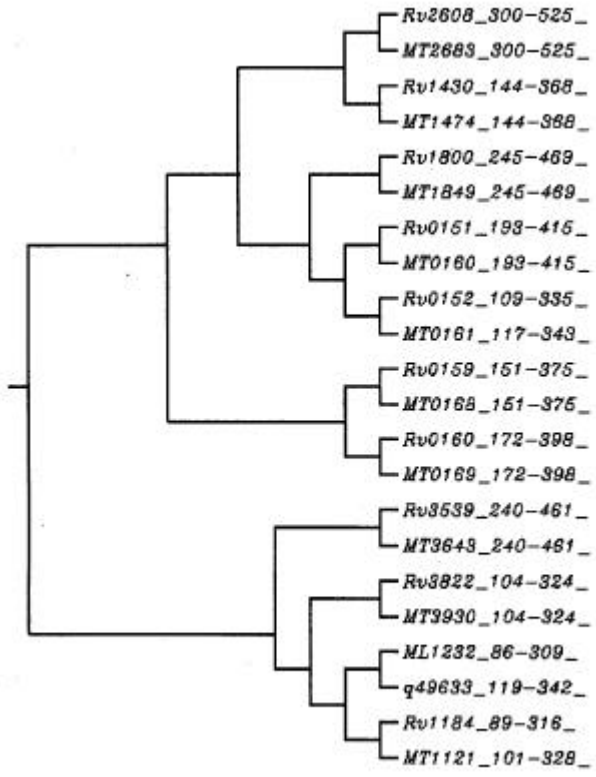

Figure 2b. Dendrogram showing relatedness of PE-PPE domain sequence produced using the CLUSTALW program available at the website (http://www.ebi.ac.uk).

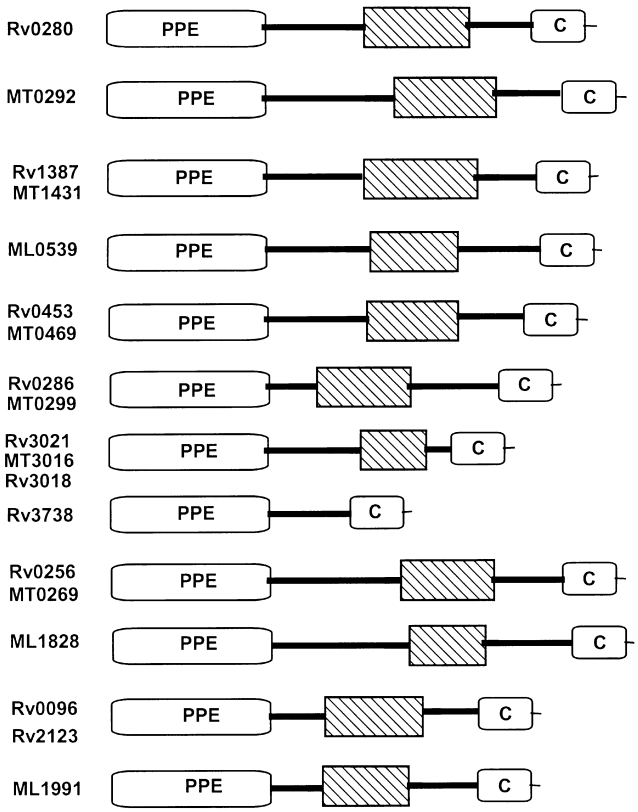

Figure 3. Schematic representation indicating the 44 amino acid-residue conserved C-terminal region (C), in some PPE proteins. Hydrophobic regions are indicated by hashed box and charged residues region flanking this by a thick line.

```
Rv0256(504-547)    QGAGTLGFAGTTHKASPGQVAGLITLPNDAFGGSPRTPMMPGTW (556)
MT0269(504-547)    QGAGTLGFAGTTHKASPGQVAGLITLPNDAFGGSPRTPMMPGTW (556)
Rv3018(376-418)    RGAGALGFVGTAGKESVGQPAGLTVLAD-EFGDGAPVPMLPGSW (434)
Rv3021(377-419)    RGAGALGFVGTAGKESVGQPAGLTVLAD-EFGDGAPVPMLPGSW (435)
MT3016(377-420)    RGAGALGFVGTAGKESVGQPAGLTVLAD-EFGDGAPVPMLPGSW (435)
Rv3738(262-305)    RGAGVLGFAGTAGKESVGRPAGLTTLAGGEFGGSPSVPMVPASW (315)
Rv0286(444-487)    RGAGTLGFAGTATKERRVRAVGLTALAGDEFGNGPRMPMVPGTW (513)
MT0299(444-487)    RGAGTLGFAGTATKERRVRAVGLTALAGDEFGNGPRMPMVPGTW (513)
Rv0280(477-520)    RGAGHLGFAGTARREAVADAAGMTTLAGDDFGDGPTTPMVPGSW (536)
MT0292(516-559)    RGAGHLGFAGTARREAVADAAGMTTLAGDDFGDGPTTPMVPGSW (575)
Rv1387(485-528)    CGAGPIGFAGTVRKEAVVKAAGLTTLAGDDFGGGPTMPMMPGTW (539)
MT1431(485-528)    CGAGPIGFAGTVRKEAVVKAAGLTTLAGDDFGGGPTMPMMPGTW (539)
Rv0453(466-509)    SGVGGLGFAGTASNETVAAPAGLTTLADDEFQCGPRMPMLPGAW (518)
MT0469(466-509)    SGVGGLGFAGTASNETVAAPAGLTTLADDEFQCGPRMPMLPGAW (518)
Rv2123(427-469)    RGLDALGFAGTIPKSAPGSATGLTHLGG-GFADVLSQPMLPHTW (473)
Rv0096(418-457)    -GAGTLGFAGTA-PTTSGAAAGMVQLS--SHSTSTTVPLLPTTW (463)
#ML1991(421-457)   HSAGPIGFAGTV-PTTASTPTGMAA------STSSSVPLLPTTW (468)
#ML1828(507-550)   QGAGTLGLAGTARQASSGPATGLVTLSNGTFSDSPRAPMLPRTW (572)
#ML0539(484-527)   CGAGSFGFAGTVSKEAVVEVAGLTTLAGDDFGGGPTMPMVPGTW (538)
^q9ans8 (95-137)   SGGSIQGFAGTLPKSG-VRASGLNTLAGNESGDGARIPMLPESW (140)
                    . .  *:.**         *:             *::* :*
```

Figure 4. Multiple sequence alignment corresponding to the 44 amino acid-residue C-terminal region in PPE proteins. The start and end residue corresponding to the above region is indicated in brackets and the total length of the protein is indicated at the end of each sequence. (#), Indicates PPE proteins in *M. leprae* genome; (^), indicates PPE protein (fragment) in *M. avium* genome; (*) indicates amino acid identities; dots (**.**) indicate conservative substitutions; and colon (:) indicates amino acid identities in majority of the sequences.

Table 1. Proteins containing the AB repeat.

SWALL ID.	Organism	Protein description	No. of AB repeats
Q50246	*Methanosarcina mazei*-A	GTG start codon	4
Q977V5	*Methanosarcina mazei*-A	Surface antigen	13
Q50245	*Methanosarcina mazei*-A	ORF492	7
TA1322	*Thermoplasma acidophilum*-A	Surface antigen genes 374	12
ST2622	*Sulfolobus tokodaii*-A	Hypothetical protein ST2622	2
ST1748	*Sulfolobus tokodaii*-A	Hypothetical protein ST1748	7
ST1138	*Sulfolobus tokodaii*-A	Hypothetical protein ST1138	1
Rv1057	*Mycobacterium tuberculosis* H37Rv-B	Hypothetical 40·7 kDa protein	6
MT1087	*Mycobacterium tuberculosis* CDC1551-B	Surface antigen, putative	6
RSC0127	*Ralstonia solanacearum*-B	Putative hemagglutinin-related protein	6
Rv0980	*Mycobacterium tuberculosis* H37Rv-B	PGRS-family protein	3
MT1008	*Mycobacterium tuberculosis* CDC1551-B	PE_PGRS family protein	3
Rv0978	*Mycobacterium tuberculosis* H37Rv-B	PGRS-family protein	2
MT1006	*Mycobacterium tuberculosis* CDC1551-B	PE_PGRS family protein	2
CAC2535	*Clostridium acetobutylicum*-B	Predicted protein of beta-propeller fold	6
O30796	*Methylobacterium extorquens*-B	MXAE	4
ALR1386	*Anabaena* sp-B	ALR1386	6
Q9RJL7	*Streptomyces coelicolor*-B	Putative lipoprotein	5
RB0179	*Rhizobium meliloti*-B	Hypothetical protein SMB20179	6
Q9RAP8	*Leptospira interrogans serovar ictero-haemorrhagiae*-B	ORFC protein	5
P71003	*Bacillus subtilis*-B	Hypothetical 49·4 kDa protein	5
ALL2941	*Anabaena* sp. (strain PCC 7120)-B	ALL2941 protein	4
P71004	*Bacillus subtilis*-B	Hypothetical 49·6 kDa protein	2
LIN0567	*Listeria innocua*.-B	LIN0567 protein	4
LMO0558	*Listeria monocytogenes*-B	LMO0558 protein	4
Q93TD8	*Pseudomonas syringae* pv. *Maculicola*-B	Hypothetical 37·7 kDa protein	7
MLR3473	*Rhizobium loti*-B	MLR3473 protein	6
O86549	*Streptomyces coelicolor*-B	Hypothetical 42·2 kDa protein	2
DOE_1063	*Rhodobacter sphaeroides*-B		7
DOE_915	*Nitrosomonas europaea*-B		4
DOE_1076	*Rhodopseudomonas palustris*-B		7
DOE_985	*Cytophaga hutchinsonii*-B		8

'A' and 'B' indicate the phylogeny, *Archaea* and *Bacteria*, respectively.

protein-carbohydrate interactions with other proteins and is predicted to contain an Ig-like fold (Bycroft *et al* 1999). The distribution of AB and C repeats in various proteins are shown in figure 7. This combination of AB and C tandem repeats has also been observed in some cell-wall associated proteins in Gram-positive bacteria (Kehoe *et al* 1994).

We further observed that some proteins containing the AB repeats also contain a Gly/Thr/Ser/Asn-rich segments of variable amino acid sequence length (not shown in figure 7) in the region preceding the tandem AB repeat. The Gly/Thr/Ser/Asn-rich segment has also been observed in the C-terminal region of AmyB, PglA, XynA (glycosyl hydrolases) from *Thermoanaerobacterium thermosulfurigenes* EM1 (Matuschek 1996) which are characterized by surface layer homology (SLH) domains (Pfam accession number: PF00395) comprising 50–60 amino acid residues. Several bacterial S-layer proteins, extracellular enzymes and non-catalytic cell envelope proteins contain SLH domains which are predicted to possibly be involved in the attachment of proteins to the cell wall (Lupas *et al* 1994; Lemaire *et al* 1995; Matuschek *et al* 1996). Surface antigens with tandem repeats have been observed in prokaryotes and eukaryotes, and are believed to play a role in the pathogen's evasion of the hosts' immune defenses by generating antigenic variation representing a mechanism for the organism's ability to survive environmental changes (Dybvig 1993). The antigenic mosaic of individual species is complex (Conway de Macario and Macario 1986). The mosaic may comprise distinct antigens which are typical to a particular species and antigens which are common to different species.

Cole *et al* (1998) proposed that members of the PE, PPE protein family in *M. tuberculosis* genome might play antigenic role. Our analyses identifying tandem AB repeats associated with surface antigen proteins provides further evidence to their proposal for the PE proteins.

```
Secondary          eee    eeee          eeeeee     eeeeee
structure           βI     βII           βIII       βIV

Rv1057(19-61)     FEFGTAPGSAVVKI-PVQGGPIGGIAISRDG-SLLVVTNNGTDTV (393)
MT1087(86-128)    FEFGTAPGSAVVKI-PVQGGPIGGIAISRDG-SLLVVTNNGTDTV (460)
doe_985(405-447)  TMIDVASNSRLGLL-DSTGLAPQGLVINADG-SRLFVHNFLTRTV (827)
q9rap8(327-368)   YVIDTTTDTVKEF--WEAGNQPTGLDVSPDN-RYLVISDFLDHQI (376)
p71003(182-223)   SVIDTNTDTVVTT--IALPYNPAGIEITPDK-SAVFVLHPNNNVI (451)
TA1322(425-466)   SVLDTQTQTVLRN--IAVGEGPAGIVVNPSN-GYVYSINQLSDDV (680)
ST1748(311-352)   SVINPLTNQVIAN--ITVGNCPTGIVYDPSN-GYIYVTNSLSGSI (403)
ST2622(100-141)   SVINPQTNTLVKT--IYVGGSPQSVVVDLKN-GYLYVVDNGIEIF (702)
ALL2941(401-442)  SVVDINQNQEIKR--IKLGPYPRGIVVDPAS-ETAYVAVMGSYNI (572)
p71004(219-266)   SVIDNKTLTIINT--ITVGGGPRKIEFDPTD-EFAYVMAAGSIYV (458)
ST1138(241-282)   VVVFNSKGGLVEE--VTVGSDPNGIIYDPQN-HYIYVINYGSSTI (328)
MT1008(377-417)   SVIDPNTNTVTGS--IPVGTGAYGVAVNPGG--NIYVTNQFSNTV (476)
Rv0980(358-398)   SVIDPNTNTVTGS--IPVGTGAYGVAVNPGG--NIYVTNQFSNTV (457)
MT1006(289-331)   SVIDPTTNTVTGSP-ITVGTAPTGVAVNPVT-GEVYVTNFAGDTV (335)
Rv0978(285-327)   SVIDPTTNTVTGSP-ITVGTAPTGVAVNPVT-GEVYVTNFAGDTV (331)
CAC2535(284-325)  TKIDIQNKTVAAT--ITVGKGAHGVVVSDDN-KFTYVTNMYDNTV (352)
q50246(30-71)     SVIDTSTNTVTAT--VPAGINPLGVAITPDG-RKAYVANRYSNNV (374)
q977v5(535-576)   SVIDTSTNTVTAT--VPAGINPLGVAITPDG-RKAYVANRYSNNV (1673)
o30796(166-207)   TVIDTGTLKAIAT--VPAGAMPYGVSVSPDG-ARVFVTNQHAGTV (281)
q50245(99-140)    SIIDTATNNVIAT--VPAGSSPQGVAVSPDG-KQVYVTNMASSTL (491)
q9rjl7(172-213)   EVVDTERREVTRV--VPVGRRPFDVDVSRDG-RQVYATNHDSFDV (392)
doe_1063(180-221) TVVDVETRTVLAE--VPVGVEPEGMGVSPDG-TIVVNTSETTNMA (382)
ALR1386(400-441)  AVVDLRSGRVSGR--IPTGWYPNSVSVSQDG-RKLFVVNAKSNSG (1001)
LIN0567(226-269)  KLNVIQTIASLPEG-FDKENKGSAIHISPDG-RFLYVSNRGQDAI (346)
LMO0558(227-269)  ALKVIQTIASLPEG-FDKENKGSAIHISPDG-RFLYVSNRGQDAI (346)
q93td8(166-208)   SVIDLDHQRPVAVVPG-FSQPRQGIRVSPDG-KTVYVTNFLGDKI (351)
RB0179(35-76)     TVLDSESWEVIATFPA--GNRPRGITISPDG-KELYVCASDDDTV (324)
RSC0127(362-401)  TTGALTAVGSPVAT----GQGPIPIAIHPSG-LFAYVGNVFDNTV (480)
MLR3473(307-350)  TVIDFATRSVVAQWPIPGGGSPDMGNVSADG-RQLWLSGRFDSEV (390)
O86549(316-358)   TGRASGAPVSVVGP--QDGTVAEGMLVSPDG-ERFLVAVHEPNAR (400)
doe_915(790-834)  VVIDTRTDSIVKSLPCDPGCHGANFGAKKGGGYYAYITNKFSNRL (912)
doe_1076(444-485) SKIDPATGAKTVV--AKDLKMPEGIALAPSG-KLIVAEVGAKRVV (544)

consensus/80%     slls.tstphhtp..h.hG..s.ulslssss.t.halss.hsptl
```

Figure 5. Multiple sequence alignment corresponding to representative AB-repeat sequences from various proteins in *Archaea* and *Bacteria* generated by the CLUSTALW program. Numbers in brackets indicate the start and end amino acid residues corresponding to the AB repeat. The total number of amino acid residues in each protein is given at the end of the sequence. The consensus labelling is same as in figure 2a. The secondary structure prediction indicated at the top was derived using the PHD program. Residues forming *b*-sheets are represented by 'e' and individual *b*-strands as b_I to b_{IV}.

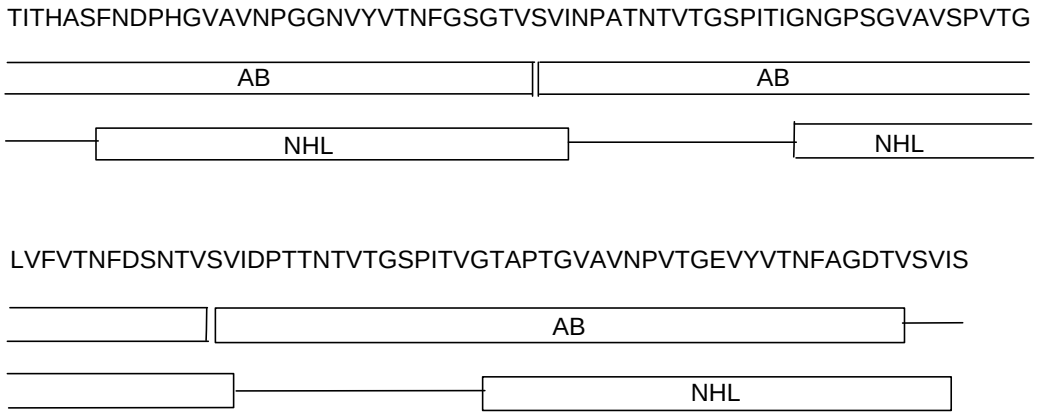

Figure 6. The C-terminal region corresponding to the protein Rv0978 as an example representing the NHL and AB repeats.

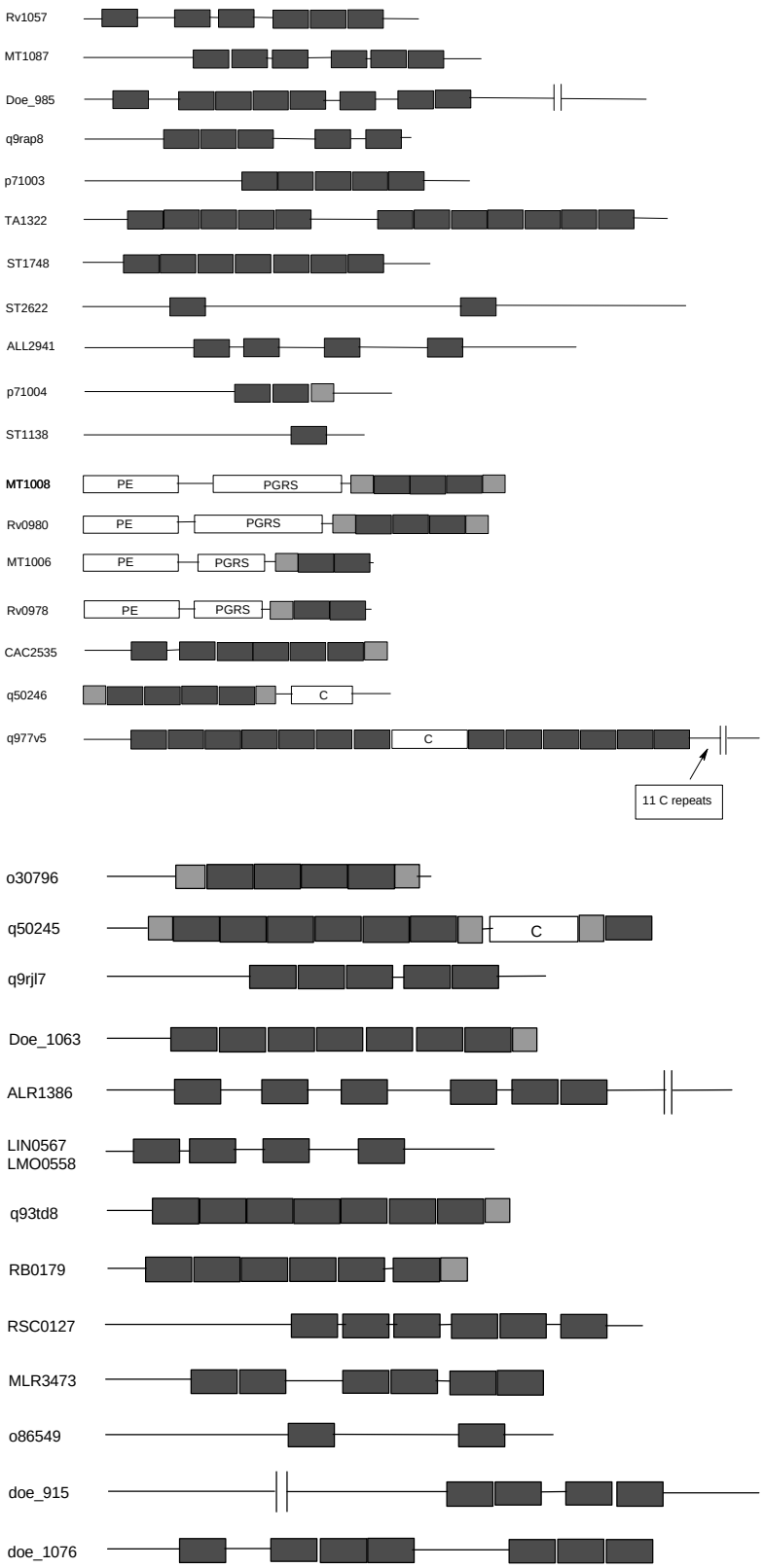

Figure 7. Distribution of AB repeats in various proteins, including the PE-PGRS proteins. Abbreviations: PE, PE domain; PGRS, PGRS domain; crossed boxes, AB repeats; hashed boxes, portion of AB repeat; C, C-repeat.

Recently, several proteins in *Methanosarcina acetivorans* genome associated as surface antigens were reported that contain the AB and C-repeats (Galagan *et al* 2002). The AB or C-repeats were not observed in *M. leprae* genome, suggesting that a different set of proteins may be recruited as surface antigens.

4. Conclusions

Some PPE proteins classified as 'unrelated' earlier have been identified to comprise a highly conserved 225 amino acid-residue region that is common to both, the PPE as well as the PE proteins and we refer to this as the 'PE-PPE' domain. Few other PPE proteins were observed to comprise highly conserved sequence motifs; GFxGT and PxxPxxW, in the C-terminus. Our observations may provide the basis to identify related members corresponding to these protein subfamilies that may be associated with a common function. The PPE proteins observed in *M. tuberculosis* but not in *M. leprae* genome may serve as potentially discriminating drug targets. Tandem AB repeat sequences first identified in some proteins of *M. mazei* genome and known to function as a surface antigen were identified in PE proteins in an attempt to characterize the variable C-terminal regions in this protein subfamily in the *M. tuberculosis* genome (H37Rv strain). Our analysis identifies PE proteins that may be associated as surface antigens in *M. tuberculosis* genome, and in several proteins of yet uncharacterized function from *Archae* and *Bacteria* genomes.

Acknowledgements

SA thanks the Council of Scientific and Industrial Research, New Delhi, for Junior Research Fellowship. LGP thanks Department of Science and Technology, New Delhi for Fast Track Young Scientist Fellowship.

References

Altschul S F, Gish W, Miller W, Myers E W and Lipman D J 1990 Basic local alignment search tool; *J. Mol. Biol.* **215** 403–410

Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W and Lipman D J 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs; *Nucleic Acids Res.* **25** 3389–3402

Bateman A, Birney E, Durbin R, Eddy S R, Howe K L and Sonnhammer E L 2000 The Pfam protein families database; *Nucleic Acids Res.* **28** 263–266

Bycroft M, Bateman A, Clarke J, Hamill S J, Sandford R, Thomas R L and Chothia C 1999 The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease; *EMBO J.* **18** 297–305

Cole S T *et al* 1998 Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence; *Nature* (*London*) **393** 537–544

Cole S T *et al* 2001 Massive gene decay in the leprosy bacillus; *Nature* (*London*) **409** 1007–1011

Conway de Macario E and Macario A J L 1986 Immunology of Archea-bacteria: identification, antigenic relationships and immunochemistry of surface structures; *Syst. Appl. Microbiol.* **7** 320–324

Dybvig K 1993 DNA rearrangements and phenotypic switching in prokaryotes; *Mol. Microbiol.* **10** 465–471

Fleischmann R D *et al* 2001 Whole genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains; http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/

Galagan J E *et al* 2002 The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity; *Genome Res.* **12** 532–542

Henning H, Fiona L and Rolf A 2003 CCP11 newsletter; http://wserv1.dl.ac.uk/CCP/CCP11/newsletter/vol2_3/sptr.html

Kehoe M 1994 Cell wall associated proteins in Gram-positive bacteria; in *Bacterial cell wall, new comprehensive biochemistry* (eds) J M Ghuysen and R Hakenbeck (New York: Elsevier) vol. 27, pp 217–261

Lupas A, Englehardt H, Peters J, Santarius U, Volker S and Baumeister W 1994 Domain structure of the *Acetogenium kivui* surface layer revealed by electron crystallography and sequence analysis; *J. Bacteriol.* **176** 1224–1233

Lemaire M, Ohayon H, Gounnon P, Fujino T and Beguin P 1995 OlpB. A new outer layer protein of *Clostridium thermocellum* and binding of its S-layer-like domains to components of the cell envelope; *J. Bacteriol.* **177** 2451–2459

Matuschek M, Sahm K, Zibat A and Bahl H 1996 Characterization of genes from *Thermoanaerobacterium thermosulfurigenes* EM1 that encode two glycosyl hydrolases with conserved S-layer-like domains; *Mol. Gen. Genet.* **252** 493–496

Mayerhofer L E, de Macario E C and Macario A J L 1995 Conservation and variability in Archaea: Protein antigens with tandem repeats encoded by a cluster of genes with common motifs in *Methanosarcina mazei* S-6; *Gene* **165** 87–91

Rost B, Sander C and Schneider R 1994 PHD – an automatic mail server for protein secondary structure prediction; *CABIOS* **10** 53–60

Schaftenaar G, Cuelenaere K, Noordik J H and Etzold T 1996 A Tcl-based SRS v. 4 interface; *Comput. Appl. Biosci.* **12** 151–155

Slack F J and Ruvkun G 1998 A novel repeat domain that is often associated with RING finger and B-box motifs; *Trends Biochem. Sci.* **23** 474–475

Thompson J D, Higgins D G and Gibson T J 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice; *Nucleic Acids Res.* **22** 4673–4680

Vega Lopez F, Brooks L A, Dockrell H M, De Smet K A, Thompson J K, Hussain R and Stoker N G 1993 Sequence and immunological characterization of a serine-rich antigen from *Mycobacterium leprae*; *Infect. Immun.* **61** 2145–2153

MS received 7 May 2002; accepted 17 December 2002

Corresponding editor: SEYED E HASNAIN