

Mycobacterial PE_PGRS Proteins Contain Calcium-Binding Motifs with Parallel β -roll Folds

Nandita Bachhawat^{1*#} and Balvinder Singh²

¹*G.N. Ramachandran Knowledge Center for Genome Informatics, Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research, Delhi 110007, India;* ²*Institute of Microbial Technology, Chandigarh 160036, India.*

[#]*Current address: Department of Bioinformatics, DAV College, Chandigarh 160012, India.*

The PE_PGRS family of proteins unique to mycobacteria is demonstrated to contain multiple calcium-binding and glycine-rich sequence motifs GGXGXD/NXUX. This sequence repeat constitutes a calcium-binding parallel β -roll or parallel β -helix structure and is found in RTX toxins secreted by many Gram-negative bacteria. It is predicted that the highly homologous PE_PGRS proteins containing multiple copies of the nona-peptide motif could fold into similar calcium-binding structures. The implication of the predicted calcium-binding property of PE_PGRS proteins in the light of macrophage-pathogen interaction and pathogenesis is presented.

Key words: *Mycobacterium tuberculosis*, virulence factors, PE_PGRS, calcium-binding motif, parallel β -roll fold

Introduction

The PE/PE_PGRS multigene family unique to mycobacteria accounts for about 5% of the *Mycobacterium tuberculosis* genome. The family includes 38 PE-encoding genes and 61 PE_PGRS-encoding genes scattered throughout the genome, characterized by a relatively conserved NH₂-terminus of \sim 110 amino acids (aa) followed by a COOH-terminal glycine-rich repeat region ranging from \sim 100 aa to over 500 aa in length (1, 2). Several PE_PGRS proteins have been demonstrated to be associated with the replication and survival of *M. tuberculosis* within macrophages of infected host tissues and are important for pathogenesis (3). One of the functions proposed for these proteins is that they are a source of antigenic variability to evade host immune responses (1). Other observations suggest the possibility that PE_PGRS proteins could either be cell surface constituents (adhesins) (4) that can influence bacterial cell structure (5) or could interfere with immune responses by inhibiting antigen processing (6). A recent study has shown that the evolution and expansion of the PE and PPE families is closely associated with the ESAT-6 (*esx*) gene cluster and has suggested a functional interdependence (7). As the ESAT-6 cluster encodes proteins involved in secretion and membrane pore formation, it is likely

that the ESAT-6 gene cluster-encoded proteins might also be required for secretion of some of the PE family proteins.

Despite these reports, the precise function and underlying mechanism of action still remain unknown for this large family of proteins. The aim of this study was therefore to investigate all of the 61 PE_PGRS proteins in the PE subfamily to predict whether they carry any distinct function in *M. tuberculosis*. Attempts were made to determine their underlying mechanism of action, and an explanation was sought for answering why there are so many different PE_PGRS proteins in *M. tuberculosis*.

Results

Attempts to identify similar proteins in databases to provide insights into their functions have been difficult owing to the repetitive nature of these proteins. Similarity searches of PE_PGRS proteins invariably identify only other PE_PGRS family members or glycine-rich proteins such as those found in highly elastic plant cell wall proteins (8). To search specifically for non-mycobacterial microbial homologues that have so far not been identified, we analyzed the microbial genome databases (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi) excluding

***Corresponding author.**

E-mail: nanditabachhawat45@gmail.com

Mycobacterium species (9). Using the widely investigated PE_PGRS protein Rv1818c as query and without masking the low complexity regions, one of the proteins revealed from the BLAST results was an RTX (repeat in toxin) toxin of *Magnetococcus* sp. MC-1 (COG2931), which is a related Ca²⁺-binding protein. The Rv1818c protein has 42% identity over 100% stretch of its PGRS domain to the glycine-rich domain of the COG2931 protein (data not shown). Although this domain corresponds to only a small region (5.8%) of the large RTX protein, the fact that some PE_PGRS proteins were reported as surface-exposed adhesin-like molecules (5, 10) prompted us to examine this similarity more carefully.

The family of RTX toxins includes haemolysin, cyclolysin, leukotoxin, and metallopeptidase, and these proteins seem to have two properties in common. First they bind calcium, and second they contain a multiple tandem repeat of nona-peptides, GGXGXN/NXUX, where X = any amino acid and U = unpolar/large hydrophobic residue at the C-terminal end of each protein (11, 12). It was subsequently found that these sequence repeats constitute a Ca²⁺-binding structure called a parallel β -helix or

parallel β -roll and might participate in host cell binding (13). To examine whether Rv1818c also contains this nona-peptide repeat corresponding to a Ca²⁺-binding motif, we searched for GGXGXN/NXUX as the calcium-binding motif and detected numerous such motifs in Rv1818c (Figure 1). Subsequently the presence of this motif was examined in all the 61 PE_PGRS proteins. Among the different PE_PGRS proteins grouped according to domains (not shown), 56 PE_PGRS proteins were found to contain this motif. The other five PE_PGRS proteins that lacked this motif were Rv0742, Rv0832, Rv0978c, Rv3652, and Rv3812. However, among the five proteins, Rv0832 (137 aa) may be frame-shifted in *M. tuberculosis* H37Rv to be fused with a protein of 749 aa encoded by Rv0833; Rv0978c is a protein of 331 aa that contains an unusually short (78 aa) PRGS motif; while Rv3652 (104 aa) may also be a frame-shifted PE_PGRS protein. All these proteins belong to both classical and non-classical types of PE_PGRS proteins. The maximum number of repeats was found in Rv3345c where 77 copies of this calcium-binding parallel β -helix or β -roll motif GGXGXN/NXUX were present (Table S1). A total of 911 such calcium-binding motifs, among

Rv1818c

MSFVVTIPEALAAVATDLGIGSTIGTANAAAAPVTTTVLAAADEVSAAMAALFSGHAQAYQALSAQAAL
 FHEQFVRALTAGAGSYAAAAEASAAPLEGVLDVINAPALALLGRPLIGNGANGAPGTGANGDGGILIGNG
 GAGGSGAAGMPGGNGGAAGLFGNGG**GGAGGNVAS**GTAGFGGAGGAGGLLYGAGGAGGAGGRAGGGVGGIG
 G**GGAGGNGGL**LFAGAGGAGVGGLAADAGD**GGAGGDGGL**FFGVGGAGG**GGTGTNVTG**G**GGAGGNGGL**LF
 GAGGV**GGVGDGVA**FLGTAPGGPGGAGGAGGLFGVGGAGGAGGIGLVNGGAGGSGSALLWGDGGAGGAG
 GVGSTTGGAG**GGAGGNAGL**LVGAGGAGGAGALGGGATGV**GGAGGNGGT**AGLLFGAGGAGGFVGGAGGAGGL
 GKGAGLIGDGGD**GGAGGNGTG**AKGGDGGAGGGAILVGN**GGNGGNAGS**GTPNGSAGTGGAGGLLKNGMNGL
 P

<u>Nine amino acid motifs</u>		<u>Inter-motif distance (amino acid)</u>
168	G G A G G N V A S 176	-
215	G G A G G N G G L 223	39
243	G G A G G D G G L 251	20
262	G G T G T N V T G 270	11
273	G G A G G N G G L 281	3
289	G G V G G D G V A 297	8
364	G G A G G N A G L 372	67
393	G G A G G N G G T 401	19
438	G G A G G N G T G 446	37

G G X G X D/N X U X = Consensus

X = arbitrary amino acid, U = unpolar/large hydrophobic residue

Fig. 1 The GGXGXN/NXUX calcium-binding motifs (bold) identified in Rv1818c and their alignments.

which 403 are GGXGXDXUX type and the remaining 508 are GGXGXNXUX type, were detected among all the PE_PGRS family members. The number of repetitive motifs varied according to the length of the open reading frames (ORFs), and the inter-motif distance was not fixed. For example, Rv1818c (498 aa) contains 9 motifs and the inter-motif distance varies from a minimum of 3 aa to as high as 67 aa (Figure 1). This is unlike RTX toxins where the calcium-binding motifs are not distributed so randomly. In this context, it is worth mentioning that the U residues of GGXGXDX/NXUX motifs in PE_PGRS proteins are not always unpolar/large hydrophobic residues in nature (Table S1). In more than 70% of these cases, U = G. In the remaining cases, U could be any arbitrary residue except for cysteine (data not shown). We also determined the secondary structure of all the 61 PE_PGRS proteins including Rv1818c to reaffirm the presence of β -strands intercepted by coils in the C-terminal PGRS domain, which is a characteristic of parallel β -roll structure (data not shown).

To identify structurally important folds of PE_PGRS proteins, we attempted to associate them with structurally similar bacterial protein(s). Employment of the 3D-Jury system (14) to predict a possible protein fold of known function, using the C-terminal PGRS domain (382 aa) of Rv1818c as query sequence, revealed a Ca^{2+} -binding fold with a statistically moderate score (Table S2). The highest fold recognition scores were compiled next for the PGRS domains of all 61 proteins (Table S2). Interestingly, ~70% of the total PE_PGRS proteins exhibited a common fold to the C-terminal β -roll Ca^{2+} -binding domain of *Serratia marcescens* metalloprotease [PDB ID: 1SRP (15) and 1SAT (16)] and the alkaline protease of *Pseudomonas aeruginosa* IFO3080 [PDB ID: 1AKL (17) and 1KAP (18)], all of them belong to the RTX family.

The Rv3344c protein with fold prediction data greater than the confidence threshold limit set by 3D-Jury system (Table S2) was used subsequently to generate an optimized 3D molecular model using the alkaline protease of *P. aeruginosa* (PDB ID: 1AKL)(17) as the template. It is important to mention here that although according to the GenBank report, the *Rv3344c* gene (Accession No. YP_177961) might be a gene fragment that should be in-frame with a following ORF (MTV016.45c), no frame-shift was found when checked in BAC and cosmid clones (as of GenBank latest update: 24-May-2007). The

MODELER program was used to create the model (18, 19). The predicted model (Figure 2) was found to adopt Ca^{2+} -binding parallel β -helix or parallel β -roll structures (13) located at the turn of coil regions. The models (18, 19) created for the PE_PGRS proteins with 1AKL/1KAP/1SAT/1SRP (PDB IDs) fold scores greater than 25.0 were also predicted to hold calcium ions in the individual glycine-rich nona-sequence motifs (data not shown).

Discussion

Our results presented here suggest that the highly homologous PGRS domain of the majority of the PE subfamily proteins have calcium-binding motifs and are therefore likely to be calcium-binding proteins. Calcium-dependent adhesins are known to exist in the soil bacterium *Rhizobium* species where their attachment to the developing root hairs of leguminous plants is considered to be the first step in the host-specific infection process that leads to a nitrogen-fixing symbiosis (20). Analogous to the *Rhizobium* nodulation gene *nodO* that encodes a Ca^{2+} -binding protein involved in interactions with plant root cells in a Ca^{2+} -dependent way, it is possible that a similar Ca^{2+} -dependent PE_PGRS-mediated interaction may exist in *M. tuberculosis* with host cells (4, 10).

The earliest interactions of *M. tuberculosis* with macrophages are known to result in a number of alterations in Ca^{2+} signaling events critical for phagosome maturation (21–23). In macrophages, activation of cytosolic Ca^{2+} -regulated enzyme Ca^{2+} /calmodulin-dependent protein kinase II (CaMKII) is essential for the phagosome-lysosome fusion (24). Other studies have demonstrated that *M. tuberculosis* blocks this pathway via inhibition of sphingosine kinase, a macrophage enzyme that increases cytosolic Ca^{2+} levels (22, 25). Furthermore, the maturation and the acidification of mycobacterial phagosome could be restored by artificially raising the cytosolic Ca^{2+} levels using Ca^{2+} ionophores (21) or receptor stimulation with ATP (26). This has suggested that *M. tuberculosis* depletes internal Ca^{2+} stores in infected human macrophages (21, 26).

In the light of these observations, a role for Ca^{2+} -binding PE_PGRS proteins could be envisaged as follows. The initial non-specific attachment of *M. tuberculosis* to the host alveolar macrophages via Ca^{2+} -dependent PE_PGRS proteins will cause a sudden dip in calcium concentration at the focal point of host

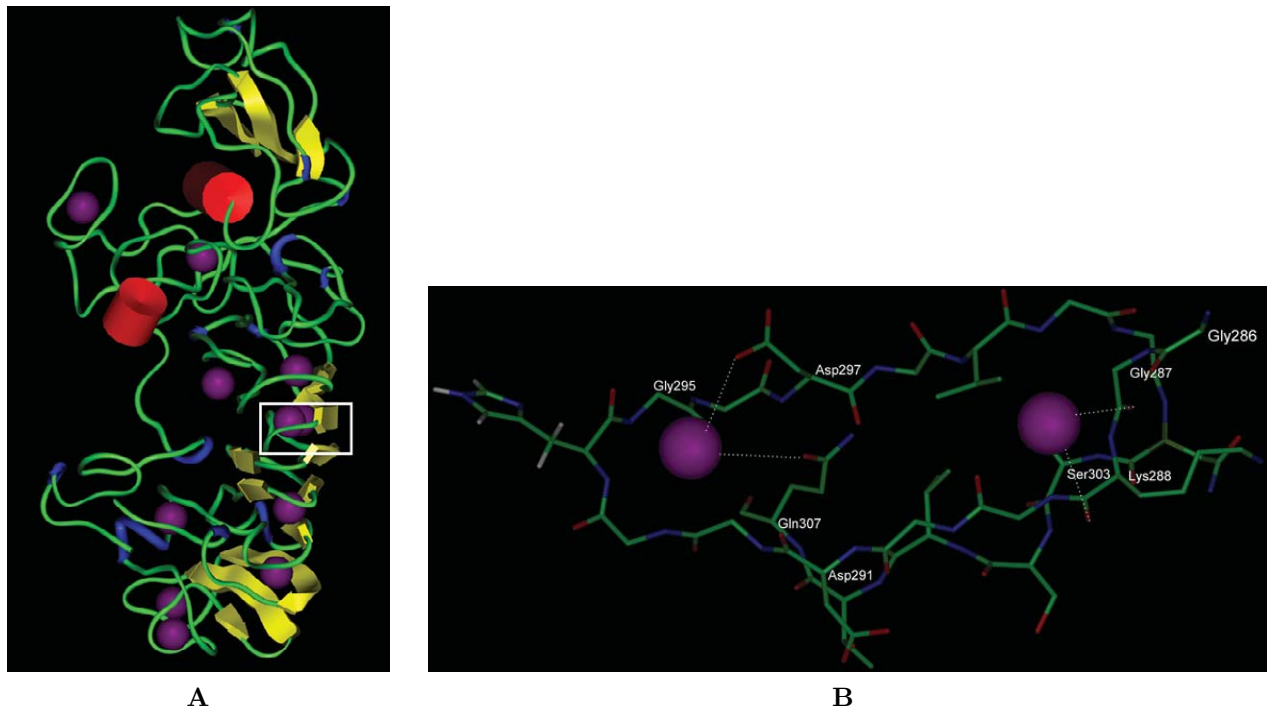


Fig. 2 The model of parallel β -roll structure of PGRS domain of Rv3344c with potential calciums interacting with glycine-rich nona-peptide motifs. **A.** Secondary structure rendering by InsightII software (Purple: calciums; Red: helices; Yellow: β -strands; Blue: turns; Green: random coils). The calcium-complexed structure was stabilized by energy minimization using InsightII. **B.** Magnified top view (of the box marked in Panel **A**) of one complete ring (aa 286–307) containing overlapping motifs (aa 286–294 and aa 292–300) interacting with calcium. In one motif, it is interacted via the side chain of Asp297, Gln307, and Gly295. In another motif, the side chain of Gly287, Lys288, and Ser303 (from below) holds onto the second calcium in the ring.

pathogen interaction (27, 28). Such an event would lead to a fall in the cytosolic Ca^{2+} levels in macrophages, ultimately preventing phagolysosome fusion. We thus postulate that the initial host-pathogen interactions could play a very crucial role to the sensing and establishment of the bacilli's intracellular pathogenesis.

The generalized calcium-dependent adhesion ability of PE_PGRS proteins does not rule out the possibility that these proteins have additional functions. Existence of “non-classical” type of PE_PGRS proteins (data not shown) with protein folds other than “conserved parallel β -roll calcium-binding folds” (Table S2) indicates that some PE_PGRS proteins might have yet undiscovered additional virulence-related functions that help the bacilli to survive in infected host (6).

Additionally, it will also be of worth to see if these initial PE subfamily protein(s)-host cell interactions can promote membrane cholesterol accumulation at the site of mycobacterial entry (29) that might even-

tually modulate the membrane depolarization event needed for the entry of external Ca^{2+} .

Materials and Methods

The genome information of *M. tuberculosis* strain H37Rv was retrieved from NCBI genome database as Accession NC_000962 (1). Subsequently, the re-annotated *M. tuberculosis* genome sequence was consulted from Camus *et al* (30). An indigenously developed algorithm was written in C++ language to extract all the 61 PE_PGRS ORF sequences.

To search specifically for non-mycobacterial microbial homologues of PE_PGRS proteins, we analyzed the microbial genome databases (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi) excluding *Mycobacterium* species (9). Classification of all the 61 PE_PGRS proteins from *M. tuberculosis* H37Rv based on domain patterns was performed using the PROSITE database (31) at <http://au.expasy.org/prosite/>. Fold recognition data for these proteins

were compiled using the 3D-Jury system (14) available via the Structure Prediction Meta Server (<http://meta.bioinfo.pl/>) to predict a possible protein fold of known function in each of the PE_PGRS proteins. No predictions were made for Rv0832 and Rv3652. The model of parallel β -roll structure of PGRS domain of Rv3344c with potential calciums interacting with glycine-rich nona-peptide motifs was generated by InsightII software (Accelrys Inc., San Diego, USA). The calcium-complex structure was stabilized by energy minimization using the same software.

Acknowledgements

NB thanks Prof. Samir K. Brahmachari for his encouragement and support during the course of this work, and also thanks Davinder Kohli for all his help. NB is a recipient of the IGIB Knowledge Center (CSIR) Fellowship.

Authors' contributions

NB conceived and supervised the study, collected and analyzed the data, and prepared the manuscript. BS collected the data relating to the fold, created the models, and assisted in manuscript preparation. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

- Cole, S.T., *et al.* 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537-544.
- Poulet, S. and Cole, S.T. 1995. Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Arch. Microbiol.* 163: 87-95.
- Ramakrishnan, L., *et al.* 2000. Granuloma-specific expression of *Mycobacterium* virulence proteins from the glycine-rich PE-PGRS family. *Science* 288: 1436-1439.
- Brennan, M.J., *et al.* 2001. Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect. Immun.* 69: 7326-7333.
- Delogu, G., *et al.* 2004. Rv1818c-encoded PE_PGRS protein of *Mycobacterium tuberculosis* is surface exposed and influences bacterial cell structure. *Mol. Microbiol.* 52: 725-733.
- Brennan, M.J. and Delogu, G. 2002. The PE multi-gene family: a 'molecular mantra' for mycobacteria. *Trends Microbiol.* 10: 246-249.
- Gey van Pittius, N.C., *et al.* 2006. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (*esx*) gene cluster regions. *BMC Evol. Biol.* 6: 95.
- Matsui, M., *et al.* 1994. Novel N-terminal sequence of a glycine-rich protein in the aleurone layer of soybean seeds. *Biosci. Biotechnol. Biochem.* 58: 1920-1922.
- Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
- Sachdeva, G., *et al.* 2005. SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* 21: 483-491.
- Coote, J.G. 1992. Structural and functional relationships among the RTX toxin determinants of Gram-negative bacteria. *FEMS Microbiol. Rev.* 8: 137-161.
- Baumann, U., *et al.* 1993. Three-dimensional structure of the alkaline protease of *Pseudomonas aeruginosa*: a two-domain protein with a calcium binding parallel beta roll motif. *EMBO J.* 12: 3357-3364.
- Lilie, H., *et al.* 2000. Folding of a synthetic parallel beta-roll protein. *FEBS Lett.* 470: 173-177.
- Ginalski, K., *et al.* 2003. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19: 1015-1018.
- Hamada, K., *et al.* 1996. Crystal structure of *Serratia* protease, a zinc-dependent proteinase from *Serratia* sp. E-15, containing a beta-sheet coil motif at 2.0 Å resolution. *J. Biochem.* 119: 844-851.
- Baumann, U. 1994. Crystal structure of the 50 kDa metallo protease from *Serratia marcescens*. *J. Mol. Biol.* 242: 244-251.
- Miyatake, H., *et al.* 1995. Crystal structure of the unliganded alkaline protease from *Pseudomonas aeruginosa* IFO3080 and its conformational changes on ligand binding. *J. Biochem.* 118: 474-479.
- Marti-Renom, M.A., *et al.* 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29: 291-325.
- Sali, A., *et al.* 1995. Evaluation of comparative protein modeling by MODELLER. *Proteins* 23: 318-326.
- Smit, G., *et al.* 1987. Involvement of both cellulose fibrils and a Ca²⁺-dependent adhesin in the attachment of *Rhizobium leguminosarum* to pea root hair tips. *J. Bacteriol.* 169: 4294-4301.
- Malik, Z.A., *et al.* 2000. Inhibition of Ca²⁺ signaling by *Mycobacterium tuberculosis* is associated with

- reduced phagosome-lysosome fusion and increased survival within human macrophages. *J. Exp. Med.* 191: 287-302.
22. Kusner, D.J. 2005. Mechanisms of mycobacterial persistence in tuberculosis. *Clin. Immunol.* 114: 239-247.
23. Vergne, I., *et al.* 2004. Cell biology of *Mycobacterium tuberculosis* phagosome. *Annu. Rev. Cell Dev. Biol.* 20: 367-394.
24. Malik, Z.A., *et al.* 2001. *Mycobacterium tuberculosis* phagosomes exhibit altered calmodulin-dependent signal transduction: contribution to inhibition of phagosome-lysosome fusion and intracellular survival in human macrophages. *J. Immunol.* 166: 3392-3401.
25. Malik, Z.A., *et al.* 2003. Cutting edge: *Mycobacterium tuberculosis* blocks Ca^{2+} signaling and phagosome maturation in human macrophages via specific inhibition of sphingosine kinase. *J. Immunol.* 170: 2811-2815.
26. Stober, C.B., *et al.* 2001. ATP-mediated killing of *Mycobacterium bovis* bacille Calmette-Guérin within human macrophages is calcium dependent and associated with the acidification of mycobacteria-containing phagosomes. *J. Immunol.* 166: 6276-6286.
27. Wagner, D., *et al.* 2005. Elemental analysis of *Mycobacterium avium*-, *Mycobacterium tuberculosis*-, and *Mycobacterium smegmatis*-containing phagosomes indicates pathogen-induced microenvironments within the host cell's endosomal system. *J. Immunol.* 174: 1491-1500.
28. Wagner, D., *et al.* 2005. Changes of the phagosomal elemental concentrations by *Mycobacterium tuberculosis* Mramp. *Microbiology* 151: 323-332.
29. Gartfield, J. and Pieters, J. 2000. Essential role for cholesterol in entry of mycobacteria into macrophage. *Science* 288: 1647-1650.
30. Camus, J.C., *et al.* 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 148: 2967-2973.
31. Hulo, N., *et al.* 2006. The PROSITE database. *Nucleic Acids Res.* 34: D227-230.

Supporting Online Material

Tables S1 and S2

<http://www.imtech.res.in/bvs/PE-PGRS-Mtb/>