The internal gene duplication and interrupted coding sequences in the MmpL genes of Mycobacterium tuberculosis: Towards understanding the multidrug transport in an evolutionary per…

ARTICLE *in* INTERNATIONAL JOURNAL OF MEDICAL MICROBIOLOGY: IJMM · MARCH 2015

Impact Factor: 3.61 · DOI: 10.1016/j.ijmm.2015.03.005

2 AUTHORS:

Padmani Sandhu
Central University of Himachal Pradesh
7 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Yusuf Akhter
Central University of Himachal Pradesh
25 PUBLICATIONS **277** CITATIONS
SEE PROFILE

International Journal of Medical Microbiology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Medical Microbiology

journal homepage: www.elsevier.com/locate/ijmm

The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective

Padmani Sandhu, Yusuf Akhter*

Centre for Computational Biology and Bioinformatics, School of Life Sciences, Central University of Himachal Pradesh, Shahpur District, Kangra 176206, Himachal Pradesh, India

ARTICLE INFO

Article history:
Received 20 October 2014
Received in revised form
18 December 2014
Accepted 10 March 2015

Keywords:
MmpL genes
RND transporters
Mycobacterium tuberculosis
Reductive evolution
Multidrug resistance
Efflux pumps

ABSTRACT

The multidrug resistance has emerged as a major problem in the treatment of many of the infectious diseases. Tuberculosis (TB) is one of such disease caused by *Mycobacterium tuberculosis*. There is short term chemotherapy to treat the infection, but the main hurdle is the development of the resistance to antibiotics. This resistance is primarily due to the impermeable mycolic acid rich cell wall of the bacteria and other factors such as efflux of antibiotics from the bacterial cell. The MmpL (Mycobacterial Membrane Protein Large) proteins of mycobacteria are involved in the lipid transport and antibiotic efflux as indicated by the preliminary reports. We present here, comprehensive comparative sequence and structural analysis, which revealed topological signatures shared by the MmpL proteins and RND (Resistance Nodulation Division) multidrug efflux transporters. This provides evidence in support of the notion that they belong to the extended RND permeases superfamily. *In silico* modelled tertiary structures are in homology with an integral membrane component present in all of the RND efflux pumps. We document internal gene duplication and gene splitting events happened in the *MmpL* genes, which further elucidate the molecular functions of these putative transporters in an evolutionary perspective.

Introduction

Mycobacterium tuberculosis the causative bacteria of tuberculosis (TB) in humans, is calculated to infect over 9.0 million people worldwide, causes the mortality of nearly 1.5 million people in the year 2013, and is thus positioned as the leading bacterial infectious agent (WHO, 2014). Many of the *M. tuberculosis* strains are naturally resistant to most of the antibiotics commonly used against the bacterial infection, due to the slow uptake of the drugs across the highly impermeable lipid rich cell envelope (Brennan and Nikaido, 1995). Other factors contributing to this resistance are the enzymatic inactivation, target alteration and efflux mechanism (Schweizer, 2003). Efflux is conferred by many membrane transport systems, which have been demonstrated to play an important role in both the bacterial and eukaryotic resistance by expulsion of the antibiotic and antiseptic drugs (Levy, 1992).

These multidrug efflux systems are characteristically energy dependent and can be classified into two mechanistically distinct kinds, depending upon the types of energy source they use. Primary transporters that couple drug extrusion from the cells with ATP hydrolysis (e.g. ABC transporters) and the secondary transporters are powered by transmembrane electrochemical gradients of either protons or sodium ions (Lubelski et al., 2007). These include, the Multidrug and Toxic compound Extrusion (MATE) family (Kuroda and Tsuchiya, 2009), the Major Facilitator Superfamily (MFS) (Pao et al., 1998), the Small Multidrug Resistance (SMR) family (Chung and Saier, 2001), and the RND permeases superfamily (Tseng et al., 1999).

The RND pumps of Gram-negative bacteria display efflux of a wide spectrum of the lipophilic and amphiphilic substrates (Tseng et al., 1999). These are the secondary transporters in which, drug efflux is coupled with the proton (H^+) influx and are often referred as H^+ drug antiporters (Paulsen et al., 1996). Characterized proteins of the RND family are involved in the transport of a range of substrates across the membrane by using the transmembrane

Abbreviations: *Mtb/M. tuberculosis*, *Mycobacterium tuberculosis*; MmpL, Mycobacterial Membrane Protein Large; MDR, Multi Drug Resistance; RND, Resistance Nodulation Division; TM, Trans Membrane; TMD, Trans Membrane Domain; TB, Tuberculosis; PS, Partial Sequences; ICDSs, Interrupted Coding Sequences.

* Corresponding author. Tel.: +91 1892 237288x309/+91 1892 237288x310/91 1892 237288x311; fax: +91 1892 2372286.
E-mail addresses: yusuf@daad-alumni.de, yusuf@cuhimachal.ac.in, yusuf.akhter@gmail.com (Y. Akhter).

http://dx.doi.org/10.1016/j.ijmm.2015.03.005
1438-4221/

proton gradient (Paulsen et al., 1996). Various examples of the RND transporters in the Gram-negative bacteria have been reported, some of these include AcrAB-TolC in *Escherichia coli* (Tikhonova and Zgurskaya, 2004) and MexAB-OprM in *Pseudomonas aeruginosa* (Poole et al., 1993). There are reported evidences revealing the presence of antibiotic efflux pumps of ABC and SMR family in *Mtb* (Balganesh et al., 2012). Until now, no such data have been presented in details on the RND pumps. In RND pumps, there are 10–12 Trans Membrane Domains (TMDs) present, which seems to be required for the activity. The functional unit of efflux pump systems of the RND family in Gram-negative bacteria are organized in a complex three component structure, which traverse both the inner and the outer membranes of the bacteria. They are composed of an integral membrane transporter, embedded in the cytoplasmic membrane, an outer membrane channel, located in the outer membrane and a periplasmic accessory protein, which brings the other two components into contact (Lee et al., 2003; Murakami, 2008; Rana et al., 2014). Comparative genomics of RND family proteins within a species and among different bacterial species have shown high degrees of homology at both sequence and structural levels (Piddock, 2006).

One of the common characteristic of membrane transporters is the internal gene duplication; it is an evolutionary process, leading to the formation of a functional unit of the transport machinery. It provides stability to the encoded proteins, assistance in the attainment of correct protein folding, conformational changes and in acquiring structural symmetry (Choi et al., 2008). Presence of the partial diploid and diploid type duplication is reported in integral membrane component of ABC and MFS transporters (Shimizu et al., 2004) and sequence analysis of two other components of the RND efflux pumps, i.e. membrane fusion protein and outer membrane protein, also reveal the presence of exactly duplicated repeats in them (Johnson and Church, 1999).

One more phenomenon, playing role in the evolution of the pathogenicity in prokaryotes is the presence of Interrupted Coding Sequences (ICDSs). 1 to 5% of the ICDS's are present in each of the bacterial genome (Perrodou et al., 2006). ICDSs also have an important role in functioning of the genes as well as, may result in loss of function as a result of the pseudogenization. About 30% ICDS's were observed in the case of *M. leprae* resulted in genome decay and pseudogenization (Deshayes et al., 2008). This genome decay in pathogenic bacterial species is a usual process which is indispensable for the pathogens to be more adapted within the host cells. It results in reductive evolution, which means losing genes and hence the respective functions, which results in decrease in genome size, thus preserving only those traits which are essential for the survival of the pathogen inside the host. This reduction in genome size makes these bacterial parasites more pathogenic as they become strictly dependent on the host and cannot survive outside independently as evident in case of *M. leprae* (Cole et al., 2001). This kind of strict parasitism makes the pathogen more virulent so that they could invade the host cells more aggressively as that is the only strategy for their survival.

The genome of *M. tuberculosis* is revealing the presence of 14 putative transmembrane protein sequences annotated as MmpL (Mycobacterium Membrane Protein Large), due to their large size and their putative localization in the inner membrane. They are speculated as members of the RND permeases superfamily on the basis of their reported role in the fatty acid transport and shared topological features (Cole et al., 1998; Tekaia et al., 1999). One of the MmpL proteins involved in antibiotic efflux is MmpL7, which is reported to be related to the isoniazid efflux in *Mycobacterium smegmatis* (Pasca et al., 2005), that is a first line drug against TB. The MmpL7 is predicted to contain 12 TMDs (like other RND transporters) with two periplasmic loops between TMD1/TMD2 and between TMD7/TMD8 (Sandhu, 2011). In addition to the isoniazid efflux by MmpL7, other proteins of this family like MmpL3 reported to play a role in the mycolic acid transport and in the intracellular accumulation of TMM (Trehalose Monomycolate) as reported in the *M. smegmatis* in a comparative analysis between the wild type and the conditional mutant of *MmpL3* gene (Varela et al., 2012). MmpL3 is also reported as a cellular target for many antitubercular drugs such as pyrrole derivative BM212 (La Rosa et al., 2012), SQ109, tetrahydropyrazolo [1,5-*a*] pyrimidine-3-carboxamides (THPPs), benzimidazole C215, indolcarboxamides, *N*-benzyl-6′,7′-dihydrospiropiperidine-4,4′-thieno-3,2-cpyrans and adamantyl ureas (AUs) (Li et al., 2014). All of these inhibitors are reported to destroy the proton motive force, thus affecting the activity of MmpL3 protein, which is a proton driven antiporter efflux pump (Li et al., 2014).

Taking into consideration all these experimental evidences regarding their putative role in fatty acid transport and drug efflux, we document here extensive evidences derived through *in silico* analysis for the association of the MmpL proteins with the RND permeases superfamily on the basis of the topological features, amino acid sequence and functional domains they share in detail. We further report their phylogenetic classification into subfamilies of RND transporters. Gene splitting and the presence of internally duplicated repeats in *MmpL* genes were also observed, which give rise to generation of multiple transmembrane helices that form the channel of the transporter as well as provide special symmetry to the overall structure to serve its designated function.

Materials and methods

Transmembrane topology analysis

The transmembrane topology analysis was performed using following four programs: TMHMM 2.0 (Krogh et al., 2001), HMMTOP (Tusnády and Simon, 1998), DAS (Cserzö et al., 1997), and MEMSAT (Jones et al., 1994). The TMHMM and HMMTOP tools are based on construction of the Hidden Markov Model, which helped us to localize the transmembrane segments and to carry out the topology prediction of various structural parts of these proteins. Two other programs DAS and MEMSAT were also used in this analysis, for finding out the TMDs in integral parts of the proteins embedded in the membrane. DAS program is based on low stringency dot plot of the query sequence against a collection of the non-homologous membrane proteins using a previously derived scoring matrix and MEMSAT is a Support Vector Machine (SVM) based method. All the programs were used in default settings and require a query amino acid sequence in FASTA format.

Hydrophobicity profile construction

The hydrophobicity profile also gives topological information like knowledge of hydrophobic residues embedded in the lipid bilayer. The hydrophobicity profile can be constructed using Protscale (Gasteiger et al., 2005). Protscale include various scales for the hydropathy calculations. Here, the Kyte and Doolittle (1982) hydrophobicity scale was used for the determination of the hydropathy index. It resulted in a hydropathy plot which helped in calculation of hydrophobicity and hydrophillicity scores over the length of a peptide sequence.

Analysis of the RND drug transporter specific signatures in the MmpL proteins

To identify the drug transporter characteristics of MmpL protein sequences from *M. tuberculosis*, they were scanned for the presence of the transporter family signature motifs characterized by sequence of the strongly conserved residues. The sequences of the

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

RND family specific signature motifs are: motif A (G x s x v T v x F x x g t D x x x A q v q V q n k L q x A x p x L P x x V q x q g x x v x k), motif B (a l v l s a V F l P m a f f g G x t G x i y r q f s i T x v s A m a l S v x v a l t l t P A l c A), motif C (x x x G k x l x e A x x x a a x x R L R P l L M T s L a f i l G v l P l a i a t G x A G a), and motif D (S i N t l T l f g l v l a i G L l v D D A l V v V E N v e R v l a e) (Paulsen et al., 1996), where x indicates any amino acid residue, capital letters show amino acids most frequently observed in a single position in more than 70% of the considered transport proteins, and lowercase letters display amino acid residues occurring in more than 40% of the RND members analyzed. The MmpL proteins were also scanned for the ABC family specific nucleotide binding motif (LSGGQ) (Hewitt and Lehner, 2003) and the MFS family specific motif (RXGRR) (Henderson and Maiden, 1990), present between TMD2/TMD3 involved in interaction with the negative charges. The BLAST function of the Tuberculist database (Lew et al., 2011) of Pasteur Institute was used for finding signature motifs of the transporters submitting each motif as a query against the MmpL protein sequences of the *M. tuberculosis* and the search was carried out using default parameters.

Conserved motifs and domains analysis within the MmpL proteins

In order to find out characteristic signature motifs within MmpL proteins, they were analyzed using the Gibbs Motif Sampler (Lawrence et al., 1993). These motifs were aligned using the COBALT program (Papadopoulos and Agarwala, 2007) and then subjected to the WEBLOGO tool (Crooks et al., 2004) to prepare a sequence logo. The position of these motifs in MmpL proteins was visualised using EsPript 3.0 (Gouet et al., 1999) output, with multiple alignments of the MmpL sequences along with the overlaying consensus secondary structure.

A conserved domain search was performed using sequences of the MmpL proteins as a query against the Conserved Domain Database (CDD) (Marchler-Bauer et al., 2007), Conserved Domain Architecture Retrieval Tool (CDART) (Geer et al., 2002), Clusters of Orthologous Groups (COG) (Tatusov et al., 1997) and the Interproscan (Zdobnov and Apweiler, 2001). These are protein annotation tools, which annotate proteins on the basis of the multiple alignments of the functional domains and these information are linked to a number of other electronic resources, which helped in finding domains related to the query sequence based on the homology.

Phylogentic analysis

Phylogenetic analysis requires multiple alignments of sequences to construct the guided trees. Multiple alignments of MmpL sequences were obtained using ClustalW2 program (Larkin et al., 2007). Gap parameters like opening penalty and extension penalty were modified to optimize the alignments. The phylogenetic tree was constructed in MEGA6 (Tamura et al., 2013) and evolutionary history was inferred using the maximum likelihood method based on the Poisson correction model (Zuckerkandl and Pauling, 1965). The bootstrap consensus tree was obtained from the 100 replicates (Felsenstein, 1985) was taken to represent the evolutionary history of the taxa analyzed. Initial tree for the heuristic search was obtained by applying the neighbour-joining (Saitou and Nei, 1987) to a matrix of pairwise distances estimated using a JTT model.

Homology modelling of the MmpL proteins

The amino acid sequences of proteins were retrieved from NCBI (http://www.ncbi.nlm.nih.gov). The three dimensional structure of any of the MmpL proteins are not available in the Protein Data Bank (PDB), so homology modelling of the MmpL proteins were carried out to analyze their structures. All these structures were modelled using Phyre2 program (Kelley and Sternberg, 2009). This program uses, protein threading techniques for prediction of the correct template/fold using the PSI-BLAST (Altschul et al., 1997) and construct an HMM model. This HMM model is scanned against an HMM database of known protein structures. It also models those regions of protein which are not homologous to any known protein structure using a new *ab-initio* folding simulation. All the MmpL protein sequences were modelled as an independent query, except for the MmpL13a and MmpL13b, which were subjected to modelling after fusing their amino acid sequences together as explained in another section. Three dimensional models were generated using multiple templates having 13-20% identity with each of the MmpL protein sequence with 100% confidence level. The query coverage for each modelled sequence was 80% to 90%. PDB structures generated were checked for the stereochemical validity using RAMPAGE (Lovell et al., 2003). Energy minimization was performed by GROMACS (Van Der Spoel et al., 2005), using parameters as a distance-dependent dielectric constant $e = 1.0$, Gromos 53A6 force field, all of the hydrogen atoms were included during the calculations and the nsteps used during energy minimization were 1000. The quality of initial models was observed to be improved after energy minimization. On the basis of the oligomeric structure of their templates, all of the verified 3D structures were subjected to the oligomerization using Symmdock program (Schneidman-Duhovny et al., 2005) and homotrimers of all the MmpL proteins were obtained. Out of all the MmpL homotrimer structures obtained, only those were selected having physical conformation relevant to biological functions. These MmpL oligomers were then subjected to energy minimization using the same protocol as described above.

Detection of the internal gene duplication

In order to analyze the internal gene duplication event in the *MmpL* genes, the Partial Sequence (PS) libraries were constructed for each of the TM protein sequence, including one TMD to six TMDs as explained earlier (Shimizu et al., 2004). To identify whether or not two PSs are homologous in each of the TM protein sequence, we have compared non-overlapping PSs of the same size within a TM protein sequence by aligning both of the sequences with each other using "align two or more sequences" option of the PSI- BLAST (Altschul et al., 1997). To judge the presence of the internal repeats, alignment coverage of >50% with the significant levels of identity were considered as criteria for designating the PSs as an internal repeat.

Results

The MmpL proteins of M. tuberculosis share topological features in functional domains and sequence motifs with the RND permeases superfamily

We have analyzed topological similarities of the MmpL proteins of *M. tuberculosis* with the RND permeases superfamily members using four different methods (Table 1). In contrast to earlier reports all the MmpL proteins were found, not to have mandatory 12 TMDs and two large periplasmic loops (Tekaia et al., 1999) but the results of topology analysis verified that out of 14 MmpL proteins, MmpL1, MmpL2, MmpL3, MmpL4, MmpL5, MmpL7, MmpL8, MmpL9, MmpL10, MmpL11 and MmpL12 possess 11 or 12 transmembrane domains as identified with the help of various programs (Table 1). Each of the MmpL proteins has 10-12 TMDs and also contain 1-2 large hydrophilic periplasmic/cytoplasmic loops between TMD1/TMD2 and between TMD7/TMD8, whereas MmpL10 which has 11 TMDs have only one large periplasmic

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

Table 1
Transmembrane helices assignment using different topology tools.

S. no.	Gene name	Rv no.	TMHMM 2.0	HMMTOP	DAS	MEMSAT
1	MmpL1	Rv0402c	12	12	12	12
2	MmpL2	Rv0507	11	12	12	12
3	MmpL3	Rv0206c	11	12	10	12
4	MmpL4	Rv0450c	11	11	10	12
5	MmpL5	Rv0676c	12	11	11	12
6	MmpL6	Rv1557	5	7	5	5
7	MmpL7	Rv2942	12	12	12	12
8	MmpL8	Rv3823c	12	13	12	12
9	MmpL9	Rv2339	11	12	11	12
10	MmpL10	Rv1183	11	12	12	11
11	MmpL11	Rv0202c	12	11	11	12
12	MmpL12	Rv1522c	11	13	12	12
13	MmpL13a	Rv1145	4	3	4	4
14	MmpL13b	Rv1146	6	7	7	7

loop between TMD6/TMD7. In MmpL3, MmpL4, MmpL5, MmpL9 and MmpL11, second large loop is present between TMD6/TMD7 instead of between TMD7/TMD8. MmpL6, MmpL13a, MmpL13b proteins possess 5, 4 and 7 TMDs respectively, and a single large loop is present in case of MmpL13a between TMD1/TMD2 while in MmpL13b the loop exists between TMD2/TMD3, whereas no large periplasmic loop is present in case of MmpL6. The loops with amino acid residues length more than 150 and 300 could be considered as large periplasmic loops present between TMD1/TMD2 and between TMD7/TMD8 respectively as reported elsewhere (Jain and Cox, 2005) and same has been observed in case of MmpL proteins.

Hydropathy analysis using Kyte and Doolittle scale also strengthened the notion in support of this assumption that the MmpL proteins from *M. tuberculosis* have 12 or less TMDs similar to RND proteins. The graph (Supplementary Fig. 1) shows that MmpL1, MmpL2, MmpL3, MmpL4, MmpL5, MmpL7, MmpL8, MmpL9, MmpL10, MmpL11 and MmpL12, each has more than 10 numbers of peaks with the positive score ≥1.8 indicating the presence of transmembrane regions and two large peaks with long stretches of residues having −ve score indicates the presence of hydrophilic residues that represent the large periplasmic or cytoplasmic loops.

When *M. tuberculosis* ORF's were scanned for the presence of RND family specific transporter signature motifs, MmpL10 and MmpL12 showed similarity at C-terminal with motif D (S i N t l T l f g l v l a i G L l v D D A l V v V E N v e R v l a e); one of the characteristic motif reported in RND transporters (Paulsen et al., 1996). This motif shows a considerable homology, similarity (65% and 68%) and identity (43% and 36%) with MmpL10 and MmpL12 respectively (Supplementary Fig. 2). This signature motif is located in TMD4 in case of RND family proteins as reported earlier, but their function in transport mechanism is yet to be shown. However, all MmpL proteins do not share this signature motif equally and levels of identities are varied. No significant similarities were observed with the ABC and MFS family specific signature motifs.

MmpL proteins having 12 transmembrane helices and two large loops, when aligned they were found to have significant similarities with each other. Two conserved motifs were found in MmpL proteins with more than 11 TMDs. These motifs are located in TMD4 [s x l x I a A g T D Y a i F l i g R Y x E] and TMD8 [D l x I x x v x x i c l i f i I m l x x x R] (Fig. 1). These motifs have prevalent conservation of aspartate (D), phenylalanine (F), tyrosine (Y), arginine (R) and glutamate (E).

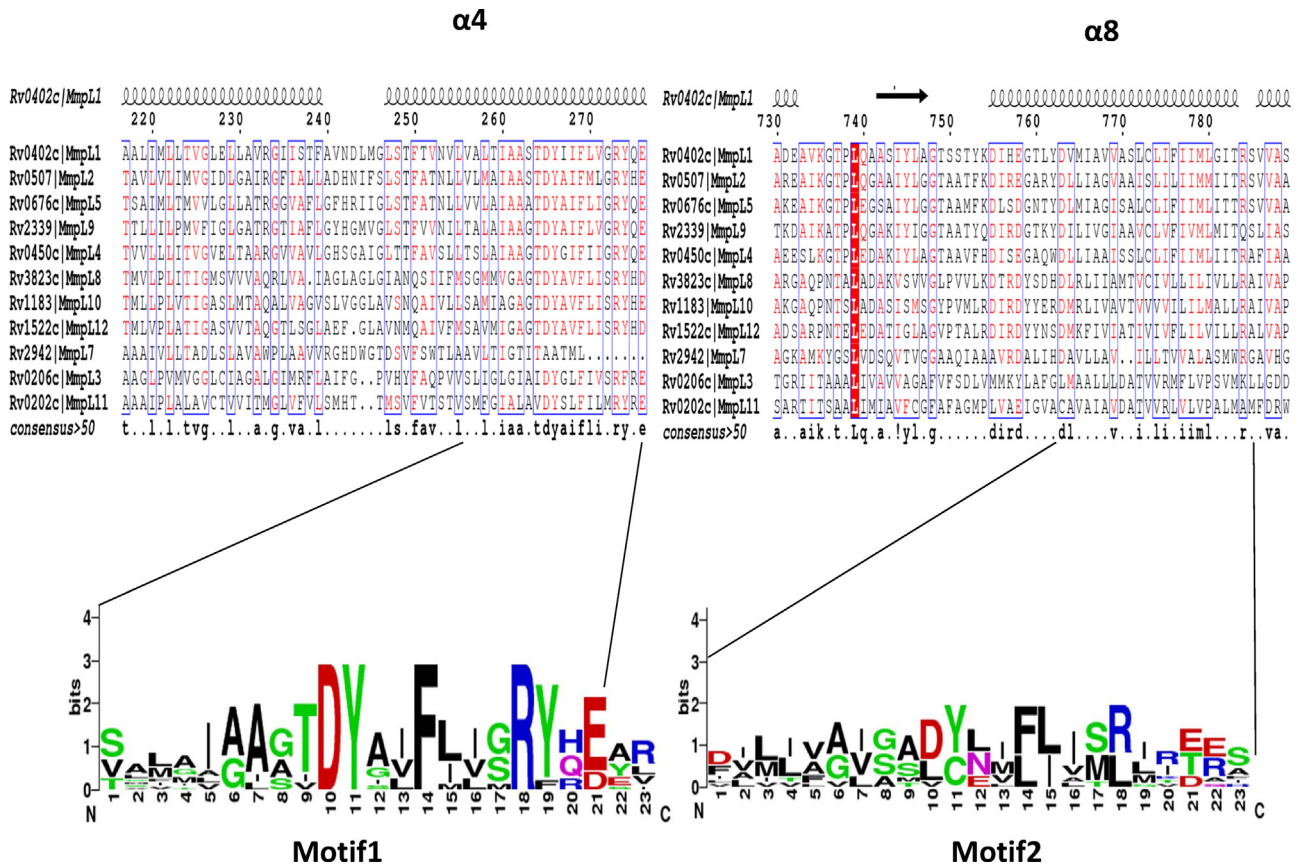

Fig. 1. *Conserved motifs in the MmpL proteins indicating functionally important amino acid residues*: The sequence logo of two motifs found in transmembrane helices of the MmpL proteins using the Gibbs sampling algorithm (Lawrence et al., 1993). The figure is showing position of the Motif1 and Motif2 in transmembrane region, i.e. the TMD2 and TMD8. This consensus sequence was generated using the WEBLOGO tool (Crooks et al., 2004) and indicates the prevalence of amino acids at the specific positions. The alignment was obtained using the MULTALIN program (Corpet, 1988) and location of transmembrane helices was visualized using ESPript 3.0 (Gouet et al., 1999).

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

Conserved protein domains are the recurring units in molecular evolution and appear in protein sequence as a conserved block of amino acid residues that may have distinct functions. These can fold, function and exist independent of the rest of protein structure. In Supplementary Table 1 we are showing the conserved domains present in different MmpL proteins. Acr_tran domain shared by Acr family proteins related to Acriflavin resistance (Ma et al., 1993), Act II domain present in proteins related to Actirhodin synthesis (Fernandez-Moreno et al., 1991), HpnN (Hopanoid) domain related to Hopanoid biosynthesis (Doughty et al., 2011), SecD domain (Matsuyama et al., 1993) of proteins related to Sec protein of Type II secretory pathway, IPR000731 corresponding to Sterol sensing domain (Millard et al., 2005), COG 2409 domain present in predicted drug exporters of RND superfamily, COG3696 domain present in putative silver efflux pumps, COG4258 and COG1033 are not assigned to any particular function but are shared by all transport proteins.

The MmpL proteins from M. tuberculosis could be classified in the subgroups of hydrophobe/amphiphile efflux (HAE) family as HAE2 and HAE3 of the RND transporters superfamily

The multiple sequence alignments (MSA) of all MmpL proteins with well-known RND proteins from other organisms show conservation at a greater extent near the C- than at N-terminal. It is evident from MSA that MmpL1, MmpL2, MmpL4, MmpL5, MmpL8, MmpL9, MmpL10, MmpL12 share greater similarity and conservation with the ActII protein and Acr family proteins, whereas MmpL3, MmpL7 and MmpL11 have conservation with the Sterol regulatory element binding protein, Hopanoid biosynthesis related proteins and SecD protein of type-II secretory pathway (Fig. 2a).

On the basis of alignment with RND proteins and sequence conservation MmpL proteins were found to be clustered with the proteins of HAE2 and HAE3 family, subfamilies of RND permeases superfamily. MmpL1, MmpL2, MmpL4, MmpL5, MmpL6, MmpL8, MmpL9, MmpL10 and MmpL12 clustered on one branch sharing sequence similarity with the members of HAE2 subfamily while MmpL3, MmpL7 and MmpL11 clustered with members of HAE3 subfamily (Fig. 2b). MmpL7 and MmpL8 were clustered into two different groups on the basis of their phylogeny. Interestingly, they are reported to have similar functions i.e. transportation of methyl branched fatty acid chains containing molecules (Cox et al., 1999; Converse et al., 2003).

In silico modelled three dimensional structures of the MmpL proteins have homologous structural architecture to the RND transporters

The homology modelling using a protein threading technique yielded three dimensional structures of the MmpL proteins showed resemblance with the structure of well-studied proteins of RND permeases superfamily from eubacteria.

However, the RMSD values between MmpL structures and template RND proteins were high due to only remote homology among them. From all the modelled and analyzed three dimensional structures, we present here MmpL5 trimer, comprises of 958 residue long protomer as a representative structure (Fig. 4a). The appearance of trimer is similar to that of oligomeric crystal structure of another RND transporter AcrB (PDB: 1IWG) with a threefold symmetry axis perpendicular to the transmembrane plane with periplasmic headpiece and a transmembrane region. Transmembrane region of oligomer consists of 36 alpha helices in which 12 transmembrane helices are contributed by each monomer unit. These helices are arranged together to form a channel which runs across the membrane. Alpha helices and beta strands extended in periplasm including large loops forming a hydrophilic head piece that interacts with the outer membrane channel in a typical fashion like an RND efflux pump. Periplasmic loops were thought to have a role in substrate binding as described in earlier studies of RND transporters from Gram-negative bacteria. Three alpha helices at the distal end of the channel, one from each monomer unit, form a triangular gate like opening. At the centre of the headpiece, when viewed from a proximal end, a pore is visible that runs through the length of the oligomer and ends at the base of the trimer (Fig. 4b).

MmpL13a and MmpL13b: A case of gene splitting with interrupted sequences in M. tuberculosis

The *MmpL13* gene from *M. tuberculosis* is consist of two different open reading frames *MmpL13a* and *MmpL13b*, both of them separately encode for the two different hypothetical proteins consisting of 4 and 6 TMDs respectively, whereas all of the other MmpL proteins have 11 or 12 TMDs and two large periplasmic loops. The MmpL13a protein has a single large periplasmic loop between TMD1/TMD2, whereas MmpL13b also have a single large periplasmic loop between TMD2/TMD3 (Fig. 3). When MmpL13a and MmpL13b were fused together to yield a fusion protein, henceforth named as *Mtb*MmpL13 and subjected to the TMHMM program for topology computation (Supplementary Fig. 3), it was observed that the *Mtb*MmpL13 consisted of 10 TMDs and two large periplasmic loops between TMD1/TMD2 and TMD6/TMD7. This appears to have an identical topology as the other MmpL proteins. When *Mtb*MmpL13, the fusion protein was subjected to the homology modelling it yielded a structure which is very similar to the CusA monomer from *E. coli* (PDB ID: 3K07). When we carried out comparative structural analysis of monomeric unit of the *Mtb*MmpL13 fusion protein, with the CusA, it showed obvious similarities in architecture, but the periplasmic part looks quite different (Fig. 3b). When compared for the RMSD values with the three dimensional models of the rest of the MmpL proteins, it showed a best RMSD value of 0.55 with the MmpL3 structure. When we looked at the individual gene structures of the *MmpL13a* and *MmpL13b* are appeared to be functional ORFs with start and stop codons. The ribosome binding sites like pyrimidine rich sequence GCACUUCCU is present immediately upstream to the start codon in the *MmpL13a* and in case of *MmpL13b*, this site is present at −9 bases upstream to start codon. This indicates that both of these genes may transcribe, and translate individually, irrespective of their ortholog from the *M. bovis*, which is a single gene (Fig. 3c). From these results it was observed that in the course of evolution, the *MmpL13* was a single unit which gave rise to the *MmpL13a* and *MmpL13b* and these two separate open reading frames of the *MmpL* genes might be a result of gene splitting and may have broken apart during the molecular evolution in the *M. tuberculosis* complex.

The internal gene duplication indicating presence of repeating domains in the MmpL proteins

The internal gene duplication is a common mechanism in the evolution of transmembrane alpha helices containing transporters in order to acquire a suitable symmetry to carry out its function. It results in a twofold symmetrical structure oriented opposite to each other. Internal symmetry is required to form a functional complex and to attain conformational changes required in case of proteins involved in the transport function as evident in case of AcrB protein of *E. coli* (Hennerdal et al., 2010). Some previous studies reported that this phenomenon occurred in the case of 12 TMD carrying transporters like MFS and ABC family (Shimizu et al., 2004). The MmpL protein sequences were divided into partial libraries, including only TMDs and analyzed for duplication patterns. Two types of duplication patterns were observed. The partial

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

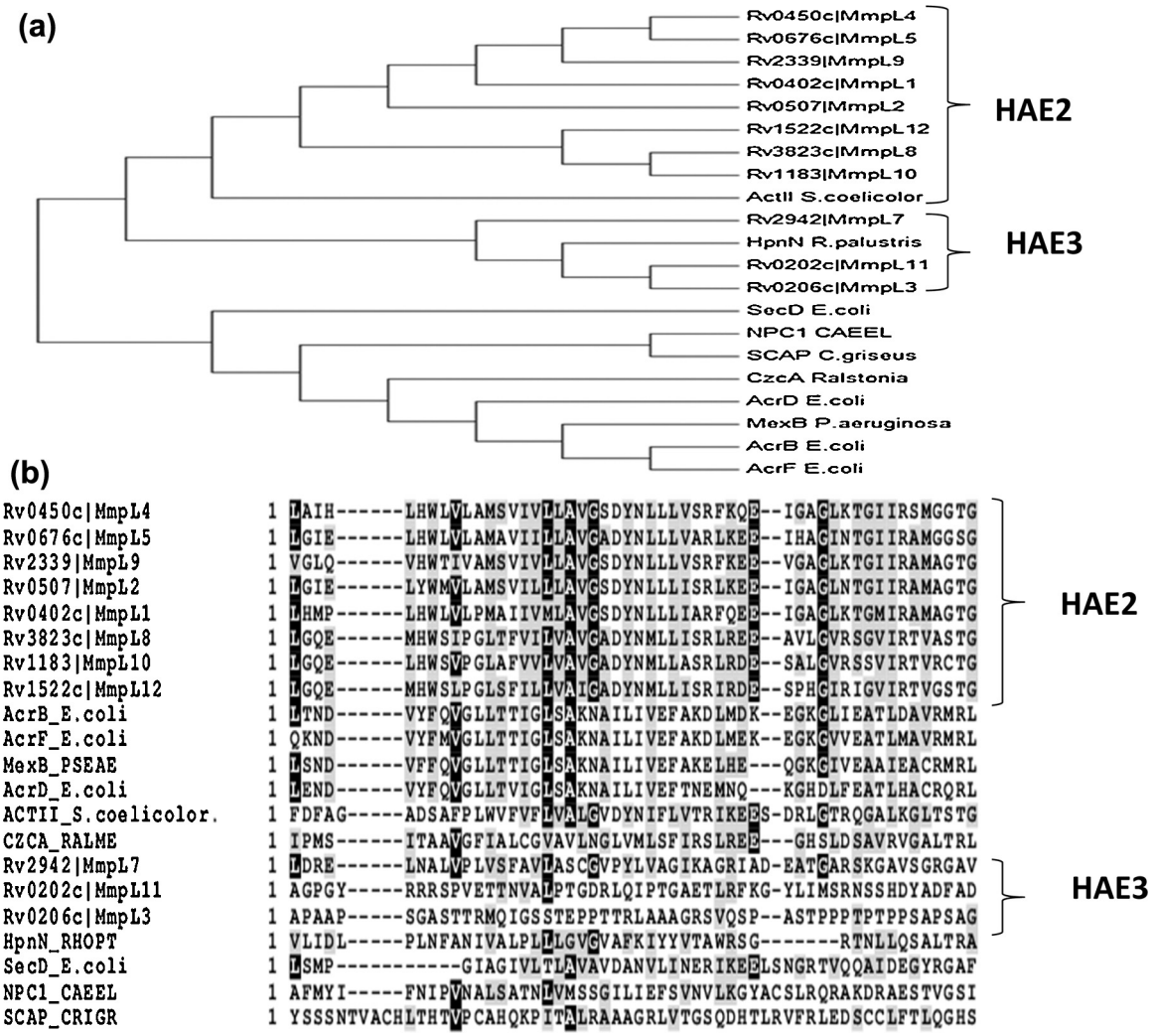

Fig. 2. *Phylogenetic classification of the MmpL proteins of M. tuberculosis*: (a) Phylogenetic tree prepared by MEGA6 (Tamura et al., 2013). 11 sequences related to the MmpL proteins and 10 sequences of the other well-known RND family proteins were subjected to the multiple alignment using ClustalW2 (Larkin et al., 2007). Initial tree(s) for the heuristic search were obtained by applying the neighbor-joining method to a matrix of pairwise distances estimated using a JTT model (Saitou and Nei, 1987). The bootstrap consensus tree inferred from the 100 replicates (Felsenstein, 1985) and evolutionary history was interpreted by the maximum likelihood method based on the Poisson correction model (Zuckerkandl and Pauling, 1965). Different MmpL members clustered into two groups, the HAE2 and HAE3 on the basis of sequence conservation. (b) Conservation of the MmpL protein sequences with the well-studied RND proteins from *Mtb* and other well-studied proteins are shown. Alignments were prepared using the ClustalW2 (Larkin et al., 2007) and box shade program was used to view the conservation patterns.

duplication, in which a part of the TMD encoding DNA sequence was duplicated and the diploid type, in which the whole of the gene was duplicated, resulted in the evolution of two similar halves in TM proteins encoding genes (Table 2). The diploid type duplication in which 6 TMDs duplicated in to 12 TMDs e.g. {1′-2′-3′-4′-5′-6′} to {1-2-3-4-5-6-7-8-9-10-11-12} was observed in MmpL11. Partial duplication was observed in MmpL 2, MmpL5, MmpL8, MmpL9, and MmpL12, where only 5 TMDs out of 7 TMDs were duplicated and gave rise to 12 TMDs system, i.e. {1-2′-3′-4′-5′-6′-7} to {1-2-3-4-5-6-7-8-9-10-11-12}. In addition to this in MmpL1, MmpL2, MmpL3 and MmpL12, out of the total TMDs only 3 or 4 TMDs were found to be duplicated. None of the duplication patterns were observed in MmpL6, MmpL10, MmpL13a and MmpL13b.

Discussion

The aim of the present study was to obtain evidence that may be used to classify *MmpL* genes of *M. tuberculosis* into membrane transporter families and to get an insight of the molecular evolution of these genes, which may be responsible for the putative drug transport. Out of these 14 MmpL proteins MmpL1, MmpL2, MmpL3, MmpL4, MmpL5, MmpL7, MmpL8, MmpL9, MmpL10, MmpL11, MmpL12 fulfil the criteria to have 11 or 12 TMDs and having two large loops, except for the MmpL10 which has a single loop. These features are characteristic of RND family proteins which have 12 TMDs, with two large loops between TMD1/TMD2, and between TMD7/TMD8 (Jain and Cox, 2005) and are reported to be associated with antibiotic and fatty acid transports (Tseng et al., 1999). However, in case of MmpL3, MmpL4, MmpL5, MmpL9 and MmpL11 the second large loop is present in between TMD6/TMD7.

According to the conserved domain studies, it is evident that the MmpL proteins possess AcrB, ActII, Sterol sensing and Hopanoid biosynthesis related domains. These domains signify similarity of the MmpL proteins with the well-known RND family members. ActII domain is preserved in all of the MmpL proteins and is a characteristic feature of Actirhodin cluster activating proteins from *Streptomyces coelicolor* involved in Actirhodin biosynthesis and transport (Fernandez-Moreno et al., 1991), similarity of *MmpL* genes with it indicates that the MmpL proteins may have some role in the transport of the Streptomycin or similar molecules, which is a well-known antitubercular aminoglycosidic drug. *S. coelicolor* bacterium is of the same taxonomic order (Actinomyceteales) as of the

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

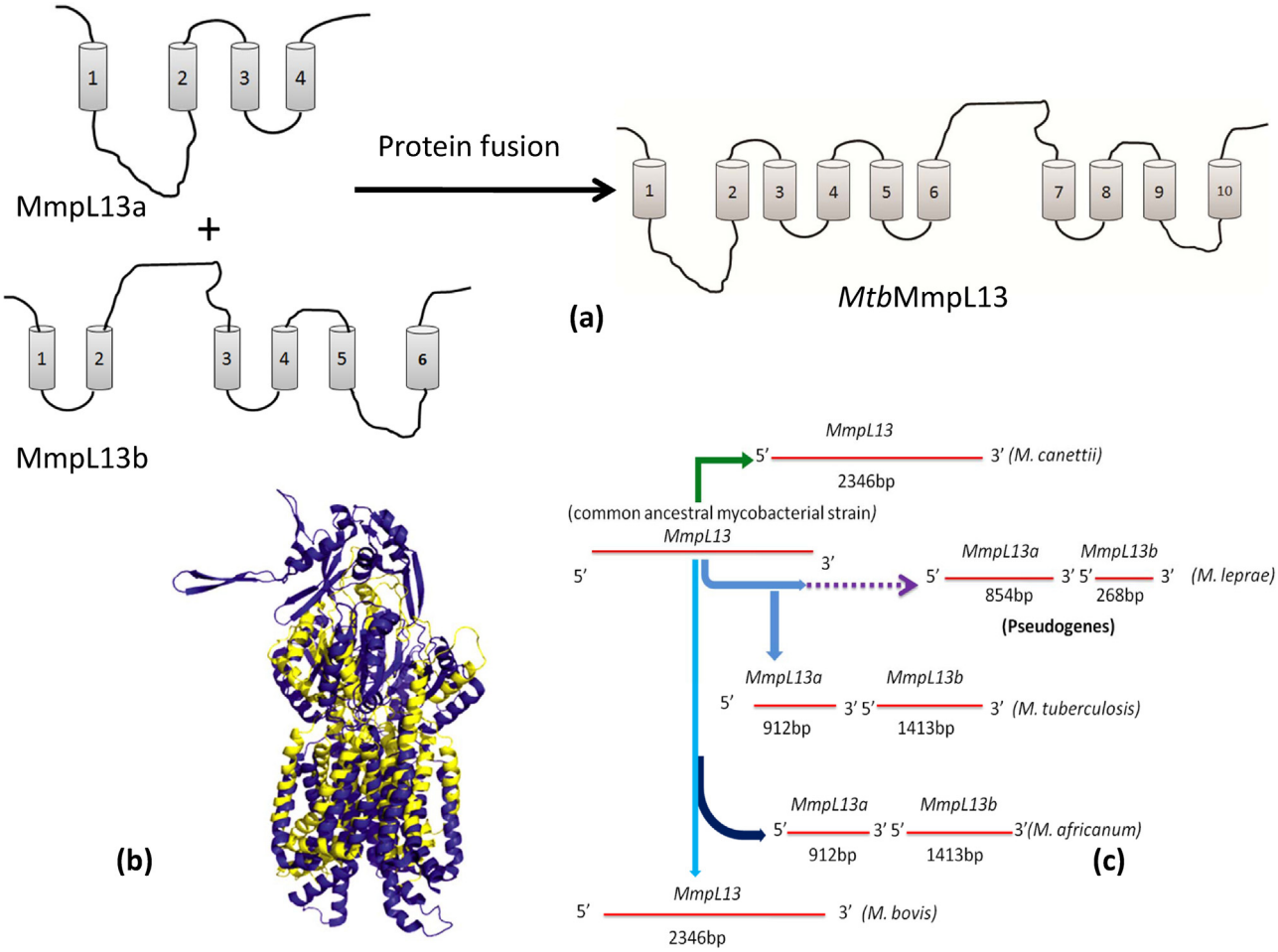

Fig. 3. *Reductive evolution of the Mmpl13 gene within M. tuberculosis complex: from intact to split genes:* (a) Topological representation of the MmpL13a and MmpL13b, respectively. *MmpL13a* and *MmpL13b*, the genes are present adjacent to each other in an operon but do not carry coding sequences for 10–12 TMDs and two large loops like the other *MmpL* genes. When their sequences were fused artificially in frame with codons and translated sequence was subsequently subjected to topology analysis they yielded 10 TMDs and two large loops. (b) Ribbon diagram of the modelled *Mtb*MmpL13, aligned with the tertiary structure of CusA, cation efflux pump from *E. coli* (PDB ID: 3K07) of *E. coli* (Long et al., 2010), a member of RND permeases superfamily. Structures were aligned in Pymol (DeLano, 2002) and both the proteins were represented using different colours. Blue coloured ribbon represents CusA monomer and yellow coloured ribbon represents the *Mtb*MmpL13 structure. It is evident from structure alignment that all of the TM helices are aligned correctly, but there is divergence in periplasmic part due to the difference in larger lengths of the loop regions of both the proteins. (c) The *MmpL13a* and *MmpL13b*, which are two separate genes in the case of *M. tuberculosis* and *M. africanum*, pseudogene in case of *M. leprae* and intact gene *MmpL13* in the case of *M. canettii* and *M. bovis*. For the interpretation of references to colours here, the audience is referred to the web version of the article. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

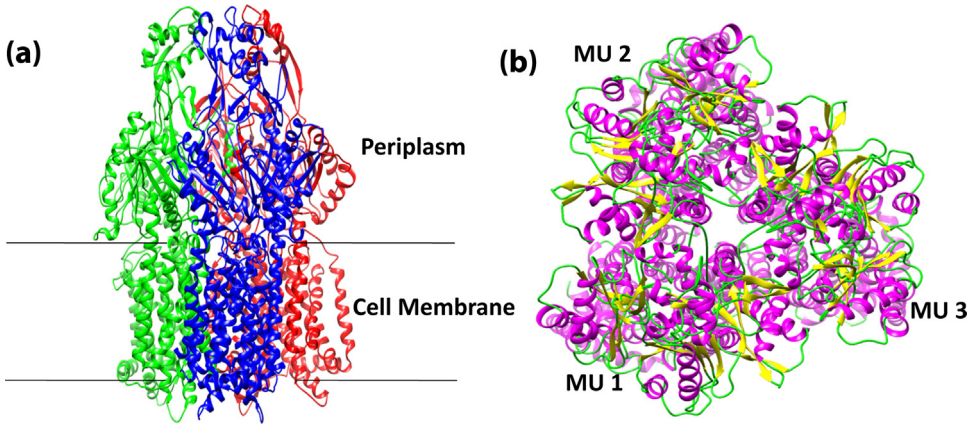

Fig. 4. *Oligomers of the MmpL proteins forming channel across the membrane*: (a) the MmpL structure positioned in the cell membrane. Each of the monomer unit is represented by different colour. The basic architecture of 3D structure resembles with the other RND proteins like CusA (PDB ID: 3K07) and AcrB (PDB ID: 1IWG). Transmembrane region is embedded in the cell membrane and consist of 12 transmembrane helices in each protomer and a total of 36 transmembrane helices form the transmembrane part of the channel. (b) The figure is showing front view of the MmpL oligomer as a functional unit. A pore at both the ends formed with a continuous channel by trimerization of the MmpL monomers. The proximal part is localized in the periplasm and the channel is followed across the length of the oilgomer up to the transmembrane region till the distal end. For the interpretation of references to colours here, the audience is referred to the web version of the article.

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

Table 2
Internal repeats in the *MmpL* genes relevant to the molecular evolution.

Gene name	Loci	Evolutionary pathway	Duplication pattern	Identity (%)	Similarity (%)
MmpL1	Rv0402c	7–12	{2–6}{8–12}	25	45
		6–12	{1–6} {7–12}	24	40
MmpL2	Rv0507	5–11	{3–5}{8–9}	39	48
		7–11	{3–6}{8–11}	26	48
		7–12	{2–6}{8–12}	27	49
MmpL3	Rv0206c	8–12	{3–4}{9–10}	37	52
		9–11	{7–8}{10–11}	46	68
		4–7	{1–3}{5–7}	33	50
MmpL4	Rv0450	7–10	{4–6}{8–10}	26	47
		7–11	{2–6}{7–11}	23	51
MmpL5	Rv0676c	4–6	{3–4}{5–6}	27	53
		7–10	{5–6}{9–10}	34	45
		7–10	{3–5}{8–10}	40	60
		8–11	{4–6}{9–11}	23	41
		7–11	{3–6}{8–11}	24	45
		7–12	{2–6}{8–11}	25	50
MmpL7	Rv2942	4–6	{3–4}{5–6}	40	50
		4–12	{3–4}{11–12}	30	46
		7–12	{2–5}{9–12}	30	46
MmpL8	Rv3832c	8–11	{2–4}{9–11}	42	63
		8–12	{3–5}{10–12}	35	51
		7–12	{2–6}{8–12}	35	51
MmpL9	Rv2339	7–11	{2–5}{8–12}	28	48
		7–12	{2–6}{8–12}	25	48
MmpL11	Rv0202c	6–11	{1–5}{7–11}	29	46
		7–12	{1–6}{8–12}	29	47
		7–12	{3–7}{8–12}	26	47
		6–12	{1–6}{7–12}	26	47
MmpL12	Rv1522c	6–10	{2–5}{7–10}	29	50

M. tuberculosis and *M. leprae*. Genomes of *M. tuberculosis* and *Streptomyces coelicolor* have high level similarities and share many gene clusters corresponding to the identical cellular functions (Bentley et al., 2002).

Four conserved transporter signature motifs have been identified in the protein families involved in the antibiotic and antiseptic resistance (Paulsen et al., 1996). The conservation of one of these motifs, motif D in the TMD4 of MmpL10 and MmpL12 proteins with significant homology (Supplementary Fig. 2) indicating sequence similarities in MmpL and RND proteins. The presence of this motif in TMD4 indicates that this may be essential for the function of MmpL proteins as TMD4 was reported to be important for the proton translocation in case of the AcrB protein of *E. coli* (Piddock, 2006). However, only two of the MmpL proteins were observed to possess this motif.

MmpL proteins show the presence of two common motifs in TMD4 and TMD10 (Fig. 1) which was obtained by pairwise comparison of MmpL proteins having 11 or 12 TMDs with each other. These motifs have prevalent aspartate (D) and glutamate (E) residues. These two residues have negatively charged carboxyl chain required for the interaction with the positively charged protons and are able to transport protons or any of the +vely charged substrate in the symport/antiport fashion as reported earlier in the case of the Melibinose Permease of *E. coli* (Zani et al., 1993) and Myoinsiotol/H^+ symporter from *Leshmania donovani* (Seyfang et al., 1997).

The multiple sequence alignments followed by phylogenetic clustering of MmpL sequences with the RND permease superfamily proteins yielded into its further classification (Fig. 2b). MmpL1, MmpL2, MmpL4, MmpL5, MmpL8, MmpL9, MmpL10 and MmpL12 clustered with the ActII protein of *Streptomyces coelicolor*, a member of the HAE2 family, while the MmpL3, the MmpL7 and the MmpL11 was clustered with the HAE3 family, Hopanoid biosynthesis related protein (Fig. 2a). The HAE families are subfamilies of RND permeases superfamily and are involved in lipid and fatty acid transport along with transportation of drug molecules across the cell membrane (Tseng et al., 1999). Phylogenetic classification of MmpL proteins in the different RND subfamilies was carried out on the basis of the homology in their amino acid sequences, resulting in clustering of MmpL7 and MmpL8 with the transporters of methyl branched fatty acid chains, Pthioserol dimycocerosate (PDIM) and Sulfolipids-1 (SL-1) in two different groups. MmpL8 shares homology with members of HAE2 family, while MmpL7 shares homology with the members of HAE2 family (Cox et al., 1999; Converse et al., 2003).

The modelled tertiary structure of the MmpL proteins were found homologous to the RND proteins, all three dimensional models were generated using multiple templates, the CusA cation efflux transporter of *E. coli* (PDB ID: 3K07), MexB (PDB ID: 2V50) integral membrane component of the MexAB-oprM efflux pump and Acriflavin resistance protein AcrB (PDB ID: 1IWG) structures. The MmpL proteins acquire a homotrimeric conformation similar to the RND efflux proteins, when subjected to oligomerization in virtual settings using Symmdock program (Schneidman-Duhovny et al., 2005). They form a channel like arrangement in trimeric structures similar to AcrB of *E. coli* (Murakami et al., 2002) and MexB of *P. aeruginosa* (Sennhauser et al., 2009). However, all of the MmpL proteins were modelled and the oligomers yielded into conformations similar to the integral membrane component of the RND efflux pumps. Out of all these structures, the MmpL5 oligomer was more refined to choose as representative structure.

The MmpL13a and MmpL13b individually do not qualify as the RND proteins on the basis of their topological features, but when

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

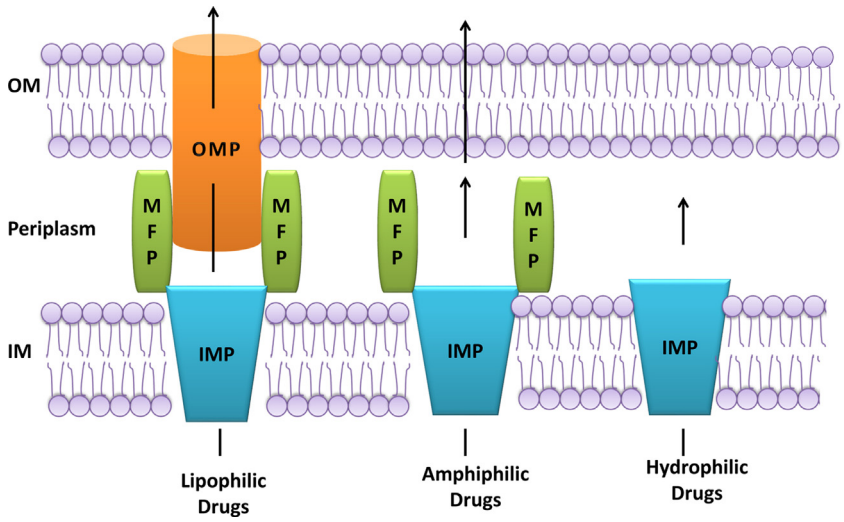

Fig. 5. *The classification of the RND pumps on the basis of constituent protein subunit components involved in the transport*: the RND pumps are multidrug efflux pumps. On the basis of evidences available, they may be divided into three categories (a) Tripartite or three component efflux pumps of the RND family in which integral membrane protein work in association with the membrane fusion protein and outer membrane channel (e.g. AcrAB-TolC in *E. coli* and MexAB-OprM in *Pseudomonas aeruginosa*) (Tikhonova and Zgurskaya, 2004; Poole et al., 1993) and function in expulsion of lipophillic drugs. (b) Putative two component pumps consist of integral membrane protein and the membrane fusion protein, which carry out the expulsion of amphiphillic drugs, but there is no experimental evidence of such pumps, these observations are based on the sequences of these genes from completed and unfinished genomes in the databases (c) Single component RND efflux pumps in which only integral membrane protein carry out the expulsion of the hydrophillic drugs like aminoglycosidic drugs to the periplasm for eg. AcrD of *E. coli* (Aires and Nikaido, 2005).

both the proteins were fused together and analyzed for the topological and structural similarities, they show homology with the well-known RND permeases superfamily proteins and with the other MmpL proteins (Fig. 3a), except in the periplasmic region due to the presence of shorter loops in the fusion protein shown in yellow (Fig. 3b). We were unable to generate the trimeric structure of the fusion protein, which might be due to the differences in periplasmic part, which contributes to the symmetry of the RND oligomers. Hereby, we speculate that this shortening of the loop regions could be an effect of ongoing reductive molecular evolution of the *MmpL13* gene in the *M. tuberculosis* complex. The short loop regions require less energy for folding, thereby increases the thermal stability of the protein**.** Therefore, a protein structure could be stable at higher temperatures. This could give an evolutionary advantage to the *M. tuberculosis* being a strict parasite on a warm blooded host like humans. When orthologs of the *MmpL13* from *Mtb* were analyzed, we observed the presence of the intact *MmpL13* in case of *M. bovis* and two split pseudogenes in case of the *M. leprae* (Fig. 3c), both of these are the members of the *M. tuberculosis* complex. This shows that the *MmpL13* gene may represent the case of evolution from intact to split genes as it is found intact in the case of *M. canetti* and *M. bovis*, while split in case of *M. africanum* (as observed in Uniprot database), *M. tuberculosis H37Rv* and *M. tuberculosis* CDC1551 (Deshayes et al., 2008), members of *M. tuberculosis* complex. *M. marinum*, an outgroup of *M. tuberculosis* complex and some other mycobacterial species like *M. ulcerans, M. tuberculosis* FJ0194, *M. tuberculosis* T46 (as observed in Uniprot database), also show intact form of *MmpL13*. These evidences are suggesting that the *MmpL13* could be an example of strain specific gene splitting not the species specific. This interruption in *MmpL13* gene might be a result of reductive genome evolution and regarded as gene under pseudogenization process in case of *M. tuberculosis*. It is evident from sequence analysis from database searches that they had already become pseudogenes in case of present day *M. leprae*, where they have probably lost their functions. Most of the decay in *M. leprae* genome happened through the process of deletion and pseudogenization. However, all the genes have not followed this identical mechanism of molecular evolution. *M. leprae* has been evolved as a champion pathogen, if we analyze its virulence, despite of massive genome decay it retained most of its mainstay genome intact. Therefore, regulation of virulence is remained the crucial issue in *M. leprae*, which makes it an intriguing pathogen when we discuss about comparative phylogenetics of slow growing mycobacteria. Comparative study of the gene structures of the present day 'live' genes of *Mtb* to its homologous pseudogenized genes in *M. leprae* makes this observation, even more interesting as it may have happened in case of *MmpL13* in *Mtb* complex. While recently, an opposite case of gene fusion was also documented for *MmpS6* and *MmpL6* genes. Where the "ancestral strain" of *Mtb* undeleted TbD1 (2153 bp) region encodes for two independent genes *MmpS6* and *MmpL6*. In "modern strain" in which TbD1 region is deleted, *MmpS6* and *MmpL6* fused in to a single gene as *MmpS6–MmpL6* (Boritsch et al., 2014).

RND family proteins are generally reported to work in association with an outer membrane factor and a periplasmic efflux protein to export molecules of interest out of the cells, but in some cases two components and one component systems were also reported. Two component systems possess an inner membrane component and a periplasmic efflux protein, whereas one component system have been reported with only one inner membrane component alone involved in transport mechanism (Schweizer, 2003). On this basis these transporters could be classified into three different types as shown in Fig. 5. MmpL proteins may be designated as RND proteins of *M. tuberculosis* on the basis of above discussed evidences but we were unable to find any traces of periplasmic and outer membrane components encoding genes in the vicinity of *MmpL* loci or far from them in the *M. tuberculosis* genome, so they may represent one component type of RND proteins like AcrD of *E. coli* (Aires and Nikaido, 2005).

In addition to these, MmpL proteins also exhibit internal duplication of the TMDs, a feature exhibited by many of the transporter proteins. The presence of internal duplicated repeats was indicated in the case of MmpL proteins earlier (Tekaia et al., 1999). We have validated internal duplication event in the case of MmpL proteins encoding genes using a different approach, involving construction and comparison of partial sequence libraries which is a more extensive analysis than the earlier work (Tekaia et al., 1999). Internal duplication results in molecular evolution as reported in the case of other prokaryotic transmembrane proteins, many of them are involved in transport across the membrane. Internal

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

gene duplication was reported in the case of an ABC transporter (Rv1217c) of *M. tuberculosis* (Shimizu et al., 2004). Similar duplication was also reported in AcrB efflux pump which Exhibit 2 folds symmetry around an axis perpendicular to the membrane bilayer (Choi et al., 2008). Similarly, our results clearly show that MmpL proteins may have evolved as a result of such a duplication process involving 3, 4, 5 and 6 TMDs. These evolutionary events equipped these pumps with special structural features such as the manyfold duplicated transmemberane helices and the characteristic symmetry, which help to acquire the structure of the channels to serve its transport functions.

Conclusions

In conclusion, 10 of the MmpL proteins may belong to the extended RND permeases superfamily of bacterial transporters, in the light of their topological similarities, presence of motifs in TMD 4 and TMD 10, which may be involved in proton translocation. They may be working as a single component pump, in contrast with the conventional three component RND efflux system due to the lack of evidence for the presence of related outer membrane factors and periplasmic efflux proteins in the *M. tuberculosis* genome. They have evolved with the internal gene duplication events, a characteristic feature of transport proteins. The observation of presence of split genes within *MmpL* genes, which are present as respective intact genes in the ancient genomes of *M. tuberculosis* complex, representing the evolutionary snapshot of these in the process of pseudogenization. Experimental studies on Integral Membrane Proteins (IMPs) are difficult and time consuming; the conventional approaches used for analysing the membrane associated proteins are usually achieved by isolating the cell envelope and separating the components by providing harsh conditions which results in mixing of the constituents. This ultimately hinders to achieve experimental results. Further, it is also very difficult to express IMPs heterologously. Instead of studying randomly chosen biochemical and biophysical properties of these transporters, the rationally scrutinized *in silico* analysis as achieved by the presented work will be proven to be better start-up data for experimental validation experiments. The presented data on one of the possible molecular mechanisms behind the alarming dissemination of antibiotic resistance in *M. tuberculosis* strains will be proven helpful for the experimentalists and clinicians to devise efficient therapeutic strategies and to further validate the molecular evolution and its implications on the transport function of this important family of proteins in mycobacteria.

Acknowledgements

We thank Central University of Himachal Pradesh for providing Infrastructure to carry out research work and UGC for providing financial support in the form of stipend to PS. Research in YA lab. is supported by Indian government's extramural research funds from UGC, ICMR and SERB (DST).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmm.2015.03.005.

References

Aires, J.R., Nikaido, H., 2005. Aminoglycosides are captured from both periplasm and cytoplasm by the AcrD multidrug efflux transporter of *Escherichia coli*. J. Bacteriol. 187, 1923–1929, http://dx.doi.org/10.1128/JB.187.6.1923-1929.2005.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Balganesh, M., Dinesh, N., Sharma, S., Kuruppath, S., Nair, A.V., Sharma, U., 2012. Efflux pumps of *Mycobacterium tuberculosis* play a significant role in antituberculosis activity of potential drug candidates. Antimicrob. Agents Chemother. 56, 2643–2651, http://dx.doi.org/10.1128/AAC.06003-11.

Bentley, S.D., Chater, K.F., Cerdeno-Tarraga, A.M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C.W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B.G., Parkhill, J., Hopwood, D.A., 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417, 141–147, http://dx.doi.org/10.1038/417141a.

Brennan, P.J., Nikaido, H., 1995. The envelope of mycobacteria. Annu. Rev. Biochem. 64, 29–63, http://dx.doi.org/10.1146/annurev.bi.64.070195.000333.

Boritsch, E.C., Supply, P., Honoré, N., Seeman, T., Stinear, T.P., Brosch, R., 2014. A glimpse into the past and predictions for the future: the molecular evolution of the tuberculosis agent. Micro Rev. 93, 835–852, http://dx.doi.org/10.1111/mmi.12720.

Choi, S., Jeon, J., Yang, J.S., Kim, S., 2008. Common occurrence of internal repeat symmetry in membrane proteins. Proteins 71 (1), 68–80, http://dx.doi.org/10.1002/prot.21656.

Chung, Y.J., Saier Jr., M.H., 2001. SMR type multidrug resistance pumps. Curr. Opin. Drug Discovery Dev. 4, 237–245.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry 3rd, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., Barrell, B.G., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544, http://dx.doi.org/10.1038/31159.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R., Barrell, B.G., 2001. Massive gene decay in the leprosy bacillus. Nature 409 (6823), 1007–1011.

Corpet, F., 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 16 (22), 10881–10890, http://dx.doi.org/10.1093/nar/16.22.10881.

Cox, J.S., Chen, B., McNeil, M., Jacobs Jr., W.R., 1999. Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. Nature 402, 79–83.

Converse, S.E., Mougous, J.D., Leavell, M.D., Leary, J.A., Bertozzi, C.R., Cox, J.S., 2003. MmpL8 is required for sulfolipid-1 biosynthesis and *Mycobacterium tuberculosis* virulence. Proc. Natl. Acad. Sci. U.S.A. 100, 6121–6126.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190, http://dx.doi.org/10.1101/gr.849004.

Cserzö, M., Wallin, E., Simon, I., Von Heijne, G., Elofsson, A., 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng. 10 (6), 673–676.

DeLano, W.L., 2002. The PyMOL Molecular Graphics System. DeLano Scientific LLC, San Carlos, CA.

Deshayes, C., Perrodou, E., Euphrasie, D., Frapy, E., Poch, O., Bifani, P., Lecompte, O., Reyrat, J.M., 2008. Detecting the molecular scars of evolution in the *Mycobacterium tuberculosis* complex by analyzing interrupted coding sequences. BMC Evol. Biol. 8 (1), 78, http://dx.doi.org/10.1186/1471-2148-8-78.

Doughty, D.M., Coleman, M.L., Hunter, R.C., Sessions, A.L., Summons, R.E., Newman, D.K., 2011. The RND-family transporter, HpnN, is required for hopanoid localization to the outer membrane of Rhodopseudomonas palustris TIE-1. Proc. Nat. Acad. Sci. 108 (45), E1045–E1051.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Fernandez-Moreno, M.A., CaballBro, J.L., Hopwood, D.A., Malpartida, F., 1991. The act cluster contains regulatory and antibiotic export genes, direct targets for translational control by the *bld*A tRNA gene of *Streptomyces*. Cell 66, 769–780, http://dx.doi.org/10.1016/0092-8674(91)90120-N.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., Bairoch, A., 2005. Protein identification and analysis tool on the ExPASy server. In: Walker, J.M. (Ed.), The Proteomics Protocols Handbook. Humana Press, New York City, pp. 571–607.

Geer, L.Y., Domrachev, M., Lipman, D.J., Bryant, S.H., 2002. CDART: protein homology by domain architecture. Genome Res. 12 (10), 1619–1623, http://dx.doi.org/10.1101/gr.278202.

WHO, 2014. Global Tuberculosis Report. WHO.

Gouet, P., Courcelle, E., Stuart, D.I., Mètoz, F., 1999. ESPript: analysis of multiple sequence alignment in Postscript. Bioinformatics 15, 305–308.

Henderson, P.J.F., Maiden, M.C.J., 1990. Homologous sugar transport proteins in *Escherichia coli* and their relatives in prokaryotes and eukaryotes. Philos. Trans. R. Soc. London Ser. B326, 391–410.

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005

Hennerdal, A., Falk, J., Lindahl, E., Elofsson, A., 2010. Internal duplications in α-helical membrane protein topologies are common but the nonduplicated forms are rare. Protein Sci. 19, 2305–2318, http://dx.doi.org/10.1002/pro.510.

Hewitt, E.W., Lehner, P.J., 2003. The ABC-transporter signature motif is required for peptide translocation but not peptide binding by TAP. Eur. J. Immunol. 33, 422–427.

Jain, M., Cox, J.S., 2005. Interaction between polyketide synthase and transporter suggests coupled synthesis and export of virulence lipid in *M. tuberculosis*. PLoS Pathog. 1 (1), e2.

Johnson, J.M., Church, G.M., 1999. Alignment and structure prediction of divergent protein families: periplasmic and outer membrane proteins of bacterial efflux pumps. J. Mol. Biol. 287 (3), 695–715, http://dx.doi.org/10.1186/1471-2180-6-66.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33, 3038–3049.

Kelley, L.A., Sternberg, M.J.E., 2009. Protein structure prediction on the web: a case study using the Phyre server. Nat. Protoc. 4, 363–371, http://dx.doi.org/10.1038/nprot.2009.2.

Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a Hidden Markov Model: application to complete genomes. J. Mol. Biol. 305 (3), 567–580, http://dx.doi.org/10.1006/jmbi.2000.4315.

Kuroda, T., Tsuchiya, T., 2009. Multidrug efflux transporters in the MATE family. Biochim. Biophys. Acta 1794, 763–768, http://dx.doi.org/10.1016/j.bbapap.2008.11.012.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157 (1), 105–132.

La Rosa, V., Poce, G., Canseco, J.O., Buroni, S., Pasca, M.R., Biava, M., Raju, R.M., Porretta, G.C., Alfonso, S., Battilocchio, C., Javid, B., Sorrentino, F., Ioerger, T.R., Sacchettini, J.C., Manetti, F., Botta, M., De Logu, A., Rubin, E.J., De Rossi, E., 2012. MmpL3 is the cellular target of the antitubercular pyrrole derivative BM212. Antimicrob. Agents Chemother. 56 (1), 324–331, http://dx.doi.org/10.1128/AAC.05270-11.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. ClustalW and ClustalX version 2. Bioinformatics 23 (21), 2947–2948, http://dx.doi.org/10.1002/0471250953.bi0203s00.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C., 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262 (5131), 208–214.

Lee, E.W., Chen, J., Huda, M.N., Kuroda, T., Mizushima, T., Tsuchiya, T., 2003. Functional cloning and expression of *EmeA* and characterization of EmeA, a multidrug efflux pump from *Enterococcus faecalis*. Biol. Pharm. Bull. 26, 266–270, http://dx.doi.org/10.1248/bpb.26.266.

Levy, S.B., 1992. Active efflux mechanisms for antimicrobial resistance. Antimicrob. Agents Chemother. 36, 695–703.

Lew, J.M., Kapopoulou, A., Jones, L.M., Cole, S.T., 2011. TubercuList – 10 years after. Tuberculosis 91 (1), 1–7, http://dx.doi.org/10.1016/j.tube.2010.09.008.

Li, W., Upadhyay, A., Fontes, F.L., North, E.J., Wang, Y., Crans, D.C., Grzegorzewicz, A.E., Jones, V., Franzblau, S.G., Lee, R.E., Crick, D.C., Jackson, M., 2014. Novel Insights into the mechanism of inhibition of MmpL3, a target of multiple pharmacophores in *Mycobacterium tuberculosis*. Antimicrob. Agents Chemother. 58 (11), 6413–6423.

Long, F., Su, C.C., Zimmermann, M.T., Boyken, S.E., Rajashankar, K.R., Jernigan, R.L., Edward, W.Y., 2010. Crystal structures of the CusA efflux pump suggest methionine-mediated metal transport. Nature 467 (7314), 484–488, http://dx.doi.org/10.1038/nature09395.

Lovell, S.C., Davis, I.W., Arendall III, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C., 2003. Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. Proteins: Struct., Funct., Genet. 50, 437, http://dx.doi.org/10.1002/prot.10286.

Lubelski, J., Konings, W.N., Driessen, A.J., 2007. Distribution and physiology of ABC-type transporters contributing to multidrug resistance in bacteria. Microbiol. Mol. Biol. Rev. 71, 463–476, http://dx.doi.org/10.1128/MMBR.00001-07.

Ma, D., Cook, D.N., Alberti, M., Pon, N.G., Nikaido, H., Hearst, J.E., 1993. Molecular cloning and characterization of acrA and acrE genes of *Escherichia coli*. J. Bacteriol. 175 (19), 6299–6313.

Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Krylov, D., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Thanki, N., Yamashita, R.A., Yin, J.J., Zhang, D., Bryant, S.H., 2007. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35, D237–D240.

Matsuyama, S.I., Fujita, Y., Mizushima, S., 1993. SecD is involved in the release of translocated secretory proteins from the cytoplasmic membrane of *Escherichia coli*. EMBO J. 12 (1), 265.

Millard, E.E., Gale, S.E., Dudley, N., Zhang, J., Schaffer, J.E., Ory, D.S., 2005. The Sterol-sensing domain of the Niemann-Pick C1 (NPC1) protein regulates trafficking of low density lipoprotein cholesterol. J. Biol. Chem. 280 (31), 28581–28590, http://dx.doi.org/10.1074/jbc.M414024200.

Murakami, S., 2008. Multidrug efflux transporter, AcrB – the pumping mechanism. Curr. Opin. Struct. Biol. 18 (4), 459–465, http://dx.doi.org/10.1016/j.sbi.2008.06.007.

Murakami, S., Nakashima, R., Yamashita, E., Yamaguchi, A., 2002. Crystal structure of bacterial multidrug efflux transporter AcrB. Nature 419 (6907), 587–593, http://dx.doi.org/10.1038/nature01050.

Pao, S.S., Paulsen, I.T., Saier Jr., M.H., 1998. Major facilitator superfamily. Microbiol. Mol. Biol. Rev. 62, 1–34.

Papadopoulos, J.S., Agarwala, R., 2007. COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics 23 (9), 1073–1079, http://dx.doi.org/10.1093/bioinformatics/btm076.

Pasca, M.R., Guglierame, P., Rossi, E.D., Zara, F., Riccardi, G., 2005. MmpL7 Gene of *Mycobacterium tuberculosis* is responsible for isoniazid efflux in *Mycobacterium smegmatis*. Antimicrob. Agents Chemother. 49 (11), 4775–4777, http://dx.doi.org/10.1128/AAC.49.11.4775-4777.2005.

Paulsen, I.T., Brown, M.H., Skurray, R.A., 1996. Proton-dependent multidrug efflux systems. Microbiol. Rev. 60 (4), 575–608.

Perrodou, E., Deshayes, C., Muller, J., Schaeffer, C., Van Dorsselaer, A., Ripp, R., Poch, O., Reyrat, J.M., Lecompte, O., 2006. ICDS database: interrupted Coding sequences in prokaryotic genomes. Nucleic Acids Res. 34, D338–D343.

Piddock, L.V., 2006. Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. Clin. Microbiol. Rev. 19, 382–402, http://dx.doi.org/10.1128/CMR.19.2.382-402.2006.

Poole, K., Heinrichs, D.E., Neshat, S., 1993. Cloning and sequence analysis of an EnvCD homologue in *Pseudomonas aeruginosa*: regulation by iron and possible involvement in the secretion of the siderophore pyoverdine. Mol. Microbiol. 10, 529–544, http://dx.doi.org/10.1111/j.1365-2958.1993.tb00925.x.

Rana, A., Rub, A., Akhter, Y., 2014. Proteome-scale identification of outer membrane proteins in *Mycobacterium avium* subspecies paratuberculosis using a structure based combined hierarchical approach. Mol. Biosyst. 10 (9), 2329–2337, http://dx.doi.org/10.1039/c4mb00234b.

Saitou, N., Nei, M., 1987. The neighbor–joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425.

Sandhu, G.K., 2011. Tuberculosis: current situation, challenges and overview of its control programs in India. J. Global Infect. Dis. 3 (2), 143–150, http://dx.doi.org/10.4103/0974-777X.81691.

Schneidman-duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J., 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 33, 363–367, http://dx.doi.org/10.1093/nar/gki481.

Schweizer, H.P., 2003. Efflux as a mechanism of resistance to antimicrobials in *Pseudomonas aeruginosa* and related bacteria: unanswered questions. Genet. Mol. Res.: GMR 2 (1), 48–62.

Sennhauser, G., Bukowska, M.A., Briand, C., Grütter, M.G., 2009. Crystal structure of the multidrug exporter MexB from *Pseudomonas aeruginosa*. J. Mol. Biol. 389 (1), 134–145, http://dx.doi.org/10.1016/j.jmb.2009.04.001.

Seyfang, A., Kavanaugh, M.P., Landfear, S.M., 1997. Aspartate 19 and glutamate 121 are critical for transport function of the myoinositol/H+ symporter from *Leishmania donovani*. J. Biol. Chem. 272 (39), 24210–24215, http://dx.doi.org/10.1074/jbc.272.39.24210.

Shimizu, T., Mitsuke, H., Noto, K., Arai, M., 2004. Internal gene duplication in the evolution of prokaryotic transmembrane proteins. J. Mol. Biol. 339 (1), 1–15, http://dx.doi.org/10.1016/j.jmb.2004.03.048.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278 (5338), 631–637, http://dx.doi.org/10.1126/science.278.5338.631.

Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G., Cole, S.T., 1999. Analysis of the proteome of *Mycobacterium tuberculosis in silico*. Tuber. Lung Dis. 79 (6), 329–342, http://dx.doi.org/10.1054/tuld.1999.0220 (the official journal of the International Union against Tuberculosis and Lung Disease).

Tikhonova, E.B., Zgurskaya, H.I., 2004. AcrA, AcrB, and TolC of *Escherichia coli* form a stable intermembrane multidrug efflux complex. J. Biol. Chem. 279 (31), 32116–32124.

Tseng, T.T., Gratwick, K.S., Kollman, J., Park, D., Nies, D.H., Goffeau, A., Saier, M.H., 1999. The RND permeases superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins. J. Mol. Microbiol. Biotechnol. 1 (1), 107–125, http://dx.doi.org/10.1186/1471-2180-12-249.

Tusnády, G.E., Simon, I., 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J. Mol. Biol. 283, 489–506.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J.C., 2005. GROMACS: fast, flexible, and free. J. Comput. Chem. 26 (16), 1701–1718, http://dx.doi.org/10.1002/jcc.20291.

Varela, C., Rittmann, D., Singh, A., Krumbach, K., Bhatt, K., Eggeling, L., Besra, G.S., Bhatt, A., 2012. MmpL genes are associated with mycolic acid metabolism in mycobacteria and corynebacteria. Chem. Biol. 19 (4), 498–506, http://dx.doi.org/10.1016/j.chembiol.2012.03.006.

Zani, M.L., Pourcher, T., Leblanc, G., 1993. Mutagenesis of acidic residues in putative membrane-spanning segments of the melibiose permease of *Escherichia coli*. II. Effect on cationic selectivity and coupling properties. J. Biol. Chem. 268 (5), 3216–3221.

Zdobnov, E.M., Apweiler, R., 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17 (9), 847–848, http://dx.doi.org/10.1093/bioinformatics/17.9.847.

Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. Evol. Genes Proteins 97, 97–166.

Please cite this article in press as: Sandhu, P., Akhter, Y., The internal gene duplication and interrupted coding sequences in the *MmpL* genes of *Mycobacterium tuberculosis*: Towards understanding the multidrug transport in an evolutionary perspective. Int. J. Med. Microbiol. (2015), http://dx.doi.org/10.1016/j.ijmm.2015.03.005