ELSEVIER

Using molecular-mimicry-inducing pathways of pathogens as novel drug targets

Anjali Garg[1], Bandana Kumari[1], Neelja Singhal[2] and Manish Kumar[1]

[1] Department of Biophysics, University of Delhi South Campus, New Delhi 110021, India
[2] Department of Microbiology, University of Delhi South Campus, New Delhi 110021, India

Several microbial pathogens cause autoimmune diseases in humans by exhibiting molecular mimicry with the host proteins. However, the contribution of autoimmunity in microbial pathogenesis has not been evaluated critically. Clinical and experimental observations have supported and corroborated that autoimmunity was a fundamental process underlying pathology of human tuberculosis bacteria. In the current review, we propose novel drug targets based on a pathogen's molecular-mimicry-inducing proteins. The process for identification of drug targets has been explained using *Mycobacterium tuberculosis* as a model organism. The procedure described here can be applied for repurposing other known drugs and/or discovery of novel therapeutics against other pathogenic bacteria that exhibit molecular mimicry with the host's proteins.

Introduction

When macromolecules found on pathogens and in host tissues share structural, functional or immunological similarities it is called molecular mimicry [1]. Molecular mimicry can occur in the form of complete identity or homology at the protein level, or as similarity in sequences of amino acids and structure. Sequence-based molecular mimicry plays an important part in immune response to infection and in autoimmune diseases. To attribute an autoimmune disease with molecular mimicry, certain criteria should be met: (i) there should be similarity between an epitope of the host, microorganism or environmental agent; (ii) antibodies or T cells cross-reactive with both epitopes must be detected in patients with an autoimmune disease; (iii) there should be evidence of an epidemiological link between exposure to a microbe or an environmental agent and development of autoimmune disease; and (iv) an autoimmune disease should be able to develop in an animal model when sensitized with the epitopes, exposed to the environmental agent or infected with the microbe [2].

Many pathogens exhibit molecular mimicry with the host proteins and cause autoimmune diseases. These pathogens have been listed in Table S1 (see supplementary material online). A detailed and explicit study on the role of molecular mimicry in microbial pathogenesis has not been conducted for most of the pathogens, except for a few fragmentary studies. For example, it was reported that group A *Streptococcus* and group B *Neisseria meningitides* use molecular mimicry to prevent induction of a pathogen-specific immune response [3]. Autoantibodies responsible for Wegener's granulomatosis and systemic lupus erythematosus have been observed in nearly half of the patients suffering from tuberculosis (TB) [4]. A few other autoimmune diseases such as inflammatory bowel disease, Behçet's disease, ankylosing spondylitis, Crohn's disease, ulcerative colitis and sarcoidosis have been associated with pathogenesis of *Mycobacterium tuberculosis* [5]. In an analysis conducted on differential gene expression among TB patients and patients with autoimmune or infectious diseases, it was found that combination of infection and autoimmune disease signatures could explain 96.7% of the differentially expressed TB signatures [6]. Autoimmunity has not been considered as a crucial process in pathology of TB. It continues to be an overlooked event with fragmentary studies [5].

Blocking the metabolic chokepoint has been used as a successful strategy for identifying new drug targets against a particular or-

Corresponding author: Kumar, M. (manish@south.du.ac.in)

https://doi.org/10.1016/j.drudis.2018.10.010

ganism [7,8]. In the present review, we describe how blocking the chokepoint involved in production of a pathogen's mimicry proteins and their interaction partners can be used for discovery of novel targets against pathogens. In this review, this approach has been explained using *M. tuberculosis* as the model organism. The initial step in this process involves identification of interaction partners of pathogen proteins (IPPP) involved in molecular mimicry with the host proteins. The homologs of the host protein, which might be present in IPPP, are removed, and chokepoints of the metabolic pathway are identified. Finally, drug candidates targeting the chokepoint proteins are selected from the DrugBank database and their efficiency and suitability is assessed.

Schema of drug repurposing

The procedure adopted for the process is explained using *Mycobacterium* spp. as the model organism. In the present manuscript, epitopes of the pathogen and host proteins involved in molecular mimicry are referred to as path-memitope and host-memitope, respectively. Similarly, proteins carrying path-memitope and host-memitope are referred to as path-protein and host-protein, respectively. The steps involved in the process are shown in Fig. 1 and described in detail below.

Data extraction

The experimentally verified events in autoimmune diseases caused by molecular mimicry were obtained from a database developed by us earlier: miPepBase [9]. In brief, miPepBase is an indigenously developed, manually curated database containing information about proteins and peptides that exhibit molecular mimicry and autoimmune diseases. A keyword search in miPepBase using 'mycobacterium' displayed 25 entries and/or events related to mimicry (Table 1). In the 25 events, 20 distinct *Mycobacterium* proteins involved in molecular mimicry were identified. These proteins were responsible for seven different types of autoimmune diseases caused by cross-reactivity with 12 different types of host proteins. We observed that one protein of the pathogen (A0A040DMG3) was removed by UniProt, hence it was excluded from our further studies. The seven different types of autoimmune diseases caused by the remaining 24 molecular mimicry events were encephalomyelitis; leprosy; multiple sclerosis; primary biliary cirrhosis; rheumatoid arthritis; skin disease and type 1 diabetes (Table 1). Also, not all of the 24 molecular mimicry events were caused by the proteins of *M. tuberculosis*. One event was caused by proteins of *Mycobacterium avium*; six were due to proteins of *M. avium* subsp. *paratuberculosis*; four caused by proteins of *Mycobacterium leprae*; one was due to proteins of *Mycobacterium gordonae*; 11 were due to proteins of *M. tuberculosis* and one was caused by proteins of *Mycobacterium bovis*.

Protein–protein interaction search

The IPPP were found using the database STRING [10]. STRING contains information about protein interactions, established by experimental studies and by genomic analysis like domain fusion, phylogenetic profiling and gene neighborhood. We included only those interactions that scored ≥0.4 (i.e., the default value). Using STRING, of the 19 path-proteins, we were able to find interacting partners for 16 proteins. Among the 16 path-proteins, one protein (P9WQ90) was a homo-dimer whereas two proteins (P0A521 and Q49375) were oligomers. For those path-proteins (A0A045I964, A0A0E2WUC4 and Q53467), about which protein-interaction information could not be retrieved using STRING [11], a BLAST search against the UniProtKB database was used to find homologous proteins. The first hits retrieved after the BLAST search of A0A045I964 and A0A0E2WUC4 were I6XH73 and F5Z390, respectively. However, for path-protein Q53467 we did not find any hits with high sequence homology (Table S2, see supplementary material online). Hence, it was removed from further analysis. I6XH73 and F5Z390 were also mycobacterial proteins. The STRING search revealed that, for the 15 path-proteins, there were 148 interacting protein partners. In the present work, if IPPP had alignment identity <50% with alignment coverage <80% with a human protein, they were considered as non-homologous IPPP (nHIPPP). As per this guideline among 148 IPPP, five proteins were homologous IPPP. Hence, these were also excluded from further analysis (Table 2).

Pathway mapping and determination of chokepoints in mycobacterial metabolic pathways

The 143 nHIPPP belonged to *M. leprae*, *M. avium* subsp. *paratuberculosis* and *M. tuberculosis*. Each nHIPPP was mapped in their corresponding metabolic networks in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [12]. KEGG is a database resource that cross-integrates genomic, chemical and systemic functional information of an organism. Because of this, KEGG is widely used as a reference knowledge base for integration and interpretation of large-scale datasets generated by genome sequencing and other high-throughput experimental technologies. The number of pathways to which these proteins were mapped are: 12 for *M. leprae*, 14 for *M. avium* subsp. *paratuberculosis* and 18 for *M. tuberculosis*. The pathways were analyzed manually to find possible chokepoint reaction(s). Our analysis revealed that these 143 proteins were a part of 53 chokepoints.

Authentication of chokepoint targets and druggability of selected targets

The validation of essentiality of chokepoint proteins in mycobacterial metabolic pathways was done in two ways. Homologs of chokepoint proteins were searched in all known mycobacterial proteomes and a total of 45 mycobacterial reference proteomes were present in UniProtKB (in October 2017). If a chokepoint protein showed ≥50% identity over 80% of sequence length in a minimum of ten mycobacterial proteomes, it was considered as a part of the core proteome (Table S3, see supplementary material online). We found that 47 of the 53 chokepoint proteins were part of core proteins (Table 3). Alternatively, a chokepoint protein was considered as an essential protein if it had an alignment identity ≥50% with a protein contained in Database of Essential Genes (DEG), over 80% of sequence length. We also found that 31 of the 53 chokepoint proteins shared a close homolog with DEG proteins (Table 3). Proteins that could not qualify either criterion were removed from further analysis.

The potential drugs that can block the IPPP were searched using DrugBank, the most widely used database of drug molecules [13]. Currently, DrugBank contains ~8200 different categories of drugs, namely FDA-approved small-molecule drugs, FDA-approved bio-

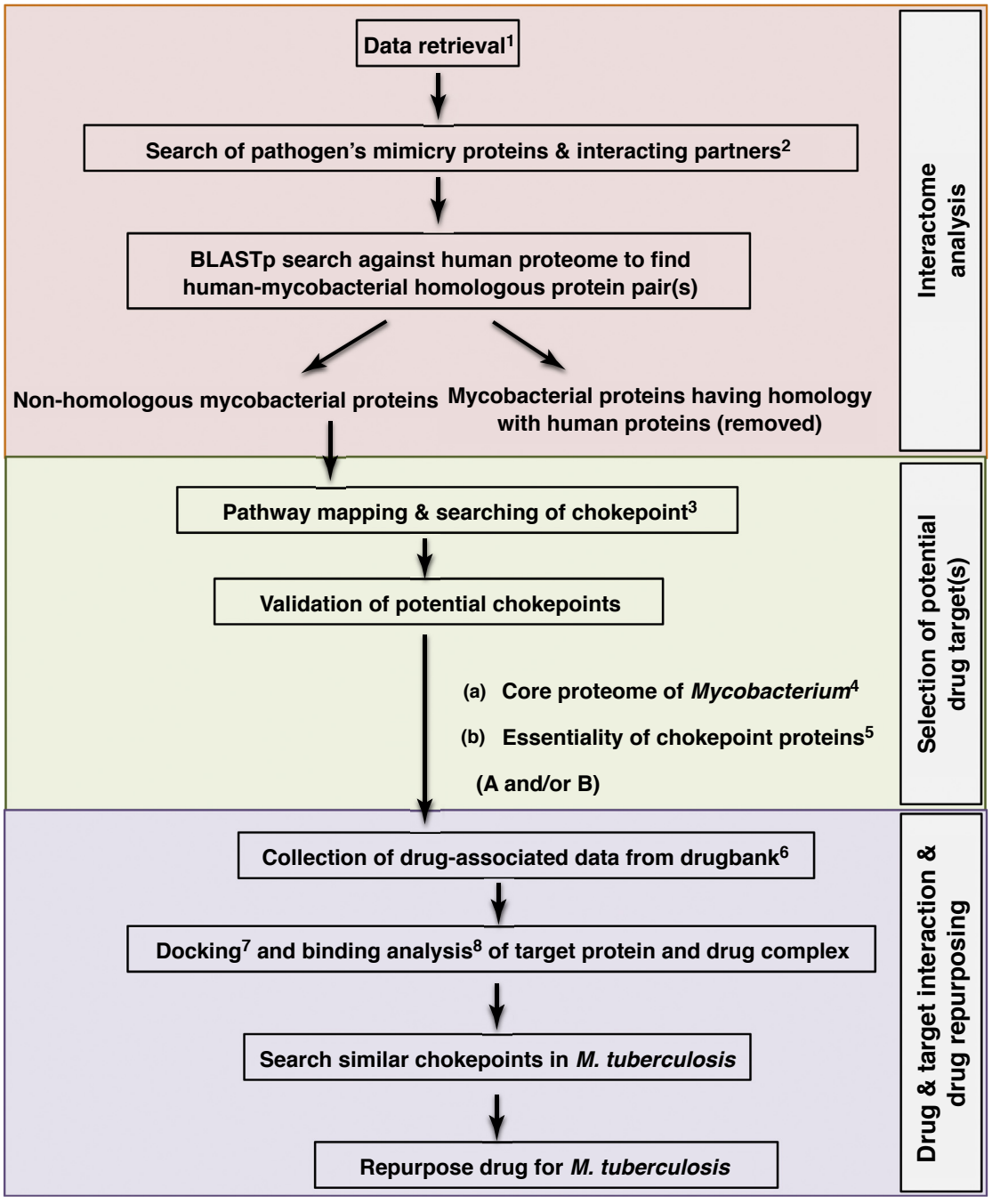

FIGURE 1

The scheme of drug repurposing proposed for *Mycobacterium tuberculosis*. In the figure we show the complete process which is clustered into three major sections. First, interactome analysis includes protein data retrieval, collection of interacting proteins and removal of path-proteins that are homologous to human protein(s). Second, selection of potential drug target(s) that include mapping of mycobacterial nHIPPP on mycobacterial metabolic pathway(s) and search of possible chokepoint protein(s). Chokepoint proteins pass through filters namely part of core proteome (A) or essential proteins (B). All chokepoint protein that crosses either filter is moved to the third step. Third, drug and target interaction and drug repurposing chokepoint proteins were searched for effective ligand(s) and their interaction was analyzed after docking. In the last step *M. tuberculosis* homolog was searched for each chokepoint protein. Superscript numbers reference a list of databases and servers used during the whole process: 1, miPepBase; 2, STRING; 3, KEGG; 4, UniProtKB; 5, DEG; 6, DrugBank; 7, PatchDock; 8, LigPlot+ v.1.4.

tech drugs, nutraceuticals and experimental drugs. To find the appropriate drug candidate, we downloaded sequences of all four types of targets: drug targets, drug enzymes, drug carriers and drug transporters, from DrugBank. Using BLAST we searched for homologs of chokepoint proteins among DrugBank target proteins. The drug molecule associated with the best hit of the DrugBank target protein was considered as a potential binder of homologous chokepoint proteins. Here too, a hit was considered as a homolo-

TABLE 1

List of *Mycobacterium spp.* proteins involved in molecular mimicry

No.	Pathogen protein entry (UniProt AC)	Mimicry peptide	Pathogen protein name	Pathogen name	Host name	Host protein entry (UniProt AC)	Host protein name	Host mimicry peptide	Autoimmune disease
1.	A0A040DMG3[a]	ACFTRPARWTL	Transmembrane protein	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
2.	A0A045I964	QRCRVHFMRNLYTAV	Transposase	*M. tuberculosis*	Human	P02686	Myelin basic protein	ENPVVHFFKNIVTPR	Multiple sclerosis
3.	A0A0E2WUC4	QRCRVHFLRNVLAQV	Transposase	*M. avium*	Human	P02686	Myelin basic protein	ENPVVHFFKNIVTPR	Multiple sclerosis
4.	A5U2C2	AAQHRQIVADF	UvrABC system protein C	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
5.	A5U956	AAQARPVKTVI	MYCTX transferase	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
6.	O32984	VSPWGKPEGRTRKPNKSSNK	50S ribosomal L2	*M. leprae*	Mouse	P02687	Myelin basic protein	VVHFFKNIVTPRTPPPSQGK	Leprosy
7.	O32984	EQANINWGKAGRMRWKGKRP	50S ribosomal L2	*M. leprae*	Mouse	P02687	Myelin basic protein	GAPKRGSGKDGHHAARTTHY	Leprosy
8.	P09239	NA	65 kDa heat shock protein	*M. leprae*	Human	P13645	Cytokeratin-10 of keratin	NA	Leprosy
9.	P0A521	AGKPLLIIAEDVEGE	HSP65	*M. bovis*	Human	P10809	HSP60	HRKPLVIIAEDVDGE	Rheumatoid arthritis
10.	P46861	NTLSAPTFVKDFPVETTPLT	Lysyl-tRNA synthetase	*M. leprae*	Mouse	P02687	Myelin basic protein	VVHFFKNIVTPRTPPPSQGK	Leprosy
11.	P9WG07	AYYGALPLIV	ABC transport	*M. tuberculosis*	Rabbit	P25274	Mid-region encephalitogen from myelin basic protein	TTHYGSLPQK	Multiple sclerosis
12.	P9WM57	ATQYRPDQLAK	Uncharacterized protein R	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
13.	P9WN15	ASMNRPNLVAL	Uncharacterized glycosyl hydrolase	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
14.	P9WPE7	STVKDLLPLL	65 kDa heat shock protein	*M. tuberculosis*	Rat	P02788	Human lactoferrin	SGQKDLLFKD	Rheumatoid arthritis
15.	P9WPE7	STVKDLLPLL	65 kDa heat shock protein	*M. tuberculosis*	Rat	P02787	Human transferrin	PHGKDLLFKD	Rheumatoid arthritis
16.	P9WPE7	VPGGGDMGG	65 kDa heat shock protein	*M. tuberculosis*	Human	P12035	Human keratin	GGYGGGMGG	Skin diseases
17.	P9WPE7	VPGGGDMGG	65 kDa heat shock protein	*M. tuberculosis*	Human	P10809	Human hsp65	GGMGGGMGG	Skin diseases
18.	P9WQ90	ASHQRQRAFAQ	Probable aspartate aminotransferase	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
19.	Q49375	GDL(IL)AE	65 kDa heat shock protein	*M. gordonae*	Human	P10515	Pyruvate dehydrogenase complex-E2	GDLIAE	Primary biliary cirrhosis
20.	Q53467	SHQIRPVCGQR	Putative transport protein	*M. avium* subsp. *paratuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis

TABLE 1 (*Continued*)

No.	Pathogen protein entry (UniProt AC)	Mimicry peptide	Pathogen protein name	Pathogen name	Host name	Host protein entry (UniProt AC)	Host protein name	Host mimicry peptide	Autoimmune disease
21.	Q73T54	MIAVALAGL	Uncharacterized protein	*M. avium* subsp. *paratuberculosis*	Human	Q8IWU4	Beta cell protein zinc transporter 8 (ZnT8)	MIIVSSCAV	Type 1 diabetes
22.	Q73T54	LAANFVVAL	Uncharacterized protein	*M. avium* subsp. *paratuberculosis*	Human	Q8IWU4	Beta cell protein zinc transporter 8 (ZnT8)	VAANIVLTV	Type 1 diabetes
23.	Q73WP1	WYIPPLSPVV	MAP_2619	*M. avium* subsp. *paratuberculosis*	Human	Q16653	Human myelin oligodendrocyte glycoprotein	MEVGWYRPPFSRVVHLYRNGK	Multiple sclerosis
24.	Q741P6	LKYGSLPLSF	SecD	*M. avium* subsp. *paratuberculosis*	Rabbit	P25274	Mid-region encephalitogen from myelin basic protein	TTHYGSLPQK	Multiple sclerosis
25.	Q745A5	PGRRPFTRKELQ	Uncharacterized protein	*M. avium* subsp. *paratuberculosis*	Human	P02686	Myelin basic protein	ENPVVNFFKNIVTP	Multiple sclerosis

Data sourced, with permission, from [9].
[a] Shows obsolete UniProtKB entry.

gous protein if it showed ≥50% identity over 80% of the sequence length. Using all DrugBank target and chokepoint protein pairs, for the five potential chokepoint proteins, we were able to identify 11 drug candidates. Proteins against which we could find drugs were mostly interaction partners of mimicry proteins responsible for multiple sclerosis. In the next stage, these probable drugs were further optimized according to Lipinski's Rule of Five scales: molecular weight ≤500, number of rotatable bonds ≤10, H-bond donors ≤5, H-bond acceptors ≤10 and logP ≤5 (Table S4, see supplementary material online). Additionally, half-life ≥60 min and toxicity information were also considered while evaluating a drug molecule. Those drug molecules that possessed a minimum of five of the seven parameters were considered as probable drugs. Drug-like compounds categorized by DrugBank as dietary supplements, micronutrients or vitamins were excluded.

After benchmarking on the basis of Lipinski's Rule of Five along with toxicity and half-life of drug molecules, we were finally left with four probable drug candidates. Of these four probable drug candidates, we noted that three were experimental approved drugs: DB08185, DB00759 and DB01930 against three chokepoint proteins of *M. leprae* rpsS, rpsC and panC respectively (Table 3). The fourth drug candidate was an experimentally verified drug: DB07349 against narH of *M. avium* subsp. *paratuberculosis*.

Of the four drugs, DB01930 is known to target the enzyme pantothenate synthetase of *M. tuberculosis* (https://www.drugbank.ca/drugs/DB01930). Pantothenate synthetase catalyzes the ATP-dependent condensation of pantoate and β-alanine to form pantothenate (vitamin B5) [14]. It is a known fact that pantothenate biosynthesis is essential for virulence of *M. tuberculosis* [15]. DB07349 is an experimental drug that targets narH and L. Several *in vivo* studies indicate that human lung granuloma, where *M. tuberculosis* resides during latency, is hypoxic and narH and L play an important part in bacterial survival in the hypoxic environment. This suggests that DB07349 can be an ideal drug candidate because it can kill *M. tuberculosis* residing in granuloma. Also, DB07349 can help in complete clearance of *M. tuberculosis* from the host, because long-term persistence of *M. tuberculosis* in the latent stage not only helps it in remaining unaffected during the antitubercular treatment but also helps the pathogen to develop resistance against currently used drugs [16]. The targets of the other two drugs (DB00759 and DB08185) are parts of the ribosomal protein complex. DB00759 (commonly known as tetracycline) is already an approved drug. It is being given to patients orally as well as by an ophthalmic ointment. These are also reported to inhibit the *M. tuberculosis* pathogen growth by binding to the 30S ribosomal subunit and blocking translation [17].

Drug–target interaction

Molecular docking is a useful tool for modeling the interaction between two biomolecules or a small molecule (could be a drug-like molecule) and a biomolecule at the atomic level. It allows us to model the behavior of binding partners in terms of binding affinity or interaction. To assess the binding potential of selected drug candidates with their target, PatchDock was used for docking drug molecules that passed the filtration criteria with their potential targets: gadB, rpsC, rpsS, panC and narH. PatchDock provides a list of receptor and ligand molecule complexes and their PatchDock

TABLE 2

List of pathogen mimicry protein, its interaction partners (IPPP), name of human homolog (if present), KEGG pathway ID to which IPPP belongs and chokepoint proteins

No.	Path-proteins	Interacting proteins of pathogen proteins (IPPP)	Human homolog of IPPP	Proteins could not be mapped on KEGG	KEGG pathway ID	Chokepoint proteins
1	A5U2C2	MRA_1028, MRA_1424, MRA_1430, MRA_1431, MRA_1432, MRA_1648, MtubH3_010100010416, mfd, uvrA, uvrB, uvrC	NA	MRA_1028, MRA_1424, MRA_1430, MRA_1431, MRA_1432, MRA_1648, MtubH3_010100010416	mtu03420	Mfd, uvrA, uvrB, uvrC
2	A5U956	MRA_0226, MRA_0381, MRA_1886, MRA_3014, MRA_3766, MRA_3767, MRA_3895, ethR	NA	NA	NA	NA
3	F5Z390	JDM601_3772, JDM601_3773, JDM601_3774	NA	NA	NA	NA
4	I6XH73	Rv3431c, gadB, nnr, Rv3434c, Rv3435c	NA	Rv3431c, nnr, Rv3434c, Rv3435c	mtu00410, mtu01100, mtu00650, mtu01120, mtu00250, mtu00430, mtu02024, mtu01110	mtu02024:gadB
5	O32984	rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS	NA	NA	mle03010	rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS
7	P09239	clpB, dnaJ1, dnaJ2, dnaK, groL2, groS, grpE, hrcA, htpG, mdh, pheT	dnaK, mdh	clpB, dnaJ1, dnaJ2 , groS, grpE, hrcA, htpG	mle03018, mle05152, mle00970	mle03018:groEL; mle05152:groEL; mle00970:pheT
6	P0A521	NA	NA	NA	NA	NA
8	P46861	argS, gltX, guaA, ileS, leuS, lysS, lysX (ML1393), metG, panC, pheT, proS	NA	NA	mle00970, mle01100, mle01110, mle00770, mle00230, mle00860, mle00450, mle00410, mle01120	mle00970:gltX, metG, leuS, ileS, lysS, argS, proS, pheT; mle00860:gltX; mle00410:panC
9	P9WG07	phoU1 (Rv3301c), phoU2 (Rv0821c), pstA1, pstA2, pstB1 or phoT (Rv0820), pstB2 or pstB (Rv0933), pstC1, pstS1, pstS2, pstS3, tcrX (Rv3765c)	NA	phoU1 (Rv3301c), phoU2 (Rv0821c), tcrX (Rv3765c)	mtu02010, mtu02020, mtu05152	mtu02010: pstB1, pstB2, pstA1, pstA2, pstC1, pstS1, pstS2, pstS3; mtu02020:pstS1, pstS2, pstS3; mtu05152:pstS1, pstS2, pstS3
10	P9WM57	Rv0184, Rv0336, Rv0515, Rv1128c, Rv1278, Rv1378c, Rv1765c, Rv2015c, Rv2100, Rv3074, Rv3776	NA	NA	NA	NA
11	P9WN15	Rv2006, Rv3400, Rv3401, aglA, glgB, glgE, glgX (Rv1564c), otsA, otsB (Rv3372), treS, treZ	NA	Rv3400, Rv3401	mtu01100, mtu00500, mtu01110, mtu00052	mtu00500:glgB; mtu00052:aglA
12	P9WPE7	Rv0312, Rv2264c, dnaJ1, dnaJ2, dnaK, groL2 (Rv0440), groS, hycE, metK, pheT, thrS	dnaK, metK	Rv0312, Rv2264c, dnaJ1, dnaJ2, groS, hycE	mtu00970, mtu05152, mtu03018	mtu00970:thrS, pheT; mtu05152:groL2; mtu03018:groL2
13	P9WQ90	NA	NA	NA	NA	NA
14	Q49375	NA	NA	NA	NA	NA
15	Q53467	NA	NA	NA	NA	NA
16	Q73T54	MAP_2073c, MAP_2138, MAP_2784, MAP_2925, MAP_3865c, MAP_3866c, MAP_3867c, atpA, ctpA, ctpC, nrdE	atpA	MAP_2073MAP_2138, MAP_2784, MAP_2925, MAP_3865c, MAP_3866c, MAP_3867c, ctpA, ctpC	mpa00230, mpa00240, mpa00190	NA

TABLE 2 (*Continued*)

No.	Path-proteins	Interacting proteins of pathogen proteins (IPPP)	Human homolog of IPPP	Proteins could not be mapped on KEGG	KEGG pathway ID	Chokepoint proteins
17	Q73WP1	MAP_0368, MAP_2102c (narK3_1), MAP_3636, MAP_3707c (narK3_2), MAP_4101c, fdhF, narG, narH, narI, narJ, narU	NA	MAP_3636, MAP_4101c, fdhF	mpa00910, mpa01120, mpa02020, mpa01100, mpa00630, mpa00680, mpa01200	mpa00910: narK3_1, narK3_2, narU, narH, narG
18	Q741P6	MAP_1042, MAP_1045, apt, dnaG, relA, secD, secE, secF, secG, secY, tatC	NA	NA	mpa03070, mpa03060, mpa02024, mpa00230, mpa01100, mpa03030	mpa03070: MAP_1042, secD, secE, secF, secG, secY; mpa03060:MAP_1042, secD, secE, secF, secG, secY; mpa02024: MAP_1045, MAP_1042, secE, secG, secY; mpa03030:dnaG
19	Q745A5	MAP_0105c, MAP_0106c, MAP_1410, MAP_2148, MAP_2752, MAP_2963c, MAP_3314c, ftsK, ogt, parB, topA	NA	NA	NA	NA

The table shows information related to pathogen protein involved in molecular mimicry (column 2), IPPP collected from STRING database at default parameters (column 3), HIPPP among the IPPP (column 4), IPPP which could not be mapped on KEGG (column 5), KEGG pathway ID in which IPPP mapped (column 6) and chokepoint proteins found after manual survey of KEGG pathway IDs listed in column 6 (column 7).

scores. The protein–ligand complex with the highest docked score was selected for further analysis. The structures of four potential drugs: DB08185, DB00759, DB01930 and DB07349, were downloaded from DrugBank. We observed that, among all chokepoint proteins, 3D structure of only panC was available in PDB. Hence, the 3D structures of the remaining proteins were obtained from Swiss-model (rpsS and rpsC) and modBase (narH). The intermolecular interactions and strengths, H-bonding, hydrophobic interactions and atom accessibilities are shown in Table S5 (see supplementary material online).

Drug repurposing for *M. tuberculosis*

Molecular mimicry plays an important part in primary establishment of *M. tuberculosis* inside the host. Hence, if *M. tuberculosis* mimicry-inducing proteins can be blocked, the pathogen can be eliminated, well-before it establishes itself inside the host. The steps described above can also be used to propose novel drugs against *M. tuberculosis*. As explained earlier, the 53 chokepoint proteins identified belong to three different species of mycobacteria. Hence, their homologs were searched in the proteome of *M. tuberculosis*. We observed that, of the 53 chokepoints, homologous proteins for 47 chokepoint proteins (14 of *M. tuberculosis*, 20 of *M. leprae* and 13 of *M. avium* subsp. *paratuberculosis*) were present in the proteome of *M. tuberculosis* (Table 3). Hence, we anticipate that these four drugs (DB08185, DB00759, DB01930 and DB07349) might be useful in the treatment of *M. tuberculosis*.

Prospects for the current approach

A lot of research has been done to discover novel drug targets and potent drugs against TB [18–24]. The current approach is different from earlier approaches, because our target here is not an active physiological process or protein(s), which helps in establishing TB bacteria inside the host. Our target is a protein(s) (and interacting partners) that is responsible for eliciting autoimmunity inside the host. Here, the authors propose to target and/or disrupt proteins of *M. tuberculosis* that evoke autoimmune diseases (using drugs or chemical compounds) as a prophylactic measure, before the onset of active TB infection. It would be pertinent to mention here that recent research proposed that mycobacterial infections might have driven autoimmunity as an evolutionary strategy and proteins involved in molecular mimicry are produced in the host long-before the appearance of the symptoms of TB [5]. Thus, our approach might be useful in devising novel prophylactic or vaccination measures against TB.

Another prospective use for our approach is that it can be used as a follow-up remedy after a patient is cured from TB. The drug molecules identified in our current study would disrupt the growth of latent bacteria residing inside the host, which will ultimately lead to clearance of TB bacilli from the host. The other advantage of our approach is that it is in-line with the therapy used for treatment of autoimmune diseases. Tumor necrosis factor (TNF)-blocker therapy is an effective treatment for many autoimmune diseases but it also significantly increases the risk of progression of latent TB to active TB. Thus, before commencing the TNF-blocker therapy for curing autoimmune diseases, patients are first tested for TB infection. Hence, use of a drug that does not involve the use of a TNF-blocker can lead to significant improvement in treatment of pathogen-induced autoimmunity.

TABLE 3

Drug target validation

Path-protein	Chokepoint proteins	Chokepoint proteins that were part of essential gene database	Chokepoint proteins that were part of core proteome	Homolog of chokepoint proteins in *M. tuberculosis* proteome	Chokepoint proteins included as drug-target in DrugBank	Potential drug molecule as per DrugBank against target proteins	Drugs follow at least 5 of 7 drug-like properties
A5U2C2	mfd, uvrA, uvrB, uvrC	uvrC	mfd, uvrA, uvrB, uvrC	mfd, uvrA, uvrB, uvrC	NA	NA	–
I6XH73	gadB	NA	gadB	gadB	gadB	gadB: DB03553	–
O32984	rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS	rplB, rplC, rplD, rplF, rplN, rplP, rplW, rpmC, rpsC, rpsS	rplB, rplC, rplD, rplF, rplN, rplP, rplW, rplV rpmC, rpsC, rpsS	rplB, rplC, rplD, rplF, rplN, rplP, rplW, rpmC, rpsC, rpsS	rpsC, rpsS	rpsC: DB00759; DB09093 rpsS: DB08185; DB00560; DB00759; DB09093	DB08185 (2-methylthio-N6-isopentenyl-adenosine-5′-monophosphate), DB00759 (tetracycline)
P09239	mle03018:groEL; mle05152: groEL; mle00970:pheT	groL2, pheT	groL2, pheT	groL2, pheT	NA	NA	–
P46861	mle00970:gltX, metG, leuS, ileS, lysS, argS, proS, pheT; mle00860:gltX; mle00410: panC	gltX, metG, leuS, ileS, lysS, argS, pheT, panC	gltX, metG, leuS, ileS, lysS, argS, pheT, panC	gltX, metG, leuS, ileS, lysS, argS, pheT, panC	panC	panC: DB01930; DB02596; DB02694; DB03107	DB01930 ((1S)-2-{[{[[(2S)-2,3-dihydroxypropyl]oxy}(hydroxy) phosphoryl]oxy}-1-[(pentanoyloxy)methyl]ethyl octanoate)
P9WG07	mtu02010: pstA1, pstA2, pstB1, pstB2, pstC1, pstS1, pstS2, pstS3; mtu02020: pstS1, pstS2, pstS3; mtu05152:pstS1, pstS2, pstS3	NA	pstA1, pstB1,pstS2, pstS3	pstA1, pstB1,pstS2, pstS3	NA	NA	–
P9WN15	mtu00500:glgB; mtu00052: aglA	glgB	glgB, aglA	glgB, aglA	NA	NA	–
P9WPE7	mtu00970:thrS, pheT; mtu05152:groL2; mtu03018: groL2	thrS, pheT, groL2	thrS, pheT, groL2	thrS, pheT, groL2	NA	NA	–
Q73WP1	mpa00910:narK3_1, narK3_2, narU, narH, narG	NA	narK3_1, narK3_2, narH, narG	narK3_1, narK3_2, narU, narH, narG	narH	narH: DB04464; DB07349	DB07349 (2,4-dihydroxy-3,3-dimethyl-butyrate)
Q741P6	mpa03070:MAP_1042, secD, secE, secF, secG, secY; mpa03060:MAP_1042, secD, secE, secF, secG, secY; mpa02024:MAP_1045, MAP_1042, secE, secG, secY; mpa03030:dnaG	secD, secE, secF, secG, secY, dnaG	MAP_1042, MAP_1045, dnaG, secD, secE, secF, secG, secY	MAP_1042, MAP_1045, dnaG, secD, secE, secF, secG, secY	NA	NA	–

The table shows information of path-proteins (column 1), potential chokepoint found in KEGG metabolic network (column 2), chokepoint proteins which were part of essential genes (column 3) and core proteins (column 4), homologous of chokepoint proteins in *M. tuberculosis* proteome (column 5) and chokepoint protein listed as drug target in DrugBank database (column 6). Column 7 has potential drug molecule as per DrugBank target protein and column 8 contains the drugs that qualified the filter of drug candidate filter.

Concluding remarks

Computational methods and integrated omics approaches, encompassing genomics, proteomics and metabolomics, have proved a valuable tool in drug discovery. Comparative and subtractive genomics proved helpful for prediction and identification of potential therapeutic targets and vaccine candidate proteins in numerous pathogenic bacteria and fungi [25–29]. In the current review, we have described a novel approach to discover new drug targets and drug molecules using a pathogen's molecular-mimicry-inducing proteins. The identification has been done by employing a rigorous systems biology approach. The process and the workflow for identification of drug targets have been explained in detail using *M. tuberculosis* as the model organism. Our systematic analysis revealed that interacting proteins of mimicry-inducing proteins of mycobacteria contain several chokepoint proteins, which can serve as potential drug targets. Inhibitors of the chokepoint proteins were searched from DrugBank employing several stringent filters. The DrugBank search revealed three drug compounds enlisted in the experimental group and one in the approved group, which might be effective against *M. tuberculosis*. Interaction between target(s) and their cognate drug molecule(s) was further confirmed by molecular docking. The drug candidates identified during the course of this study are FDA-approved drug molecules, with proven efficacy against many microbial pathogens. The proposed drug candidates might be tested *in vitro* for assessing their efficacy against *M. tuberculosis* clinical isolates. Thus, instead of developing new chemotherapeutics, our approach helps in repurposing the known drugs against TB.

Using the interaction partners of mimicry proteins, the authors were able to discover only four drug candidates against TB. The trivial number of drugs might be because only one database was used to search drug molecules: DrugBank. DrugBank was preferred over other databases because it provides detailed information about the properties and mechanisms-of-action of ~12000 marketed or experimental drugs. However, the number of probable drug candidates would have increased if data from other relevant databases were also included in the study. For example, databases such as ChEMBL [30], PubChem [31] and ChemBank [32] could be used to provide a comprehensive collection of biological activity, whereas ZINC database [33] could be used for virtual screening. Similarly, incorporation of additional data for example protein–chemical interactions from the Therapeutic Target Database [34] and STITCH [35] can also increase the number of drug targets and candidates. Nevertheless, the authors believe that the scheme described in the current review can be applied for repurposing the known drugs and discovery of novel therapeutics against other pathogenic bacteria that exhibit molecular mimicry with a host's proteins.

Acknowledgments

A.G. is thankful to ICMR for granted scholarship [3/1/3 J.R.F.-2016/LS/HRD-(32262)] and B.K. is thankful to ICMR for granted scholarship [Grant no. BIC/11(33)/2014].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2018.10.010.

References

1 White Cunningham, M. (2009) Molecular mimicry. *eLS* doi: https://doi.org/10.1002/9780470015902.a0000958.pub2
2 Peterson, L.K. and Fujinami, R.S. *et al.* (2007) Molecular mimicry. In *Autoantibodies* (2nd ed.) (Shoenfeld, Y., ed.), pp. 13–19, Elsevier
3 Nisini, R. (2015) Targeting immune-evasion mechanisms as a possible new approach in the fight against tuberculosis. *Immunol. Disord. Immunother.* 1, 101
4 Kakumanu, P. *et al.* (2008) Patients with pulmonary tuberculosis are frequently positive for anti-cyclic citrullinated peptide antibodies, but their sera also react with unmodified arginine-containing peptide. *Arthritis Rheum.* 58, 1576–1581
5 Elkington, P. *et al.* (2016) Tuberculosis: an infection-initiated autoimmune disease? *Trends Immunol.* 37, 815–818
6 Clayton, K. *et al.* (2017) Gene expression signatures in tuberculosis have greater overlap with autoimmune diseases than with infectious diseases. *Am. J. Respir. Crit. Care Med.* 196, 655–656
7 Rahman, S.A. and Schomburg, D. (2006) Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. *Bioinformatics* 22, 1767–1774
8 Gupta, M. *et al.* (2017) Identification of phosphoribosyl-AMP cyclohydrolase, as drug target and its inhibitors in *Brucella melitensis* bv. 1 16M using metabolic pathway analysis. *J. Biomol. Struct. Dyn.* 35, 287–299
9 Garg, A. *et al.* (2017) miPepBase: a database of experimentally verified peptides involved in molecular mimicry. *Front. Microbiol.* 8, 2053
10 Szklarczyk, D. *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368
11 Apweiler, R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–119
12 Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30
13 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–906
14 Jonczyk, R. *et al.* (2008) Pantothenate synthetase is essential but not limiting for pantothenate biosynthesis in *Arabidopsis*. *Plant Mol. Biol.* 66, 1–14
15 Hung, A.W. *et al.* (2016) Optimization of inhibitors of *Mycobacterium tuberculosis* pantothenate synthetase based on group efficiency analysis. *ChemMedChem* 11, 38–42
16 Gomez, J.E. and McKinney, J.D. (2004) *M. tuberculosis* persistence, latency, and drug tolerance. *Tuberculosis* 84, 29–44
17 Hentschel, J. *et al.* (2017) The complete structure of the *Mycobacterium smegmatis* 70S ribosome. *Cell Rep.* 20, 149–160
18 Lamichhane, G. (2011) Novel targets in *M. tuberculosis*: search for new drugs. *Trends Mol. Med.* 17, 25–33
19 Anishetty, S. *et al.* (2005) Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. *Comput. Biol. Chem.* 29, 368–378
20 Singh, V. and Mizrahi, V. (2017) Identification and validation of novel drug targets in *Mycobacterium tuberculosis*. *Drug Discov. Today* 22, 503–509
21 Duncan, K. (2004) Identification and validation of novel drug targets in tuberculosis. *Curr. Pharm. Des.* 10, 3185–3194
22 Ioerger, T.R. *et al.* (2013) Identification of new drug targets and resistance mechanisms in *Mycobacterium tuberculosis*. *PLoS One* 8, e75245
23 Mdluli, K. and Spigelman, M. (2006) Novel targets for tuberculosis drug discovery. *Curr. Opin. Pharmacol.* 6, 459–467
24 Raman, K. *et al.* (2008) targetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. *BMC Syst. Biol.* 2, 109
25 Volker, C. and Brown, J.R. (2002) Bioinformatics and the discovery of novel antimicrobial targets. *Curr. Drug Targets Infect. Disord.* 2, 279–290
26 Johri, A.K. *et al.* (2006) Group B *Streptococcus*: global incidence and vaccine development. *Nat. Rev. Microbiol.* 4, 932–942
27 Doyle, M.A. *et al.* (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* 11, 222
28 Chong, C.E. *et al.* (2006) *In silico* analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. *In Silico Biol.* 6, 341–346

29 Amineni, U. *et al*. (2010) *In silico* identification of common putative drug targets in Leptospira interrogans. *J. Chem. Biol* 3, 165–173
30 Gaulton, A. *et al*. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 40, D1100–1107
31 Kim, S. *et al*. (2016) PubChem substance and compound databases. *Nucleic Acids Res*. 44, D1202–D1213
32 Seiler, K.P. *et al*. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res*. 36, D351–359
33 Irwin, J.J. and Shoichet, B.K. (2005) ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model*. 45, 177–182
34 Li, Y.H. *et al*. (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res*. 46, D1121–1127
35 Szklarczyk, D. *et al*. (2016) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res*. 44, D380–D384

ELSEVIER

Using molecular-mimicry-inducing pathways of pathogens as novel drug targets

Anjali Garg[1], Bandana Kumari[1], Neelja Singhal[2] and Manish Kumar[1]

[1] Department of Biophysics, University of Delhi South Campus, New Delhi 110021, India
[2] Department of Microbiology, University of Delhi South Campus, New Delhi 110021, India

Several microbial pathogens cause autoimmune diseases in humans by exhibiting molecular mimicry with the host proteins. However, the contribution of autoimmunity in microbial pathogenesis has not been evaluated critically. Clinical and experimental observations have supported and corroborated that autoimmunity was a fundamental process underlying pathology of human tuberculosis bacteria. In the current review, we propose novel drug targets based on a pathogen's molecular-mimicry-inducing proteins. The process for identification of drug targets has been explained using *Mycobacterium tuberculosis* as a model organism. The procedure described here can be applied for repurposing other known drugs and/or discovery of novel therapeutics against other pathogenic bacteria that exhibit molecular mimicry with the host's proteins.

Introduction

When macromolecules found on pathogens and in host tissues share structural, functional or immunological similarities it is called molecular mimicry [1]. Molecular mimicry can occur in the form of complete identity or homology at the protein level, or as similarity in sequences of amino acids and structure. Sequence-based molecular mimicry plays an important part in immune response to infection and in autoimmune diseases. To attribute an autoimmune disease with molecular mimicry, certain criteria should be met: (i) there should be similarity between an epitope of the host, microorganism or environmental agent; (ii) antibodies or T cells cross-reactive with both epitopes must be detected in patients with an autoimmune disease; (iii) there should be evidence of an epidemiological link between exposure to a microbe or an environmental agent and development of autoimmune disease; and (iv) an autoimmune disease should be able to develop in an animal model when sensitized with the epitopes, exposed to the environmental agent or infected with the microbe [2].

Many pathogens exhibit molecular mimicry with the host proteins and cause autoimmune diseases. These pathogens have been listed in Table S1 (see supplementary material online). A detailed and explicit study on the role of molecular mimicry in microbial pathogenesis has not been conducted for most of the pathogens, except for a few fragmentary studies. For example, it was reported that group A *Streptococcus* and group B *Neisseria meningitides* use molecular mimicry to prevent induction of a pathogen-specific immune response [3]. Autoantibodies responsible for Wegener's granulomatosis and systemic lupus erythematosus have been observed in nearly half of the patients suffering from tuberculosis (TB) [4]. A few other autoimmune diseases such as inflammatory bowel disease, Behçet's disease, ankylosing spondylitis, Crohn's disease, ulcerative colitis and sarcoidosis have been associated with pathogenesis of *Mycobacterium tuberculosis* [5]. In an analysis conducted on differential gene expression among TB patients and patients with autoimmune or infectious diseases, it was found that combination of infection and autoimmune disease signatures could explain 96.7% of the differentially expressed TB signatures [6]. Autoimmunity has not been considered as a crucial process in pathology of TB. It continues to be an overlooked event with fragmentary studies [5].

Blocking the metabolic chokepoint has been used as a successful strategy for identifying new drug targets against a particular or-

Corresponding author: Kumar, M. (manish@south.du.ac.in)

https://doi.org/10.1016/j.drudis.2018.10.010

ganism [7,8]. In the present review, we describe how blocking the chokepoint involved in production of a pathogen's mimicry proteins and their interaction partners can be used for discovery of novel targets against pathogens. In this review, this approach has been explained using *M. tuberculosis* as the model organism. The initial step in this process involves identification of interaction partners of pathogen proteins (IPPP) involved in molecular mimicry with the host proteins. The homologs of the host protein, which might be present in IPPP, are removed, and chokepoints of the metabolic pathway are identified. Finally, drug candidates targeting the chokepoint proteins are selected from the DrugBank database and their efficiency and suitability is assessed.

Schema of drug repurposing

The procedure adopted for the process is explained using *Mycobacterium* spp. as the model organism. In the present manuscript, epitopes of the pathogen and host proteins involved in molecular mimicry are referred to as path-memitope and host-memitope, respectively. Similarly, proteins carrying path-memitope and host-memitope are referred to as path-protein and host-protein, respectively. The steps involved in the process are shown in Fig. 1 and described in detail below.

Data extraction

The experimentally verified events in autoimmune diseases caused by molecular mimicry were obtained from a database developed by us earlier: miPepBase [9]. In brief, miPepBase is an indigenously developed, manually curated database containing information about proteins and peptides that exhibit molecular mimicry and autoimmune diseases. A keyword search in miPepBase using 'mycobacterium' displayed 25 entries and/or events related to mimicry (Table 1). In the 25 events, 20 distinct *Mycobacterium* proteins involved in molecular mimicry were identified. These proteins were responsible for seven different types of autoimmune diseases caused by cross-reactivity with 12 different types of host proteins. We observed that one protein of the pathogen (A0A040DMG3) was removed by UniProt, hence it was excluded from our further studies. The seven different types of autoimmune diseases caused by the remaining 24 molecular mimicry events were encephalomyelitis; leprosy; multiple sclerosis; primary biliary cirrhosis; rheumatoid arthritis; skin disease and type 1 diabetes (Table 1). Also, not all of the 24 molecular mimicry events were caused by the proteins of *M. tuberculosis*. One event was caused by proteins of *Mycobacterium avium*; six were due to proteins of *M. avium* subsp. *paratuberculosis*; four caused by proteins of *Mycobacterium leprae*; one was due to proteins of *Mycobacterium gordonae*; 11 were due to proteins of *M. tuberculosis* and one was caused by proteins of *Mycobacterium bovis*.

Protein–protein interaction search

The IPPP were found using the database STRING [10]. STRING contains information about protein interactions, established by experimental studies and by genomic analysis like domain fusion, phylogenetic profiling and gene neighborhood. We included only those interactions that scored ≥0.4 (i.e., the default value). Using STRING, of the 19 path-proteins, we were able to find interacting partners for 16 proteins. Among the 16 path-proteins, one protein (P9WQ90) was a homo-dimer whereas two proteins (P0A521 and Q49375) were oligomers. For those path-proteins (A0A045I964, A0A0E2WUC4 and Q53467), about which protein-interaction information could not be retrieved using STRING [11], a BLAST search against the UniProtKB database was used to find homologous proteins. The first hits retrieved after the BLAST search of A0A045I964 and A0A0E2WUC4 were I6XH73 and F5Z390, respectively. However, for path-protein Q53467 we did not find any hits with high sequence homology (Table S2, see supplementary material online). Hence, it was removed from further analysis. I6XH73 and F5Z390 were also mycobacterial proteins. The STRING search revealed that, for the 15 path-proteins, there were 148 interacting protein partners. In the present work, if IPPP had alignment identity <50% with alignment coverage <80% with a human protein, they were considered as non-homologous IPPP (nHIPPP). As per this guideline among 148 IPPP, five proteins were homologous IPPP. Hence, these were also excluded from further analysis (Table 2).

Pathway mapping and determination of chokepoints in mycobacterial metabolic pathways

The 143 nHIPPP belonged to *M. leprae*, *M. avium* subsp. *paratuberculosis* and *M. tuberculosis*. Each nHIPPP was mapped in their corresponding metabolic networks in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [12]. KEGG is a database resource that cross-integrates genomic, chemical and systemic functional information of an organism. Because of this, KEGG is widely used as a reference knowledge base for integration and interpretation of large-scale datasets generated by genome sequencing and other high-throughput experimental technologies. The number of pathways to which these proteins were mapped are: 12 for *M. leprae*, 14 for *M. avium* subsp. *paratuberculosis* and 18 for *M. tuberculosis*. The pathways were analyzed manually to find possible chokepoint reaction(s). Our analysis revealed that these 143 proteins were a part of 53 chokepoints.

Authentication of chokepoint targets and druggability of selected targets

The validation of essentiality of chokepoint proteins in mycobacterial metabolic pathways was done in two ways. Homologs of chokepoint proteins were searched in all known mycobacterial proteomes and a total of 45 mycobacterial reference proteomes were present in UniProtKB (in October 2017). If a chokepoint protein showed ≥50% identity over 80% of sequence length in a minimum of ten mycobacterial proteomes, it was considered as a part of the core proteome (Table S3, see supplementary material online). We found that 47 of the 53 chokepoint proteins were part of core proteins (Table 3). Alternatively, a chokepoint protein was considered as an essential protein if it had an alignment identity ≥50% with a protein contained in Database of Essential Genes (DEG), over 80% of sequence length. We also found that 31 of the 53 chokepoint proteins shared a close homolog with DEG proteins (Table 3). Proteins that could not qualify either criterion were removed from further analysis.

The potential drugs that can block the IPPP were searched using DrugBank, the most widely used database of drug molecules [13]. Currently, DrugBank contains ~8200 different categories of drugs, namely FDA-approved small-molecule drugs, FDA-approved bio-

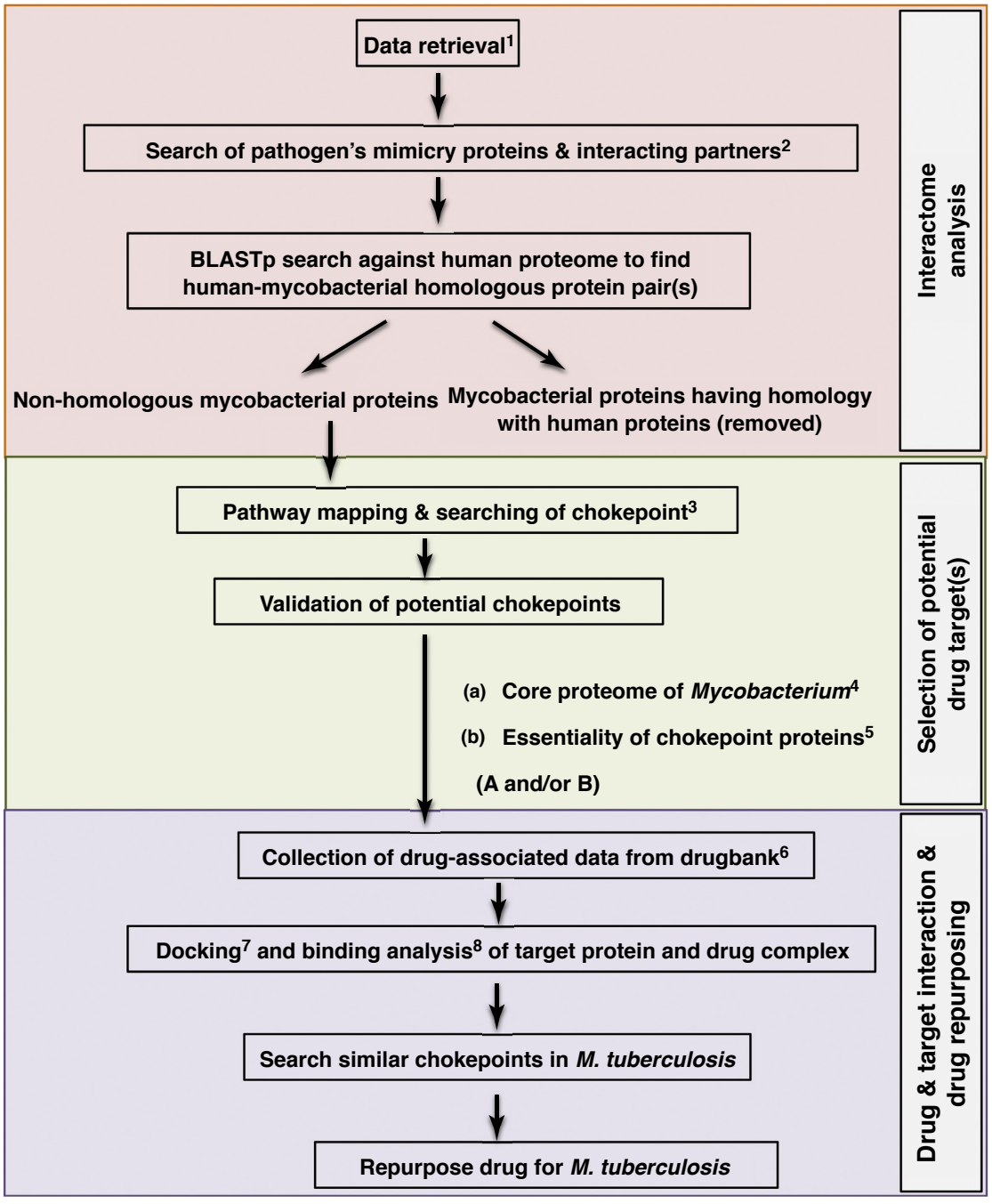

FIGURE 1

The scheme of drug repurposing proposed for *Mycobacterium tuberculosis*. In the figure we show the complete process which is clustered into three major sections. First, interactome analysis includes protein data retrieval, collection of interacting proteins and removal of path-proteins that are homologous to human protein(s). Second, selection of potential drug target(s) that include mapping of mycobacterial nHIPPP on mycobacterial metabolic pathway(s) and search of possible chokepoint protein(s). Chokepoint proteins pass through filters namely part of core proteome (A) or essential proteins (B). All chokepoint protein that crosses either filter is moved to the third step. Third, drug and target interaction and drug repurposing chokepoint proteins were searched for effective ligand(s) and their interaction was analyzed after docking. In the last step *M. tuberculosis* homolog was searched for each chokepoint protein. Superscript numbers reference a list of databases and servers used during the whole process: 1, miPepBase; 2, STRING; 3, KEGG; 4, UniProtKB; 5, DEG; 6, DrugBank; 7, PatchDock; 8, LigPlot+ v.1.4.

tech drugs, nutraceuticals and experimental drugs. To find the appropriate drug candidate, we downloaded sequences of all four types of targets: drug targets, drug enzymes, drug carriers and drug transporters, from DrugBank. Using BLAST we searched for homologs of chokepoint proteins among DrugBank target proteins. The drug molecule associated with the best hit of the DrugBank target protein was considered as a potential binder of homologous chokepoint proteins. Here too, a hit was considered as a homolo-

TABLE 1

List of *Mycobacterium spp.* proteins involved in molecular mimicry

No.	Pathogen protein entry (UniProt AC)	Mimicry peptide	Pathogen protein name	Pathogen name	Host name	Host protein entry (UniProt AC)	Host protein name	Host mimicry peptide	Autoimmune disease
1.	A0A040DMG3[a]	ACFTRPARWTL	Transmembrane protein	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
2.	A0A045I964	QRCRVHFMRNLYTAV	Transposase	*M. tuberculosis*	Human	P02686	Myelin basic protein	ENPVVHFFKNIVTPR	Multiple sclerosis
3.	A0A0E2WUC4	QRCRVHFLRNVLAQV	Transposase	*M. avium*	Human	P02686	Myelin basic protein	ENPVVHFFKNIVTPR	Multiple sclerosis
4.	A5U2C2	AAQHRQIVADF	UvrABC system protein C	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
5.	A5U956	AAQARPVKTVI	MYCTX transferase	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
6.	O32984	VSPWGKPEGRTRKPNKSSNK	50S ribosomal L2	*M. leprae*	Mouse	P02687	Myelin basic protein	VVHFFKNIVTPRTPPPSQGK	Leprosy
7.	O32984	EQANINWGKAGRMRWKGKRP	50S ribosomal L2	*M. leprae*	Mouse	P02687	Myelin basic protein	GAPKRGSGKDGHHAARTTHY	Leprosy
8.	P09239	NA	65 kDa heat shock protein	*M. leprae*	Human	P13645	Cytokeratin-10 of keratin	NA	Leprosy
9.	P0A521	AGKPLLIIAEDVEGE	HSP65	*M. bovis*	Human	P10809	HSP60	HRKPLVIIAEDVDGE	Rheumatoid arthritis
10.	P46861	NTLSAPTFVKDFPVETTPLT	Lysyl-tRNA synthetase	*M. leprae*	Mouse	P02687	Myelin basic protein	VVHFFKNIVTPRTPPPSQGK	Leprosy
11.	P9WG07	AYYGALPLIV	ABC transport	*M. tuberculosis*	Rabbit	P25274	Mid-region encephalitogen from myelin basic protein	TTHYGSLPQK	Multiple sclerosis
12.	P9WM57	ATQYRPDQLAK	Uncharacterized protein R	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
13.	P9WN15	ASMNRPNLVAL	Uncharacterized glycosyl hydrolase	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
14.	P9WPE7	STVKDLLPLL	65 kDa heat shock protein	*M. tuberculosis*	Rat	P02788	Human lactoferrin	SGQKDLLFKD	Rheumatoid arthritis
15.	P9WPE7	STVKDLLPLL	65 kDa heat shock protein	*M. tuberculosis*	Rat	P02787	Human transferrin	PHGKDLLFKD	Rheumatoid arthritis
16.	P9WPE7	VPGGGDMGG	65 kDa heat shock protein	*M. tuberculosis*	Human	P12035	Human keratin	GGYGGGMGG	Skin diseases
17.	P9WPE7	VPGGGDMGG	65 kDa heat shock protein	*M. tuberculosis*	Human	P10809	Human hsp65	GGMGGGMGG	Skin diseases
18.	P9WQ90	ASHQRQRAFAQ	Probable aspartate aminotransferase	*M. tuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis
19.	Q49375	GDL(IL)AE	65 kDa heat shock protein	*M. gordonae*	Human	P10515	Pyruvate dehydrogenase complex-E2	GDLIAE	Primary biliary cirrhosis
20.	Q53467	SHQIRPVCGQR	Putative transport protein	*M. avium* subsp. *paratuberculosis*	Mouse	F6RT34	Myelin basic protein MBPAc	ASQKRPSQRSK	Encephalomyelitis

TABLE 1 (*Continued*)

No.	Pathogen protein entry (UniProt AC)	Mimicry peptide	Pathogen protein name	Pathogen name	Host name	Host protein entry (UniProt AC)	Host protein name	Host mimicry peptide	Autoimmune disease
21.	Q73T54	MIAVALAGL	Uncharacterized protein	*M. avium* subsp. *paratuberculosis*	Human	Q8IWU4	Beta cell protein zinc transporter 8 (ZnT8)	MIIVSSCAV	Type 1 diabetes
22.	Q73T54	LAANFVVAL	Uncharacterized protein	*M. avium* subsp. *paratuberculosis*	Human	Q8IWU4	Beta cell protein zinc transporter 8 (ZnT8)	VAANIVLTV	Type 1 diabetes
23.	Q73WP1	WYIPPLSPVV	MAP_2619	*M. avium* subsp. *paratuberculosis*	Human	Q16653	Human myelin oligodendrocyte glycoprotein	MEVGWYRPPFSRVVHLYRNGK	Multiple sclerosis
24.	Q741P6	LKYGSLPLSF	SecD	*M. avium* subsp. *paratuberculosis*	Rabbit	P25274	Mid-region encephalitogen from myelin basic protein	TTHYGSLPQK	Multiple sclerosis
25.	Q745A5	PGRRPFTRKELQ	Uncharacterized protein	*M. avium* subsp. *paratuberculosis*	Human	P02686	Myelin basic protein	ENPVVNFFKNIVTP	Multiple sclerosis

Data sourced, with permission, from [9].
[a] Shows obsolete UniProtKB entry.

gous protein if it showed ≥50% identity over 80% of the sequence length. Using all DrugBank target and chokepoint protein pairs, for the five potential chokepoint proteins, we were able to identify 11 drug candidates. Proteins against which we could find drugs were mostly interaction partners of mimicry proteins responsible for multiple sclerosis. In the next stage, these probable drugs were further optimized according to Lipinski's Rule of Five scales: molecular weight ≤500, number of rotatable bonds ≤10, H-bond donors ≤5, H-bond acceptors ≤10 and logP ≤5 (Table S4, see supplementary material online). Additionally, half-life ≥60 min and toxicity information were also considered while evaluating a drug molecule. Those drug molecules that possessed a minimum of five of the seven parameters were considered as probable drugs. Drug-like compounds categorized by DrugBank as dietary supplements, micronutrients or vitamins were excluded.

After benchmarking on the basis of Lipinski's Rule of Five along with toxicity and half-life of drug molecules, we were finally left with four probable drug candidates. Of these four probable drug candidates, we noted that three were experimental approved drugs: DB08185, DB00759 and DB01930 against three chokepoint proteins of *M. leprae* rpsS, rpsC and panC respectively (Table 3). The fourth drug candidate was an experimentally verified drug: DB07349 against narH of *M. avium* subsp. *paratuberculosis*.

Of the four drugs, DB01930 is known to target the enzyme pantothenate synthetase of *M. tuberculosis* (https://www.drugbank.ca/drugs/DB01930). Pantothenate synthetase catalyzes the ATP-dependent condensation of pantoate and β-alanine to form pantothenate (vitamin B5) [14]. It is a known fact that pantothenate biosynthesis is essential for virulence of *M. tuberculosis* [15]. DB07349 is an experimental drug that targets narH and L. Several *in vivo* studies indicate that human lung granuloma, where *M. tuberculosis* resides during latency, is hypoxic and narH and L play an important part in bacterial survival in the hypoxic environment. This suggests that DB07349 can be an ideal drug candidate because it can kill *M. tuberculosis* residing in granuloma. Also, DB07349 can help in complete clearance of *M. tuberculosis* from the host, because long-term persistence of *M. tuberculosis* in the latent stage not only helps it in remaining unaffected during the antitubercular treatment but also helps the pathogen to develop resistance against currently used drugs [16]. The targets of the other two drugs (DB00759 and DB08185) are parts of the ribosomal protein complex. DB00759 (commonly known as tetracycline) is already an approved drug. It is being given to patients orally as well as by an ophthalmic ointment. These are also reported to inhibit the *M. tuberculosis* pathogen growth by binding to the 30S ribosomal subunit and blocking translation [17].

Drug–target interaction

Molecular docking is a useful tool for modeling the interaction between two biomolecules or a small molecule (could be a drug-like molecule) and a biomolecule at the atomic level. It allows us to model the behavior of binding partners in terms of binding affinity or interaction. To assess the binding potential of selected drug candidates with their target, PatchDock was used for docking drug molecules that passed the filtration criteria with their potential targets: gadB, rpsC, rpsS, panC and narH. PatchDock provides a list of receptor and ligand molecule complexes and their PatchDock

TABLE 2

List of pathogen mimicry protein, its interaction partners (IPPP), name of human homolog (if present), KEGG pathway ID to which IPPP belongs and chokepoint proteins

No.	Path-proteins	Interacting proteins of pathogen proteins (IPPP)	Human homolog of IPPP	Proteins could not be mapped on KEGG	KEGG pathway ID	Chokepoint proteins
1	A5U2C2	MRA_1028, MRA_1424, MRA_1430, MRA_1431, MRA_1432, MRA_1648, MtubH3_010100010416, mfd, uvrA, uvrB, uvrC	NA	MRA_1028, MRA_1424, MRA_1430, MRA_1431, MRA_1432, MRA_1648, MtubH3_010100010416	mtu03420	Mfd, uvrA, uvrB, uvrC
2	A5U956	MRA_0226, MRA_0381, MRA_1886, MRA_3014, MRA_3766, MRA_3767, MRA_3895, ethR	NA	NA	NA	NA
3	F5Z390	JDM601_3772, JDM601_3773, JDM601_3774	NA	NA	NA	NA
4	I6XH73	Rv3431c, gadB, nnr, Rv3434c, Rv3435c	NA	Rv3431c, nnr, Rv3434c, Rv3435c	mtu00410, mtu01100, mtu00650, mtu01120, mtu00250, mtu00430, mtu02024, mtu01110	mtu02024:gadB
5	O32984	rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS	NA	NA	mle03010	rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS
7	P09239	clpB, dnaJ1, dnaJ2, dnaK, groL2, groS, grpE, hrcA, htpG, mdh, pheT	dnaK, mdh	clpB, dnaJ1, dnaJ2 , groS, grpE, hrcA, htpG	mle03018, mle05152, mle00970	mle03018:groEL; mle05152:groEL; mle00970:pheT
6	P0A521	NA	NA	NA	NA	NA
8	P46861	argS, gltX, guaA, ileS, leuS, lysS, lysX (ML1393), metG, panC, pheT, proS	NA	NA	mle00970, mle01100, mle01110, mle00770, mle00230, mle00860, mle00450, mle00410, mle01120	mle00970:gltX, metG, leuS, ileS, lysS, argS, proS, pheT; mle00860:gltX; mle00410:panC
9	P9WG07	phoU1 (Rv3301c), phoU2 (Rv0821c), pstA1, pstA2, pstB1 or phoT (Rv0820), pstB2 or pstB (Rv0933), pstC1, pstS1, pstS2, pstS3, tcrX (Rv3765c)	NA	phoU1 (Rv3301c), phoU2 (Rv0821c), tcrX (Rv3765c)	mtu02010, mtu02020, mtu05152	mtu02010: pstB1, pstB2, pstA1, pstA2, pstC1, pstS1, pstS2, pstS3; mtu02020:pstS1, pstS2, pstS3; mtu05152:pstS1, pstS2, pstS3
10	P9WM57	Rv0184, Rv0336, Rv0515, Rv1128c, Rv1278, Rv1378c, Rv1765c, Rv2015c, Rv2100, Rv3074, Rv3776	NA	NA	NA	NA
11	P9WN15	Rv2006, Rv3400, Rv3401, aglA, glgB, glgE, glgX (Rv1564c), otsA, otsB (Rv3372), treS, treZ	NA	Rv3400, Rv3401	mtu01100, mtu00500, mtu01110, mtu00052	mtu00500:glgB; mtu00052:aglA
12	P9WPE7	Rv0312, Rv2264c, dnaJ1, dnaJ2, dnaK, groL2 (Rv0440), groS, hycE, metK, pheT, thrS	dnaK, metK	Rv0312, Rv2264c, dnaJ1, dnaJ2, groS, hycE	mtu00970, mtu05152, mtu03018	mtu00970:thrS, pheT; mtu05152:groL2; mtu03018:groL2
13	P9WQ90	NA	NA	NA	NA	NA
14	Q49375	NA	NA	NA	NA	NA
15	Q53467	NA	NA	NA	NA	NA
16	Q73T54	MAP_2073c, MAP_2138, MAP_2784, MAP_2925, MAP_3865c, MAP_3866c, MAP_3867c, atpA, ctpA, ctpC, nrdE	atpA	MAP_2073MAP_2138, MAP_2784, MAP_2925, MAP_3865c, MAP_3866c, MAP_3867c, ctpA, ctpC	mpa00230, mpa00240, mpa00190	NA

TABLE 2 (*Continued*)

No.	Path-proteins	Interacting proteins of pathogen proteins (IPPP)	Human homolog of IPPP	Proteins could not be mapped on KEGG	KEGG pathway ID	Chokepoint proteins
17	Q73WP1	MAP_0368, MAP_2102c (narK3_1), MAP_3636, MAP_3707c (narK3_2), MAP_4101c, fdhF, narG, narH, narI, narJ, narU	NA	MAP_3636, MAP_4101c, fdhF	mpa00910, mpa01120, mpa02020, mpa01100, mpa00630, mpa00680, mpa01200	mpa00910: narK3_1, narK3_2, narU, narH, narG
18	Q741P6	MAP_1042, MAP_1045, apt, dnaG, relA, secD, secE, secF, secG, secY, tatC	NA	NA	mpa03070, mpa03060, mpa02024, mpa00230, mpa01100, mpa03030	mpa03070: MAP_1042, secD, secE, secF, secG, secY; mpa03060:MAP_1042, secD, secE, secF, secG, secY; mpa02024: MAP_1045, MAP_1042, secE, secG, secY; mpa03030:dnaG
19	Q745A5	MAP_0105c, MAP_0106c, MAP_1410, MAP_2148, MAP_2752, MAP_2963c, MAP_3314c, ftsK, ogt, parB, topA	NA	NA	NA	NA

The table shows information related to pathogen protein involved in molecular mimicry (column 2), IPPP collected from STRING database at default parameters (column 3), HIPPP among the IPPP (column 4), IPPP which could not be mapped on KEGG (column 5), KEGG pathway ID in which IPPP mapped (column 6) and chokepoint proteins found after manual survey of KEGG pathway IDs listed in column 6 (column 7).

scores. The protein–ligand complex with the highest docked score was selected for further analysis. The structures of four potential drugs: DB08185, DB00759, DB01930 and DB07349, were downloaded from DrugBank. We observed that, among all chokepoint proteins, 3D structure of only panC was available in PDB. Hence, the 3D structures of the remaining proteins were obtained from Swiss-model (rpsS and rpsC) and modBase (narH). The intermolecular interactions and strengths, H-bonding, hydrophobic interactions and atom accessibilities are shown in Table S5 (see supplementary material online).

Drug repurposing for *M. tuberculosis*

Molecular mimicry plays an important part in primary establishment of *M. tuberculosis* inside the host. Hence, if *M. tuberculosis* mimicry-inducing proteins can be blocked, the pathogen can be eliminated, well-before it establishes itself inside the host. The steps described above can also be used to propose novel drugs against *M. tuberculosis*. As explained earlier, the 53 chokepoint proteins identified belong to three different species of mycobacteria. Hence, their homologs were searched in the proteome of *M. tuberculosis*. We observed that, of the 53 chokepoints, homologous proteins for 47 chokepoint proteins (14 of *M. tuberculosis*, 20 of *M. leprae* and 13 of *M. avium* subsp. *paratuberculosis*) were present in the proteome of *M. tuberculosis* (Table 3). Hence, we anticipate that these four drugs (DB08185, DB00759, DB01930 and DB07349) might be useful in the treatment of *M. tuberculosis*.

Prospects for the current approach

A lot of research has been done to discover novel drug targets and potent drugs against TB [18–24]. The current approach is different from earlier approaches, because our target here is not an active physiological process or protein(s), which helps in establishing TB bacteria inside the host. Our target is a protein(s) (and interacting partners) that is responsible for eliciting autoimmunity inside the host. Here, the authors propose to target and/or disrupt proteins of *M. tuberculosis* that evoke autoimmune diseases (using drugs or chemical compounds) as a prophylactic measure, before the onset of active TB infection. It would be pertinent to mention here that recent research proposed that mycobacterial infections might have driven autoimmunity as an evolutionary strategy and proteins involved in molecular mimicry are produced in the host long-before the appearance of the symptoms of TB [5]. Thus, our approach might be useful in devising novel prophylactic or vaccination measures against TB.

Another prospective use for our approach is that it can be used as a follow-up remedy after a patient is cured from TB. The drug molecules identified in our current study would disrupt the growth of latent bacteria residing inside the host, which will ultimately lead to clearance of TB bacilli from the host. The other advantage of our approach is that it is in-line with the therapy used for treatment of autoimmune diseases. Tumor necrosis factor (TNF)-blocker therapy is an effective treatment for many autoimmune diseases but it also significantly increases the risk of progression of latent TB to active TB. Thus, before commencing the TNF-blocker therapy for curing autoimmune diseases, patients are first tested for TB infection. Hence, use of a drug that does not involve the use of a TNF-blocker can lead to significant improvement in treatment of pathogen-induced autoimmunity.

TABLE 3

Drug target validation

Path-protein	Chokepoint proteins	Chokepoint proteins that were part of essential gene database	Chokepoint proteins that were part of core proteome	Homolog of chokepoint proteins in *M. tuberculosis* proteome	Chokepoint proteins included as drug-target in DrugBank	Potential drug molecule as per DrugBank against target proteins	Drugs follow at least 5 of 7 drug-like properties
A5U2C2	mfd, uvrA, uvrB, uvrC	uvrC	mfd, uvrA, uvrB, uvrC	mfd, uvrA, uvrB, uvrC	NA	NA	–
I6XH73	gadB	NA	gadB	gadB	gadB	gadB: DB03553	–
O32984	rplB, rplC, rplD, rplF, rplN, rplP, rplV, rplW, rpmC, rpsC, rpsS	rplB, rplC, rplD, rplF, rplN, rplP, rplW, rpmC, rpsC, rpsS	rplB, rplC, rplD, rplF, rplN, rplP, rplW, rplV rpmC, rpsC, rpsS	rplB, rplC, rplD, rplF, rplN, rplP, rplW, rpmC, rpsC, rpsS	rpsC, rpsS	rpsC: DB00759; DB09093 rpsS: DB08185; DB00560; DB00759; DB09093	DB08185 (2-methylthio-N6-isopentenyl-adenosine-5′-monophosphate), DB00759 (tetracycline)
P09239	mle03018:groEL; mle05152: groEL; mle00970:pheT	groL2, pheT	groL2, pheT	groL2, pheT	NA	NA	–
P46861	mle00970:gltX, metG, leuS, ileS, lysS, argS, proS, pheT; mle00860:gltX; mle00410: panC	gltX, metG, leuS, ileS, lysS, argS, pheT, panC	gltX, metG, leuS, ileS, lysS, argS, pheT, panC	gltX, metG, leuS, ileS, lysS, argS, pheT, panC	panC	panC: DB01930; DB02596; DB02694; DB03107	DB01930 ((1S)-2-{[{[[(2S)-2,3-dihydroxypropyl]oxy}(hydroxy) phosphoryl]oxy}-1-[(pentanoyloxy)methyl]ethyl octanoate)
P9WG07	mtu02010: pstA1, pstA2, pstB1, pstB2, pstC1, pstS1, pstS2, pstS3; mtu02020: pstS1, pstS2, pstS3; mtu05152:pstS1, pstS2, pstS3	NA	pstA1, pstB1,pstS2, pstS3	pstA1, pstB1,pstS2, pstS3	NA	NA	–
P9WN15	mtu00500:glgB; mtu00052: aglA	glgB	glgB, aglA	glgB, aglA	NA	NA	–
P9WPE7	mtu00970:thrS, pheT; mtu05152:groL2; mtu03018: groL2	thrS, pheT, groL2	thrS, pheT, groL2	thrS, pheT, groL2	NA	NA	–
Q73WP1	mpa00910:narK3_1, narK3_2, narU, narH, narG	NA	narK3_1, narK3_2, narH, narG	narK3_1, narK3_2, narU, narH, narG	narH	narH: DB04464; DB07349	DB07349 (2,4-dihydroxy-3,3-dimethyl-butyrate)
Q741P6	mpa03070:MAP_1042, secD, secE, secF, secG, secY; mpa03060:MAP_1042, secD, secE, secF, secG, secY; mpa02024:MAP_1045, MAP_1042, secE, secG, secY; mpa03030:dnaG	secD, secE, secF, secG, secY, dnaG	MAP_1042, MAP_1045, dnaG, secD, secE, secF, secG, secY	MAP_1042, MAP_1045, dnaG, secD, secE, secF, secG, secY	NA	NA	–

The table shows information of path-proteins (column 1), potential chokepoint found in KEGG metabolic network (column 2), chokepoint proteins which were part of essential genes (column 3) and core proteins (column 4), homologous of chokepoint proteins in *M. tuberculosis* proteome (column 5) and chokepoint protein listed as drug target in DrugBank database (column 6). Column 7 has potential drug molecule as per DrugBank target protein and column 8 contains the drugs that qualified the filter of drug candidate filter.

Concluding remarks

Computational methods and integrated omics approaches, encompassing genomics, proteomics and metabolomics, have proved a valuable tool in drug discovery. Comparative and subtractive genomics proved helpful for prediction and identification of potential therapeutic targets and vaccine candidate proteins in numerous pathogenic bacteria and fungi [25–29]. In the current review, we have described a novel approach to discover new drug targets and drug molecules using a pathogen's molecular-mimicry-inducing proteins. The identification has been done by employing a rigorous systems biology approach. The process and the workflow for identification of drug targets have been explained in detail using *M. tuberculosis* as the model organism. Our systematic analysis revealed that interacting proteins of mimicry-inducing proteins of mycobacteria contain several chokepoint proteins, which can serve as potential drug targets. Inhibitors of the chokepoint proteins were searched from DrugBank employing several stringent filters. The DrugBank search revealed three drug compounds enlisted in the experimental group and one in the approved group, which might be effective against *M. tuberculosis*. Interaction between target(s) and their cognate drug molecule(s) was further confirmed by molecular docking. The drug candidates identified during the course of this study are FDA-approved drug molecules, with proven efficacy against many microbial pathogens. The proposed drug candidates might be tested *in vitro* for assessing their efficacy against *M. tuberculosis* clinical isolates. Thus, instead of developing new chemotherapeutics, our approach helps in repurposing the known drugs against TB.

Using the interaction partners of mimicry proteins, the authors were able to discover only four drug candidates against TB. The trivial number of drugs might be because only one database was used to search drug molecules: DrugBank. DrugBank was preferred over other databases because it provides detailed information about the properties and mechanisms-of-action of ~12000 marketed or experimental drugs. However, the number of probable drug candidates would have increased if data from other relevant databases were also included in the study. For example, databases such as ChEMBL [30], PubChem [31] and ChemBank [32] could be used to provide a comprehensive collection of biological activity, whereas ZINC database [33] could be used for virtual screening. Similarly, incorporation of additional data for example protein–chemical interactions from the Therapeutic Target Database [34] and STITCH [35] can also increase the number of drug targets and candidates. Nevertheless, the authors believe that the scheme described in the current review can be applied for repurposing the known drugs and discovery of novel therapeutics against other pathogenic bacteria that exhibit molecular mimicry with a host's proteins.

Acknowledgments

A.G. is thankful to ICMR for granted scholarship [3/1/3 J.R.F.-2016/LS/HRD-(32262)] and B.K. is thankful to ICMR for granted scholarship [Grant no. BIC/11(33)/2014].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2018.10.010.

References

1 White Cunningham, M. (2009) Molecular mimicry. *eLS* doi: https://doi.org/10.1002/9780470015902.a0000958.pub2
2 Peterson, L.K. and Fujinami, R.S. *et al.* (2007) Molecular mimicry. In *Autoantibodies* (2nd ed.) (Shoenfeld, Y., ed.), pp. 13–19, Elsevier
3 Nisini, R. (2015) Targeting immune-evasion mechanisms as a possible new approach in the fight against tuberculosis. *Immunol. Disord. Immunother.* 1, 101
4 Kakumanu, P. *et al.* (2008) Patients with pulmonary tuberculosis are frequently positive for anti-cyclic citrullinated peptide antibodies, but their sera also react with unmodified arginine-containing peptide. *Arthritis Rheum.* 58, 1576–1581
5 Elkington, P. *et al.* (2016) Tuberculosis: an infection-initiated autoimmune disease? *Trends Immunol.* 37, 815–818
6 Clayton, K. *et al.* (2017) Gene expression signatures in tuberculosis have greater overlap with autoimmune diseases than with infectious diseases. *Am. J. Respir. Crit. Care Med.* 196, 655–656
7 Rahman, S.A. and Schomburg, D. (2006) Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. *Bioinformatics* 22, 1767–1774
8 Gupta, M. *et al.* (2017) Identification of phosphoribosyl-AMP cyclohydrolase, as drug target and its inhibitors in *Brucella melitensis* bv. 1 16M using metabolic pathway analysis. *J. Biomol. Struct. Dyn.* 35, 287–299
9 Garg, A. *et al.* (2017) miPepBase: a database of experimentally verified peptides involved in molecular mimicry. *Front. Microbiol.* 8, 2053
10 Szklarczyk, D. *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368
11 Apweiler, R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–119
12 Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30
13 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–906
14 Jonczyk, R. *et al.* (2008) Pantothenate synthetase is essential but not limiting for pantothenate biosynthesis in *Arabidopsis*. *Plant Mol. Biol.* 66, 1–14
15 Hung, A.W. *et al.* (2016) Optimization of inhibitors of *Mycobacterium tuberculosis* pantothenate synthetase based on group efficiency analysis. *ChemMedChem* 11, 38–42
16 Gomez, J.E. and McKinney, J.D. (2004) *M. tuberculosis* persistence, latency, and drug tolerance. *Tuberculosis* 84, 29–44
17 Hentschel, J. *et al.* (2017) The complete structure of the *Mycobacterium smegmatis* 70S ribosome. *Cell Rep.* 20, 149–160
18 Lamichhane, G. (2011) Novel targets in *M. tuberculosis*: search for new drugs. *Trends Mol. Med.* 17, 25–33
19 Anishetty, S. *et al.* (2005) Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. *Comput. Biol. Chem.* 29, 368–378
20 Singh, V. and Mizrahi, V. (2017) Identification and validation of novel drug targets in *Mycobacterium tuberculosis*. *Drug Discov. Today* 22, 503–509
21 Duncan, K. (2004) Identification and validation of novel drug targets in tuberculosis. *Curr. Pharm. Des.* 10, 3185–3194
22 Ioerger, T.R. *et al.* (2013) Identification of new drug targets and resistance mechanisms in *Mycobacterium tuberculosis*. *PLoS One* 8, e75245
23 Mdluli, K. and Spigelman, M. (2006) Novel targets for tuberculosis drug discovery. *Curr. Opin. Pharmacol.* 6, 459–467
24 Raman, K. *et al.* (2008) targetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. *BMC Syst. Biol.* 2, 109
25 Volker, C. and Brown, J.R. (2002) Bioinformatics and the discovery of novel antimicrobial targets. *Curr. Drug Targets Infect. Disord.* 2, 279–290
26 Johri, A.K. *et al.* (2006) Group B *Streptococcus*: global incidence and vaccine development. *Nat. Rev. Microbiol.* 4, 932–942
27 Doyle, M.A. *et al.* (2010) Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* 11, 222
28 Chong, C.E. *et al.* (2006) *In silico* analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. *In Silico Biol.* 6, 341–346

29 Amineni, U. *et al*. (2010) *In silico* identification of common putative drug targets in Leptospira interrogans. *J. Chem. Biol* 3, 165–173
30 Gaulton, A. *et al*. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 40, D1100–1107
31 Kim, S. *et al*. (2016) PubChem substance and compound databases. *Nucleic Acids Res*. 44, D1202–D1213
32 Seiler, K.P. *et al*. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res*. 36, D351–359
33 Irwin, J.J. and Shoichet, B.K. (2005) ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model*. 45, 177–182
34 Li, Y.H. *et al*. (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res*. 46, D1121–1127
35 Szklarczyk, D. *et al*. (2016) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res*. 44, D380–D384

Journal of Proteomics 200 (2019) 102–110

Contents lists available at ScienceDirect

Journal of Proteomics

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

Proteome profiling of carbapenem-resistant *K. pneumoniae* clinical isolate (NDM-4): Exploring the mechanism of resistance and potential drug targets

Divakar Sharma[a], Anjali Garg[b], Manish Kumar[b], Asad U. Khan[a,*]

[a] *Interdisciplinary Biotechnology Unit, Aligarh Muslim University, Aligarh, India*
[b] *Department of Biophysics, University of Delhi South Campus, India*

ARTICLE INFO

Keywords:
Antibiotic resistance
Carbapenem-resistant *K. pneumoniae*
NDM-4
Proteomics
Functional annotation
Pathways enrichment
STRING

ABSTRACT

The emergence of carbapenem resistance has become a major problem worldwide. This has made treatment of *K. pneumoniae* infections a difficult task. In this study, we have explored the whole proteome of the carbapenem-resistant *Klebsiella pneumonia* clinical isolate (NDM-4) under the meropenem stress. Proteomics (LC-MS/MS) and bioinformatics approaches were employed to uncover the novel mystery of the resistance over the existing mechanisms. Gene ontology, KEGG and STRING were used for functional annotation, pathway enrichment and protein–protein interaction (PPI) network respectively. LC-MS/MS analysis revealed that 52 proteins were overexpressed (≥10 log folds) under meropenem stress. These proteins belong to four major groups namely protein translational machinery complex, DNA/RNA modifying enzymes or proteins, proteins involved in carbapenems cleavage, modifications & transport and energy metabolism & intermediary metabolism-related proteins. Among the total 52 proteins 38 {matched to *Klebsiella pneumonia* subsp. pneumoniae (strain ATCC 700721/MGH 78578)} were used for functional annotation, pathways enrichment and protein–protein interaction. These were significantly enriched in the "intracellular" (14 of 38), "cytoplasm" (12 of 38) and "ribosome" (10 of 38). We suggest that these 52 over expressed proteins and their interactive proteins cumulatively contributed in survival of bacteria and meropenem resistance through various mechanisms or enriched pathways. These proteins targets and their pathways might be used for development of novel therapeutics against the resistance; therefore, the situation of the emergence of "bad-bugs" could be controlled.

1. Introduction

Worldwide resistance to broad-spectrum antimicrobials (extended-spectrum cephalosporins) is a well recognized problem among enterobacteriaceae, a family of Gram-negative bacteria [1,2]. *Klebsiella pneumoniae* is the most significant member of the enterobacteriaceae in the clinical setting. Carbapenems have served as an important antimicrobial class for the treatment of these drug resistant organisms, including those which producing extended spectrum beta-lactamases (ESBLs) [3,4]. However, due to production of carbapenemases, carbapenem resistant enterobacteriaceae (CRE) have been emerged and spread globally which is representing a serious threat to public health. Carbapenemases are specific beta-lactamases with the ability to hydrolyze carbapenems and their production is the most widespread cause of carbapenem resistance. An increasing number of class-A carbapenemases (KPC and GES enzymes), class-B metallo-beta-lactamases (VIM, IMP, and NDM beta-lactamases), class-C beta-lactamases (CMY-10 and PDC beta-lactamases) and class-D carbapenemases (OXA-23) have recently emerged. In addition, overproduction of class-C beta-lactamases (CMY-10 and PDC beta-lactamases) as well as loss of porins can also lead to carbapenem resistance [5–9]. Several explanations have been put forward to explain the mechanisms of carbapenem resistance but still our knowledge regarding resistance is fragmentary.

However less attention has been paid towards the transcriptome and proteome that might play an important role in the development of drug resistance. Proteome is the functional moiety of the cell and directly correlated to the external stimuli like drug stress and resistance [10]. During drug stress, internal harmony of the bacterial system disturbed and rapidly bacteria adopted alternative cellular functions mediated by proteome to overcome the effect. Proteomics coupled with bioinformatics are the potential strategy to explore the biological problems. Differential expression proteome analysis of drug resistant bacteria

Abbreviations: MIC, Minimum inhibitory concentration; ESBLs, Extended spectrum beta-lactamases; CLSI, Clinical and Laboratory Standards Institute; STRING, Search Tool for the Retrieval of Interacting Genes/Proteins; LB, Luria-Bertani

* Corresponding author.
E-mail address: asad.k@rediffmail.com (A.U. Khan).

https://doi.org/10.1016/j.jprot.2019.04.003
Received 20 February 2019; Received in revised form 28 March 2019; Accepted 2 April 2019
Available online 03 April 2019
1874-3919/

under drug stress could unravel the novel mechanism of the bacteria to overcome the effects of drugs. Comparative proteomic studies addressing the whole cell proteome of drug resistant microbial strains with or without drug stress have been reported [11–16]. Through liquid chromatography coupled with mass spectrometry (LC-MS/MS) approach, we have analyzed the over expressed proteome of carbapenem resistant *K. pneumonia* (NDM-4) clinical isolate under the meropenem drug stress.

2. Materials and methods

2.1. Clinical bacterial strains and drug susceptibility testing

Carbapenems resistant *Klebsiella pneumonia* clinical isolate (NDM-4) was used in the present study. Drug susceptibility testing (DST) for the meropenem drug, against carbapenems resistant *Klebsiella pneumonia* clinical isolate was determined by micro-dilution method according to the Clinical and Laboratory Standards Institute (CLSI) guidelines [17].

2.2. Culture and drug induction

K. pneumoniae (NDM-4) clinical isolate was grown in Luria Bertani broth (LB) at 37°C & 220 rpm. Sub-MIC (32 μg/ml) value of meropenem drug was added in one of the flask. Bacteria were grown up to the exponential phase (OD_{600} = 0.8), further cells were harvested by centrifugation at 8000 rpm for 8 min at 4°C and the cell pellet was stored at -80°C until required. All the experiments have replicated biologically.

2.3. Protein sample preparation

Cells were washed with normal saline and re-suspended in the lysis buffer (50 mMTris-HCl containing 10 mM $MgCl_2$, 0.1% sodium azide, 1 mM phenylmethylsulfonyl fluoride (PMSF) and 1 mM (ethylene glycol tetraacetic acid) EGTA; pH 7.4) at the concentration of 1 g wet weight per 5 ml and then broken down by intermittent sonication by sonicator on power at 35% amplitude (Sonics & Materials Inc., Newtown, CT, USA for 10 min at 4 °C. Further homogenate was centrifuged at 12,000 *g* for 20 min at 4 °C and supernatant was stored at −80 °C [12–14]. The supernatant was precipitated overnight at -20°C by adding cold acetone to supernatant in access (1,4). The precipitated protein was collected by centrifugation (12,000 rpm, 20 min) and allowed to air dry and suspended in appropriate volume of protein dissolving buffer. Protein concentration was estimated using the Bradford assay [18]. Experiments have repeated technically.

2.4. Separation and identification of proteome by nanoLC-Triple TOF5600-MS

Equal amount of protein samples were digested using trypsin. Further, digested protein samples were analyzed using a Triple TOF 5600 mass-spectrometer (AB-Sciex, USA) equipped with Eskigent MicroLC 200 system (Eskigent, Dublin, CA) having an Eskigent C18-reverse phase column (150 × 0.3 mm, 3 μm, 120 Å). In order to generate spectral library for protein identification experiment, Data dependent analysis (DDA) was performed for the individual samples, to generate high quality spectral ion libraries for SWATH analysis, by operating the mass spectrometer with specific parameters. Spectral library were generated using information dependent acquisition (IDA) mode after injecting tryptic digest of 2 g on column using Eksigent nano LC-Ultra™ 2D plus system coupled with SCIEX Triple TOF® 5600 system fitted with nano spray III source. The samples were loaded on the trap (Eksigent Chrom XP 350 μm × 0.5 mm, 3 μm 120 Å) and washed for 30 min at 3 μL/min. A 120 min gradient in multiple steps (ranging from 5 to 50% Acetonitrile in water containing 0.1% formic acid) was set up to elute the peptides from the ChromXP 3-C18, 0.075 × 150 mm, 3 μm, 120 Å analytical column. Experiment has done in technical replicates.

2.5. Information dependent acquisition (IDA) parameters

In IDA ion library generation method was used where maximum 20 most intense multiple charged ions per MS cycle were selected to perform MS/MS fragmentation. A dynamic exclusion criterion was applied to each of the ions for 10 s. The accumulation time for each MS/MS experiment was set to 70 ms.

2.6. SWATH parameters for label free quantification

In SWATH acquisition method Q1 transmission window was set to 12 Da from the mass range for 350-1250 Da. Total 75 windows was acquired independently with an accumulation time of 62 ms, along with three technical replicates for each of the sets. Total cycle time was kept constant at < 5 s. To generate spectral library Protein Pilot™ v. 5.0 was used. For Label free quantification the peak extraction and spectral alignment were performed using PeakView® 2.2 Software with the specific parameters such as; number of peptides as 2, number of transitions as 5, peptide confidence as 95%, XIC width as 30 ppm, XIC extraction window as 3 min. The data was further subjected to Marker View software V 1.3 (AB Sciex) to get statistical data interpretation. In Marker View, Peak area under the curve for the selected peptides was normalized by internal standard protein (beta galactosidase) spike during the SWATH accumulation. The results were shown as three output files containing AUC of the ions, the summed intensity of peptides for protein and the summed intensity of ions for the peptide. All SWATH Acquisition data were processed using the SWATH Acquisition MicroApp 2.0 in PeakView® Software.

2.7. Data analysis

Data was processed with Protein Pilot Software v. 5.0 (AB SCIEX, Foster City, CA) utilizing the Paragon and Progroup Algorithm. Analysis was also done using the integrated tools in Protein Pilot at the 1% false discovery rate (FDR).

2.8. Gene ontology and pathway enrichment

In this study *Klebsiella pneumonia* subsp. pneumoniae (strain ATCC 700721 / MGH 78578) was used as a reference strain to carry out all functional annotation studies. The proteins that showed an alignment identity ≥50% over 80% of sequence length were assigned as its homolog and annotated with UniProtKB accession number which have used for the subsequent functional studies. For functional annotation, we used slim version of Gene ontology (GO) term; obtained from Gene Ontology Consortium (http://www.geneontology.org/) [19]. KEGG database was used to assign the metabolic pathway to the proteins [20].

2.9. Integration of a protein–protein interaction (PPI) network

To find the interaction partners, protein–protein interaction (PPI) information obtained by STRING database v10.0 (http://www.string-db.org/) [21–26] STRING based PPI established by experimental studies as well as by genomic analysis like domain fusion, phylogenetic profiling high-throughput experiments, co-expression studies and gene neighborhood. In the present study interactions having score ≥0.4 were considered as significant.

3. Results

3.1. LC-MS/MS based proteins identification

In this study we have grown the carbapenem resistant isolate under meropenem drug stress (Sub-MIC: 32). Further we have identified the proteome of the resistant bacteria by LC-MS/MS and using SWATH workflow; 1156 proteins were quantified at 1% FDR. Among them, 52

Table 1
Details of the over expressed proteome (> 10 folds) under meropenem stress in *Klebsiella pneumonia* clinical isolates (NDM-4).

S.No.	Protein name	Accession number	Log fold change vs. *P*-value
1	Elongation factor Tu	A6TEX7	66.69
2	Branched-chain alpha-keto acid dehydrogenase subunit E2	A0A2A5PUJ8	36.43
3	Elongation factor G	A6TEX8	34.35
4	Methionine adenosyltransferase	A6TDV1	26.71
5	Type I restriction-modification system,DNA-methyltransferase subunit M	W1BCK1	26.65
6	Beta-hexosaminidase	A6T7G6	26.48
7	Alpha-ketoacid dehydrogenase subunit beta	A0A2A5PUI0	26.36
8	30S ribosomal protein S11	A6TEV0	24.57
9	Beta-lactamase	A6TIL8	22.72
10	Type I restriction-modification system,specificity subunit S	W1BEE5	21.73
11	Metallo-beta-lactamase NDM-5	U5TTL2	20.55
12	Outer membrane protein C	A6TBT2	19.39
13	50S ribosomal protein L5	A6TEW0	19.20
14	Carbamoyl-phosphate synthase (glutamine-hydrolyzing)	A6T4I0	18.94
15	30S ribosomal protein S4	A6TEU9	18.43
16	DNA-binding protein	A6TAK8	17.94
17	30S ribosomal protein S13 (Fragment)	A6TEV1	17.22
18	UDP-glucose 6-dehydrogenase	A6TBE0	16.98
19	30S ribosomal protein S7	A6TEX9	16.68
20	Formate *C*-acetyltransferase	A6T6Z6	16.47
21	50S ribosomal protein L14	A6TEW2	16.22
22	tRNA (cytidine/uridine-2-O-)-methyltransferase TrmJ	A6TCF3	15.71
23	30S ribosomal protein S17	A6TEW3	15.68
24	Acetate kinase	A6TBY3	15.43
25	Metallo-beta-lactamase	A0A0K2DRS9	15.36
26	50S ribosomal protein L2	A6TEW9	15.15
27	Iron-sulfur cluster carrier protein	W1GPR1	14.65
28	Bifunctionalpolymyxin resistance protein ArnA	A6TF98	14.40
29	30S ribosomal protein S13 (Fragment)	A6TEV1	14.30
30	Catabolite control protein A	A0A2A5PY42	13.76
31	Elongation factor G	W1BCD5	13.64
32	50S ribosomal protein L15	A6TEV3	13.63
33	Biotin carboxylase of acetyl-CoA carboxylase	A6TES3	13.54
34	tRNA (guanine-N(7)-)-methyltransferase	A6TDW8	13.14
35	RibonucleaseVapC	A6TIM9	12.95
36	mRNA interferase	W1 AM39	12.88
37	50S ribosomal protein L17	A6TEU7	12.81
38	Transaldolase	A6T4E8	12.65
39	Exoribonuclease II	A6T7Z4	12.46
40	3-deoxy-manno-octulosonate cytidylyltransferase	A6T710	12.16
41	CTP synthase	A6TD54	11.95
42	Carbamoyl-phosphate synthase large chain	W1BCF0	11.79
43	NADH oxidoreductase	A6T6X1	11.67
44	Dihydropteroate synthase	A6TIZ0	11.65
45	Chromosome (Plasmid) partitioning protein ParB	A6TIC7	11.37
46	Glycine–tRNA ligase beta subunit	A6TFH4	10.85
47	GTP-binding protein TypA/BipA	A6TG74	10.83
48	Catabolite repressor-activator	A6T4M3	10.59
49	Protein RecA	A6TCW1	10.32
50	DNA-binding protein	W1AQL9	10.10
51	Exoribonuclease 2	W1B6L8	10.06
52	Inositol-1-monophosphatase	A6TCF4	10.05

proteins were over expressed up to ten log folds change vs. *p*-value and tabulated in the Table 1. The level of over-expression has been represented as the log folds change vs. p-value ratio. In-depth analysis of over expressed 52 proteins suggested that these belong to four major groups. These proteins involved in process of protein translational machinery complex, DNA/RNA modifying enzymes or proteins, proteins involved in intermediary metabolism & respiration and proteins involved in carbapenems drugs cleavage, modifications and transport.

Table 2
List of 38 DEGs in response to meropenem in *Klebsiella pneumoniae* subsp. *pneumonia* (strain ATCC 700721 / MGH 78578).

S. No.	Gene name	Protein name	Accession number	Log fold change vs. P-value
1	tufA	Elongation factor Tu	A6TEX7	66.69
2	fusA	Elongation factor G	A6TEX8	34.35
3	metK	Methionine adenosyltransferase	A6TDV1	26.71
4	nagZ	Beta-hexosaminidase	A6T7G6	26.48
5	rpsK	30S ribosomal protein S11	A6TEV0	24.57
6	bla	Beta-lactamase	A6TIL8	22.72
7	ompC	Outer membrane protein C	A6TBT2	19.39
8	rplE	50S ribosomal protein L5	A6TEW0	19.2
9	carB	Carbamoyl-phosphate synthase (glutamine-hydrolyzing)	A6T4I0	18.94
10	rpsD	30S ribosomal protein S4	A6TEU9	18.43
11	hns	DNA-binding protein	A6TAK8	17.94
12	rpsM	30S ribosomal protein S13 (Fragment)	A6TEV1	17.22
13	ugd	UDP-glucose 6-dehydrogenase	A6TBE0	16.98
14	rpsG	30S ribosomal protein S7	A6TEX9	16.68
15	pflB	Formate *C*-acetyltransferase	A6T6Z6	16.47
16	rplN	50S ribosomal protein L14	A6TEW2	16.22
17	trmJ	tRNA (cytidine/uridine-2-O-)-methyltransferase, TrmJ	A6TCF3	15.71
18	rpsQ	30S ribosomal protein S17	A6TEW3	15.68
19	ackA	Acetate kinase	A6TBY3	15.43
20	rplB	50S ribosomal protein L2	A6TEW9	15.15
21	arnA	Bifunctionalpolymyxin resistance protein ArnA	A6TF98	14.4
22	rplO	50S ribosomal protein L15	A6TEV3	13.63
23	accC	Biotin carboxylase of acetyl-CoA carboxylase	A6TES3	13.54
24	trmB	tRNA (guanine-N(7)-)-methyltransferase	A6TDW8	13.14
25	mck	RibonucleaseVapC	A6TIM9	12.95
26	rplQ	50S ribosomal protein L17	A6TEU7	12.81
27	talB	Transaldolase	A6T4E8	12.65
28	rnb	Exoribonuclease II	A6T7Z4	12.46
29	kdsB	3-deoxy-manno-octulosonate cytidylyltransferase	A6T710	12.16
30	pyrG	CTP synthase	A6TD54	11.95
31	hcr	NADH oxidoreductase	A6T6X1	11.67
32	sul	Dihydropteroate synthase	A6TIZ0	11.65
33	sopB	Chromosome (Plasmid) partitioning protein ParB	A6TIC7	11.37
34	glyS	Glycine–tRNA ligase beta subunit	A6TFH4	10.85
35	bipA	GTP-binding protein TypA/BipA	A6TG74	10.83
36	fruR	Catabolite repressor-activator	A6T4M3	10.59
37	recA	Protein RecA	A6TCW1	10.32
38	suhB	Inositol-1-monophosphatase	A6TCF4	10.05

3.2. Functional analysis of meropenem- responsive proteins

Among 52 over expressed proteins 38 proteins were enriched by using the *Klebsiella pneumonia* subsp. pneumoniae (strain ATCC 700721/MGH 78578) as a reference strain to carry out all functional annotation studies (Table 2). Analysis of functional enrichment revealed that the most abundant biological process (BP), among the up-regulated proteins, was “translation”. 14 of 38 up-regulated proteins were found to be involved in translation. Other major processes in decreasing order of representation were biosynthetic process, cellular nitrogen compound metabolic process, and carbohydrate metabolic process. 11, 11 and 7 proteins were part of these processes respectively (Fig. 1a). In molecular function (MF) category, the upregulated proteins

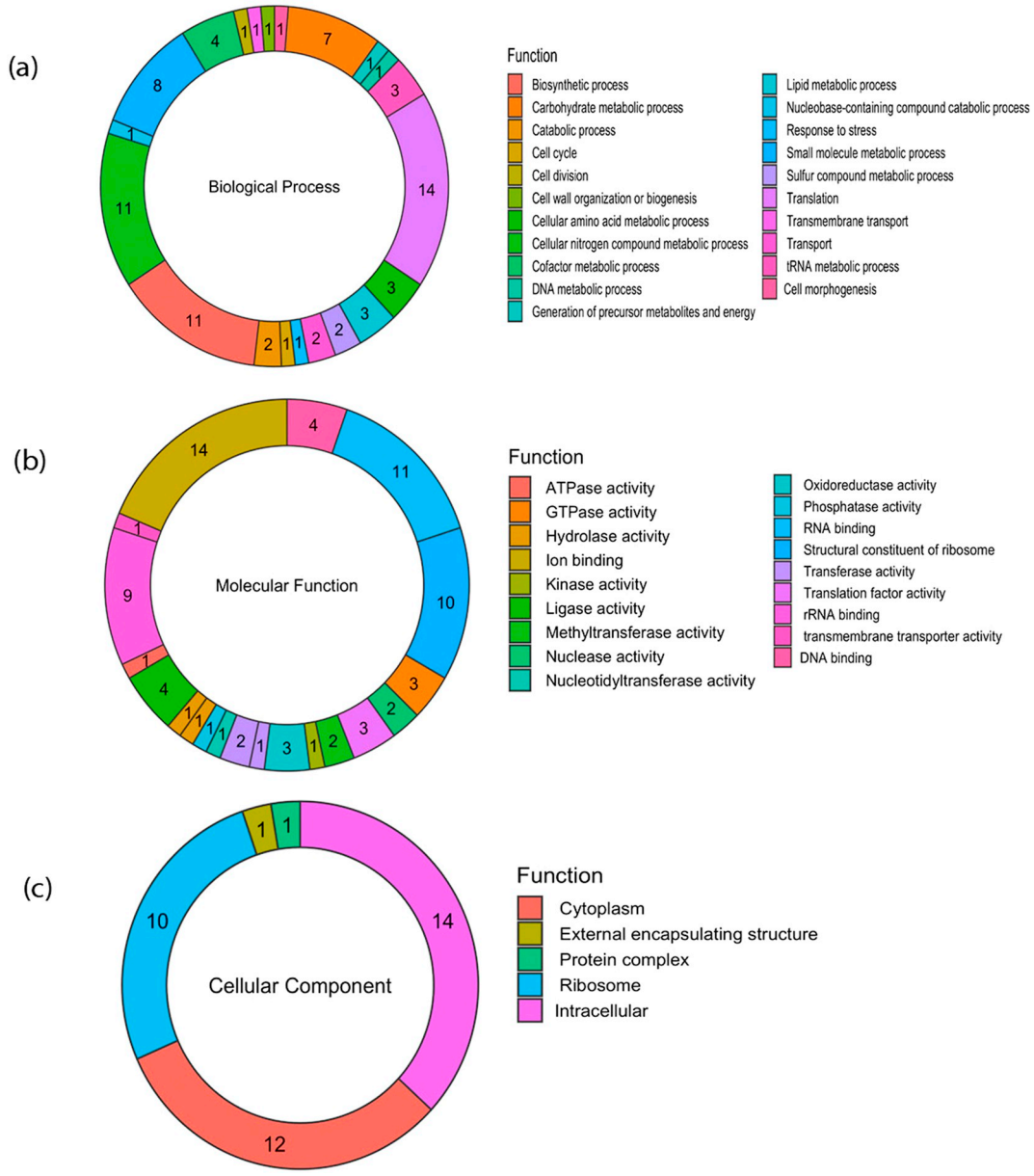

Fig. 1. Functional categories of the upregulated proteins obtained from MS/MS experiments. The 38 proteins were broadly classified based on GO analysis: (a) biological process (b) molecular function, and (c) cellular component.

were enriched in “ion binding” (14 of 38), “RNA binding” (11 of 38), “structural constituent of ribosome” (10 of 38) and “rRNA binding” (9 of 38) (Fig. 1b). Cell component (CC) enrichment analysis showed that the up-regulated genes were significantly enriched in the “intracellular” (14 of 38), “cytoplasm” (12 of 38) and “ribosome” (10 of 38) (Fig. 1c).

3.3. KEGG pathway analysis

The most significantly enriched pathways, in which the over expressed proteins belong, was analyzed by KEGG shown in Fig. 2 and Table 3. The up-regulated proteins were enriched in ribosome (10 of 38), microbial metabolism in diverse environments (4 of 38), beta-lactam resistance (3 of 38), biosynthesis of secondary metabolites (3 of 38), propanoate metabolism (3 of 38), amino sugar and nucleotide sugar metabolism (3 of 38), carbon metabolism (3 of 38), and pyruvate metabolism (3 of 38).

3.4. Protein-protein interaction network (PPIs)

Among the 38 proteins, four plasmid encoded proteins namely ribonuclease VapC (RNaseVapC), dihydropteroate synthase (DHPS), beta-lactamase (bla) and RepFiA replicon did not show any interaction. Moreover, quantitative interaction network was created for remaining 34 up-regulated genes and total of 1430 nodes and 4099 edges related to these proteins were obtained from the Cytoscape (Fig. 3).

4. Discussion

Emergence of carbapenem resistant *K. pneumoniae* has continuously risen and significantly threatens the world. Over expression of carbapenemases and loss of porins are the well known phenomena of carbapenem resistance although these elucidated resistome mechanisms are still in fragmentary phase. In the present study we have found the over expression of a panel of proteins (52 proteins) in the meropenem induced culture (sub-MIC) which might play a pivotal role in

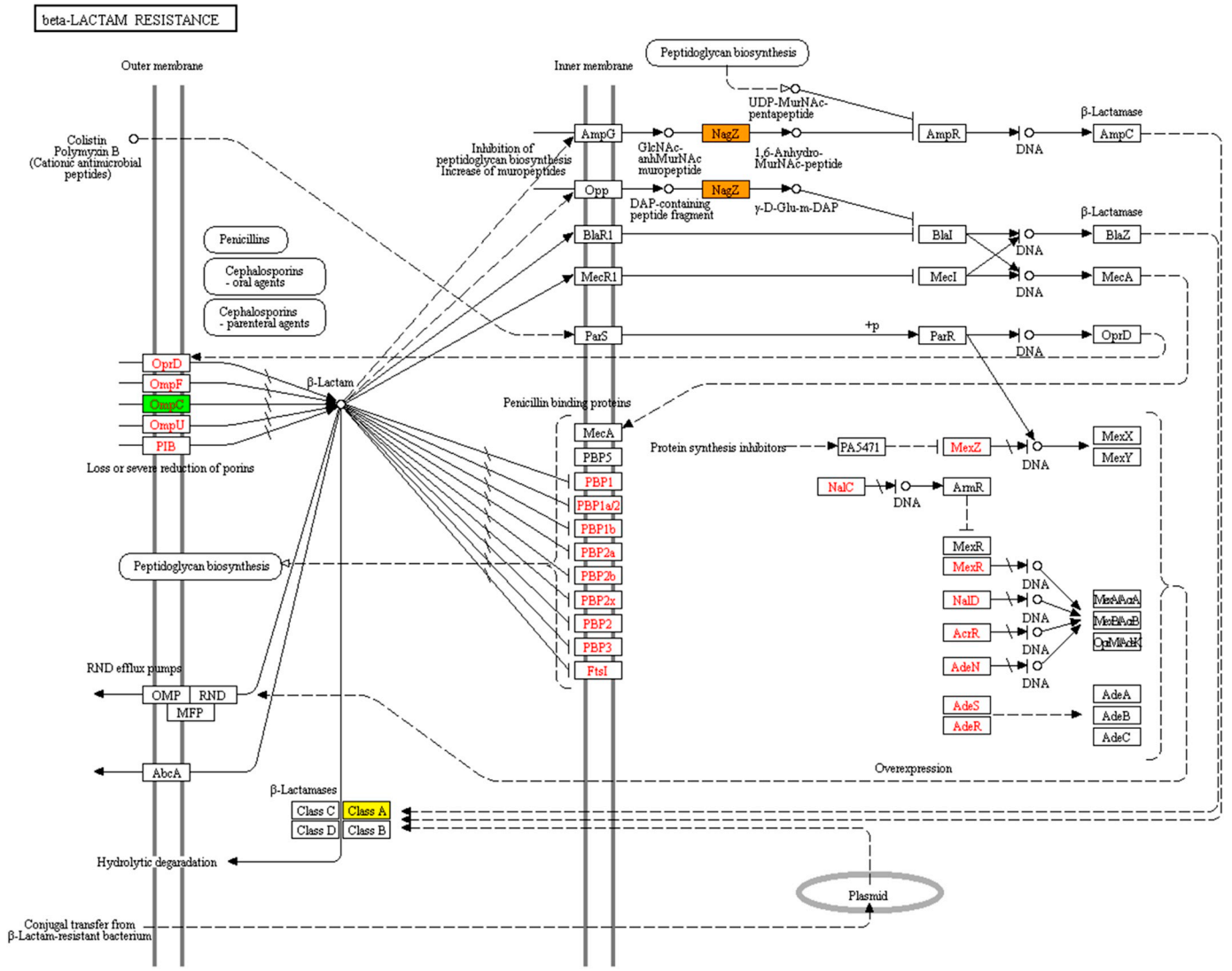

Fig. 2. beta-lactam resistance pathway manually curated using KEGG. Three over expressed genes in nagZ (Orange), bla (yellow) and ompC (green) were successfully mapped on to this pathway. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

carbapenem resistance of *K. pneumoniae* (NDM-4). These 52 proteins belong to four major classes such as protein translational machinery complex, DNA/RNA modifying enzymes or proteins, proteins involved in energy metabolism & intermediary metabolism and proteins involved in carbapenems drugs cleavage, modifications and transport.

4.1. Ribosomal protein translational machinery complex formation and regulation

A panel of proteins involved in the ribosomal proteins translational machinery complex formation and its regulation process were found to be over expressed in carbapenem resistant *K. pneumoniae* clinical strains induced with the meropanam. These proteins are Elongation factor Tu, Elongation factor G, 30S ribosomal protein S4, 30S ribosomal protein S7, 30S ribosomal protein S11, 30S ribosomal protein S13, 30S ribosomal protein S17, 50S ribosomal protein L2, 50S ribosomal protein L5, 50S ribosomal protein L14, 50S ribosomal protein L15 and 50S ribosomal protein L17. Elongation factor Tu (EF-Tu) is a GTP binding protein, involved in elongation of peptide chain during protein translation [27,28]. Elongation factor G (EF-G) regulates the peptide elongation as well as ribosome recycling which are the crucial steps of translation [29]. We also reported the over expressions of EF-Tu and EF-G in the drug resistant tuberculosis [13,30]. Remaining 30S, 50S ribosomal proteins along with EF-Tu and EF-G cumulatively make a ribosomal protein translational machinery complex which involved in proteins biosynthesis.

4.2. DNA/RNA modifying/degradation enzymes and related proteins

Over expression of the DNA/RNA modifying enzymes and its related proteins were observed in the study. These proteins are Type I restriction-modification system-DNA-methyltransferase subunit M, Type I restriction-modification system-specificity subunit S, RecA protein, DNA-binding protein, CTP synthase, Exoribonuclease-2, tRNA (cytidine/uridine-2'-*O*-)-methyltransferase (TrmJ), tRNA (guanine-N (7)-)-methyltransferase, Glycine–tRNA ligase beta subunit, Ribonuclease (VapC) and mRNA interferase (pemK). Meropenam pressure induced stress and activates the repairs systems (SOS system) for DNA and RNA repair. RecA protein involved in repair processes of damaged DNA, DNA recombination, and SOS response [31] which trigger the cells to re-enter into the cell cycle. Methylation of the DNA and RNA is another mechanism of defence through which bacteria save themselves by any stress and drug stress. Therefore, the over expression of these tRNA (cytidine/uridine-2'-O-)-methyltransferase (TrmJ) and tRNA (guanine-N (7)-)-methyltransferase might be contributed to resistance. Ribonuclease (VapC) is the toxin VapC of the two component toxin-antitoxin

Table 3
KEGG pathway analysis of DEGs associated with *Klebsiella pneumonia* subsp. *pneumoniae* (strain ATCC 700721/MGH 78578).

Category	KEGG Team	Count	Protein list
Up-regulated	Ribosome	10	rplQ,rpsD,rpsK,rpsM,rplO,rplE,rplN,rpsQ,rplB,rpsG
	Microbial metabolism in diverse environments	4	talB,pflB,ackA,accC
	beta-Lactam resistance	3	nagZ,ompC,bla
	Biosynthesis of secondary metabolites	3	talB,metK,accC
	Propanoate metabolism	3	pflB,ackA,accC
	Amino sugar and nucleotide sugar metabolism	3	nagZ,ugd,yfbG
	Carbon metabolism	3	talB,ackA,accC
	Pyruvate metabolism	3	pflB,ackA,accC
	Biosynthesis of amino acids	2	talB,metK
	Pyrimidine metabolism	2	carB,pyrG
	Biosynthesis of antibiotics	2	talB,accC
	Lipopolysaccharide biosynthesis	1	kdsB
	Aminoacyl-tRNA biosynthesis	1	glyS
	Methane metabolism	1	ackA
	Inositol phosphate metabolism	1	suhB
	Pentose and glucuronateinterconversions	1	ugd
	Cationic antimicrobial peptide (CAMP) resistance	1	yfbG
	Taurine and hypotaurine metabolism	1	ackA
	Alanine, aspartate and glutamate metabolism	1	carB
	Folate biosynthesis	1	sul
	Two-component system	1	ompC
	Ascorbate and aldarate metabolism	1	ugd
	Pentose phosphate pathway	1	talB
	Homologous recombination	1	recA
	Cysteine and methionine metabolism	1	metK
	Streptomycin biosynthesis	1	suhB
	Fatty acid biosynthesis	1	accC
	Fatty acid metabolism	1	accC
	Butanoate metabolism	1	pflB

(TA)/VapBC system in which toxins inhibit the cell growth and antitoxins repress translation of the toxin genes. In Gram positive and Gram negative bacteria, VapC ectopic expression inhibits cell growth by inhibiting the global rate of translation [32,33]. STRING analysis revealed that toxin VapC directly interact to VapB antitoxins which may neutralized the cognate VapCs through protein-protein interaction. The mRNA interferase/pemK is also the toxin of a type II toxin-antitoxin (TA) system, encoded by plasmid R100 and it is the homologous of MazF, an *E.coli* toxin [34]. They inhibit the translation by mRNA cleavage. Exoribonuclease-2 (rnb) also involved in mRNA degradation. It selectively hydrolyzes single-stranded polyribonucleotides processively in the 3′ to 5′ direction. In the present study over expression of methyltransferase, RecA proteins, VapC, mRNA interferase/pemK and exoribonuclease-2 might be involved in DNA/RNA repair, maturation and turnover.

4.3. Proteins involved in drugs modifications or drugs degradation and resistance

Higher level enzymes production (AmpC Beta-lactamase, penicillinase, Extended Spectrum Beta-Lactamase (ESBL), carbapenemase, Beta-lactamase, and Metallo-beta-lactamase) have reported to contribute in the carbapenem resistance [35,36]. ArnA/Bifunctional polymyxin resistance protein is the first enzyme specific to the lipid A-Ara4N pathway. It modifies the lipid-A through 4-amino-4-deoxy-L-arabinose (Ara4N) and makes the bacteria to resist the antibiotics such as polymyxin and other [37]. This sub-pathway is part of the pathway UDP-4-deoxy-4-formamido-beta-L-arabinose biosynthesis, nucleotide-sugar biosynthesis, lipopolysaccharide biosynthesis and bacterial outer membrane biogenesis. Outer membrane proteins played an important role in the carbapenem resistance. Brinkworth et al., 2015 reported a panel of outer membrane proteins and exportins by 2DE and LC-MS/MS, however outer membrane protein C was not reported [38]. In our study outer membrane protein C was over expressed, sequence similarity reveled that it matches 100% to porin OmpK35 which is involved in resistance [39]. Therefore, we hypothesized that over expression of beta lactamases degrade the drug inside the cell so that cell unable to get signal for the repression of ompC. This study revealed the cumulative effect of beta lactamases, bifunctional polymyxin resistance protein ArnA, and ompC which might be contributed into carbapenems resistance through different mechanisms and may generate the clue for future research.

4.4. Proteins involved in energy metabolism and intermediary metabolism

In this study we have identified a penal of proteins belong to energy and intermediary metabolism. These are branched-chain alpha-keto acid dehydrogenase subunit E2, 3-deoxy-manno-octulosonate cytidylyltransferase, Methionine adenosyltransferase, Beta-hexosaminidase, Alpha-ketoacid dehydrogenase subunit beta, Carbamoyl-phosphate synthase (glutamine-hydrolyzing), UDP-glucose 6-dehydrogenase, Acetate kinase, Catabolite control protein A, Biotin carboxylase of acetyl-CoA carboxylase, Transaldolase, 3-deoxy-manno-octulosonate cytidylyltransferase, CTP synthase, Carbamoyl-phosphate synthase large chain, NADH oxidoreductase, Dihydropteroate synthase, GTP-binding protein TypA/BipA, Catabolite repressor-activator, Carbamoyl-phosphate synthase large chain and Inositol-1-monophosphatase. These proteins involved in generation of various intermediate metabolites of many metabolic pathways such as amino acid metabolism, lipid metabolism, nucleotide metabolism, tetrahydrofolate biosynthesis, carbon catabolism and glucose or energy metabolism. CTP synthase is an essential enzyme involved in formation of CTP from UTP and ATP using glutamate as nitrogen source [40]. Methionine adenosyltransferase catalyzes the formation of S-adenosylmethionine (AdoMet) from methionine and ATP.

In the present study we also used system biology approach for integration of identified proteins. Among 52 over expressed proteins, 38 DEGs are successfully mapped on *Klebsiella pneumoniasubsp. pneumoniae* (strain ATCC 700721 / MGH 78578) strain representative proteome. Among 38 upregulated proteins, a major chunk (> 18%) were ribosomal protein encoding genes (rplB, rplE, rpsM, rplN, rpsD, rpsK, rplO, rpsQ, rpsG, rplQ) and genes involved in tRNA metabolism (trmB, trmJ

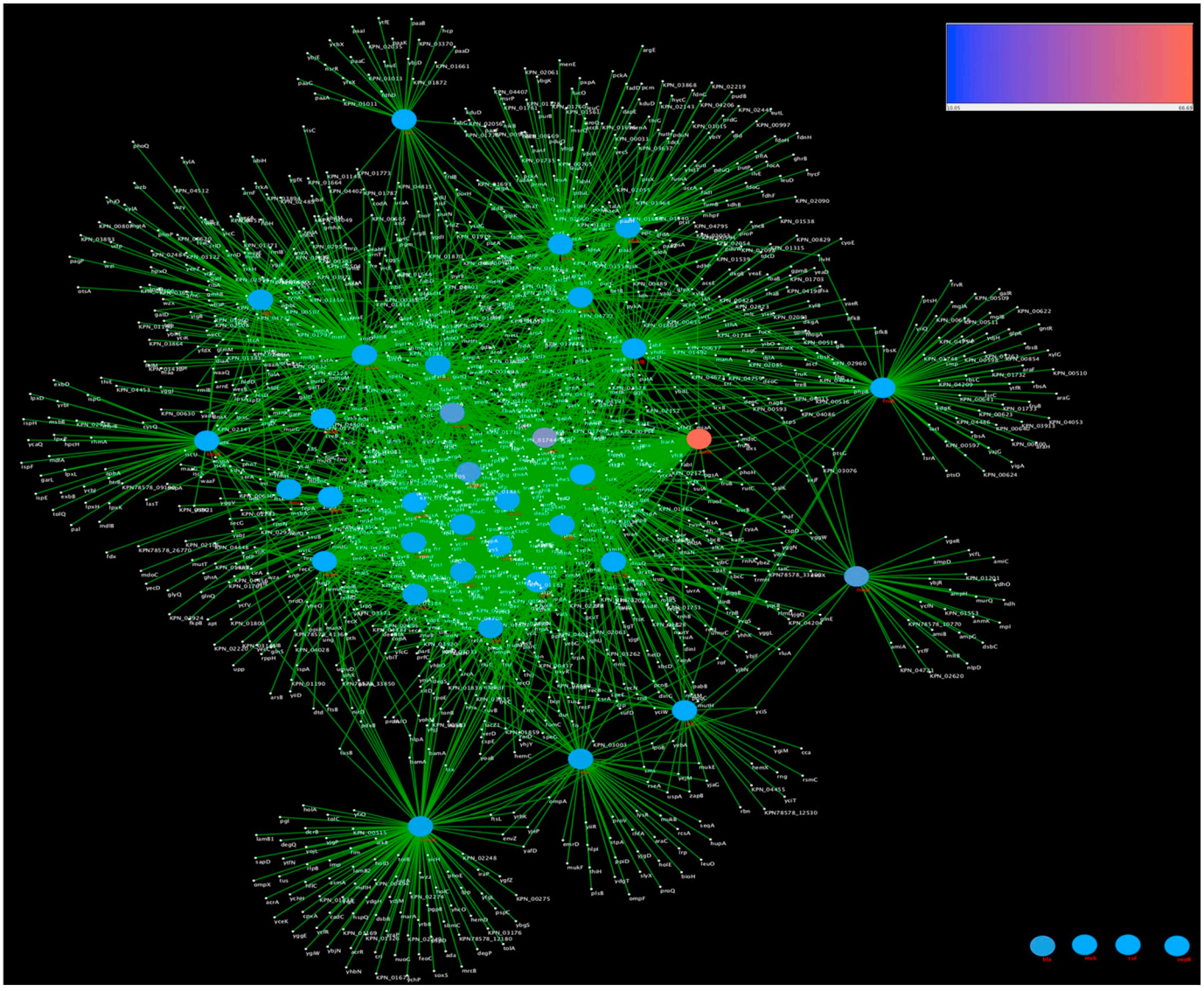

Fig. 3. The interaction networks for over expressed genes in *K. pneumonia.* Shown are changes of gene expression under meropenem stress condition, data derived from MS/MS and visualized in Cytoscape [45]. Color scale denotes log fold vs. *p*-value in gene expression.

and glyS) (Table 2). The GO-term analysis of over-expressed proteins showed that they were mainly involved in carbohydrate metabolic process, translation, biosynthetic process, cellular nitrogen compound metabolic process and small molecule metabolic process. At level of molecular function these proteins carry out RNA binding, structural constituent of ribosome, r-RNA and ion binding. The analysis also showed that these proteins were part of ribosome; cytoplasm and intracellular space (Fig. 1).The previous report suggested that antibiotics stress mostly influences the metabolic pathways and ribosomal subunits expression [41]. Interestingly, our results from high throughput LC-MS/MS confirm the assumption that the over expression of these proteins may be a compensative mechanism for bacteria to defence antibiotics. In the present work we tried to characterize the current expression profile in response to *K. pneumoniae* in presence of meropenem. We mapped the metabolic pathways of carbapenem resistant clinical strain (NDM-4) through over expressed proteins to understand the mystery of beta-lactam resistance (Fig. 2). Analysis of metabolic pathways using KEGG has also concurred with the GO-term analysis. We found meropenem influence certain pathways namely, ribosome component, microbial metabolism in diverse environments, beta-lactam resistance, biosynthesis of secondary metabolites, propanoate metabolism, amino sugar and nucleotide sugar metabolism, carbon metabolism and pyruvate metabolism (Table 3). These metabolic pathways contribute to a complex biological process (for e.g. pyruvate metabolism) that promotes some of antibiotics uptake [41]. Here, we hypothesized that overexpressed proteins were mostly affected those metabolic pathways which involved in antibiotic-resistant mechanism while it is not an antibiotic-mediated cell death process. We also constructed a PPI network of 38 over expressed proteins (Fig. 3). Our analysis revealed that some upregulated proteins were considered hub protein which indicates their important role in translation, metabolism and beta-lactam drug resistance [42]. One protein, namely recA, had degree of connectivity 320 (Fig. 3) and contributes in the homologous recombination pathway (Table 2). Earlier studies also reported the role of RecA in increased tolerance to antibiotic treatment by enhancing DNA repair that occurs either directly by antibiotic-induced DNA damage or indirectly from metabolic and oxidative stress [43]. Furthermore, our experimental results showed that recA was also highly expressed in meropenem stress condition. Therefore, we can hypothesize that this protein might contribute to the progression of antibiotic resistance with drug treatment. There are several genes that are involved in regulation of ampC expression. nagZ is one of the important gene which produces β-*N*-acetylglucosaminidase. Degradation product of peptidoglycan cell wall, GlcNAc-1,6-anhydromuropeptides are generated in periplasm.

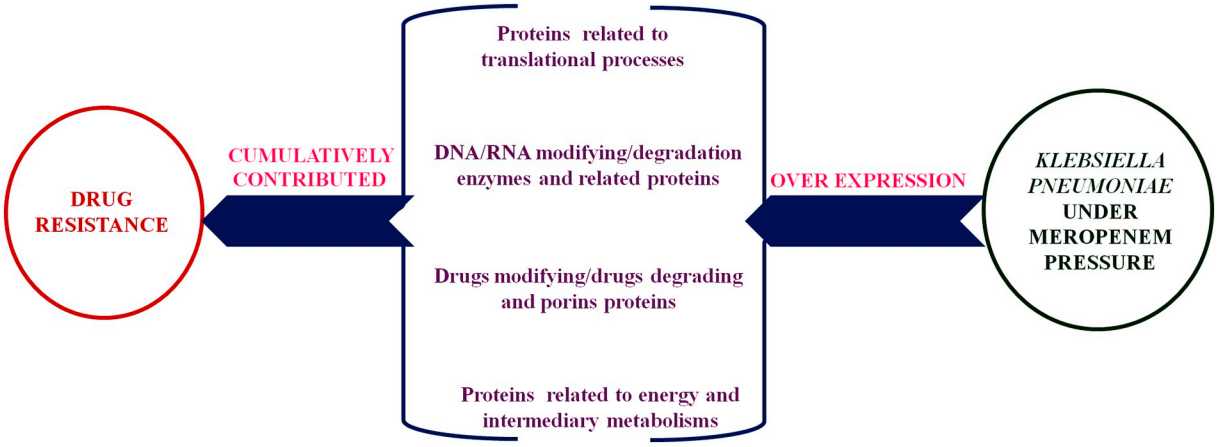

Fig. 4. Representative model shown the summarised mechanism based on the outcome of the study.

They are internalized into cytoplasm by ampG encoded inner membrane permease. In the cytoplasm NagZ transforms GlcNAc-1,6-anhydromuropeptides into 1,6-anhydromuropeptides, which in-turn induces AmpC synthesis by interacting with the LysR-type transcriptional regulator AmpR [44]. Under normal condition, 1, 6-anhydromuropeptides are processed by the *N*-acetyl-anhydromuramyl-L-alanine amidase AmpD, blocking *ampC* induction. Whereas under drug stress or beta-lactamase inducers (for e.g. meropenem), large amounts of muropeptides are generated and collected into the cytoplasm, which leads to the AmpR-mediated initiation of *ampC* expression [44]. Hence, we can foresee that three genes viz. nagZ, ompC and bla have vital role in antibiotic resistance and can be used to exploit its resistance mechanism. On the basis of these over expressed proteins, we proposed a model (Fig. 4) which may explore the new mechanism of carbapenem resistance. In the present study we suggest that cumulative effect of over expressed proteins, mapped pathways and interactive proteins partner of the over expressed proteins might be involved in carbapenem drugs cleavage, modifications and transport which may contribute in the carbapenem resistance.

5. Conclusion and future prospects

To concise, present study focused the whole proteome of carbapenem resistant *K. pneumoniae* clinical isolate (NDM-4) under meropenem stress through proteomic and bioinformatic approaches. These 52 over expressed proteins belong to four major groups such as protein translational machinery complex, DNA/RNA modifying enzymes or proteins, proteins involved in carbapenems cleavage, modifications, transport, energy metabolism and intermediary metabolism related proteins which suggested their direct or indirect role in the carbapenem resistance. Pathways analysis (GO and KEGG) and proteins-proteins interaction suggested that these 52 over expressed proteins, pathways and their interactive proteins cumulatively contributed to the survival of bacteria and meropenem resistance through various mechanisms. These proteins targets and their pathways might open a new vistas of drug resistance and could be used for the development of novel therapeutics against the resistance. Therefore the situation of emergence of bad-bugs could be prevented.

Authors' contribution

DS design the concept, experimented and wrote the manuscript. AG and MK have done the system biology analysis. AUK designed the problem and guided the study and finalized the manuscript.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgements and funding

SERB is gratefully acknowledged for providing fellowship and funds to DS (SERB-N-PDF/2016/001622) to work at IBU-AMU Aligarh. AG is JRF, supported by ICMR (3/1/3 JRF/LS/HRD-32262). Authors also acknowledge the SCIEX Pvt. Ltd. Gurgaon, India for proteomics facility support. Authors also acknowledge Dr. Neelja Singhal for helped in data analysis.

References

[1] Centers for Disease Control and Prevention, Antibiotic resistance threats in the United States, (2013) http://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf2.
[2] World Health Organization, Factsheet on Antimicrobial Resistance, http://www.who.int/mediacentre/factsheets/fs194.
[3] D.L. Paterson, Recommendation for treatment of severe infections caused by Enterobacteriaceae producing extended-spectrum beta-lactamases (ESBLs), Clin. Microbiol. Infect. 6 (2000) 460–463.
[4] D.L. Paterson, R.A. Bonomo, Extended-spectrum beta-lactamases: a clinical update, Clin. Microbiol. Rev. 18 (4) (2005) 657–686.
[5] L. Martínez-Martínez, A. Pascual, S. Hernández-Allés, D. Alvarez-Díaz, A.I. Suárez, J. Tran, et al., Roles of β-lactamases and porins in activities of carbapenems and cephalosporins against *Klebsiella pneumoniae*, Antimicrob. Agents Chemother. 43 (1999) 1669–1673.
[6] G.A. Jacoby, D.M. Mills, N. Chow, Role of beta-lactamases and porins in resistance to Ertapenem and other beta-lactams in *Klebsiella pneumoniae*, Antimicrob. Agents Chemother. 48 (2004) 3203–3206.
[7] A.M. Queenan, K. Bush, Carbapenemases: the versatile β-lactamases, Clin. Microbiol. Rev. 20 (3) (2007) 440–458, https://doi.org/10.1128/CMR.00001-07.
[8] A. Loli, L.S. Tzouvelekis, E. Tzelepi, A. Carattoli, A.C. Vatopoulos, P.T. Tassios, V. Miriagou, Sources of diversity of carbapenem resistance levels in *Klebsiella pneumoniae* carrying blaVIM-1, J. Antimicrob. Chemother. 58 (3) (2006) 669–672.
[9] R.P. Ambler, A.F. Coulson, J.M. Frère, J.M. Ghuysen, B. Joris, M. Forsman, R.C. Levesque, G. Tiraby, S.G. Waley, A standard numbering scheme for the class A beta-lactamases, Biochem J. 276 (Pt 1) (1991 May 15) 269–270.
[10] C. Freiberg, H. Brötz-Oesterhelt, H. Labischinski, The impact of transcriptome and proteome analyses on antibiotic drug discovery, Curr. Opin. Microbiol. 7 (5) (2004) 451–459.
[11] K.V. dos Santos, C.G. Diniz, C. Veloso Lde, H.M. de Andrade, S. Giusta Mda, F. Pires Sda, et al., Proteomic analysis of E.coli with experimentally induced resistance to piperacillin/tazobactam, Res. Microbiol. 161 (2010) 268–275.
[12] S. Qayyum, D. Sharma, D. Bisht, A.U. Khan, Protein translation machinery holds a key for transition of planktonic cells to biofilm state in *Enterococcus faecalis:* a proteomic approach, Biochem. Biophys. Res. Commun. 474 (2016) 652–659.
[13] M. Lata, D. Sharma, N. Deo, P.K. Tiwari, D. Bisht, K. Venkatesan, Proteomic analysis of ofloxacin-mono resistant *Mycobacterium tuberculosis* isolates, J. Proteomics 127 (Part A) (2015) 114–121.
[14] A. Khan, D. Sharma, M. Faheem, D. Bisht, A.U. Khan, Proteomic analysis of a carbapenem-resistant *Klebsiella pneumoniae* strain in response to meropenem stress, J Glob Antimicrob Resist. 8 (2017) 172–178.
[15] D. Sharma, D. Bisht, A.U. Khan, Potential alternative strategy against drug resistant tuberculosis: a proteomics prospect, Proteomes 6 (2) (2018) 26.

[16] S. Qayyum, D. Sharma, D. Bisht, A.U. Khan, Identification of factors involved in *Enterococcus faecalis* biofilm under quercetin stress, Microb. Pathog. 126 (2019) 205–211.
[17] P.A. Wayne, Performance standards for antimicrobial susceptibility testing: 24 informational supplement, CLSI M100 (2014) S24.
[18] M.M. Bradford, A rapid and sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein-dye binding, Anal. Biochem. 72 (1976) 248–254.
[19] E. Camon, D. Barrell, V. Lee, E. Dimmer, R. Apweiler, The Gene Ontology Annotation (GOA) Database-an integrated resource of GO annotations to the UniProt Knowledgebase, In Silico Biol. 4 (1) (2003) 5–6.
[20] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[21] D. Sharma, M. Lata, M. Faheem, A.U. Khan, B. Joshi, K. Venkatesan, S. Shukla, D. Bisht, M. Tuberculosis ferritin (Rv3841): potential involvement in Amikacin (AK) & kanamycin (KM) resistance, Biochem. Biophys. Res. Commun. 478 (2) (2016) 908–912.
[22] D. Sharma, D. Bisht, Secretory proteome analysis of streptomycin resistant Mycobacterium tuberculosis clinical isolates, SLAS Discov. 22 (2017) 1229–1238.
[23] D. Sharma, D.M. Bisht, *tuberculosis* hypothetical proteins and proteins of unknown function: Hope for exploring novel resistance mechanisms as well as future target of drug resistance, Front. Microbiol. 8 (2017) 465.
[24] D. Sharma, D. Bisht, Role of bacterioferritin & ferritin in *M.tuberculosis* pathogenesis and drug resistance: a future perspective by interactomic approach, Front. Cell. Infect. Microbiol. 7 (240) (2017).
[25] D. Sharma, R. Singh, N. Deo, D. Bisht, Interactome analysis of Rv0148 to predict potential targets and their pathways linked to aminoglycosides drug resistance: an insilico approach, Microb. Pathog. 121 (2018) 179–183.
[26] D. Sharma, A.U. Khan, Role of cell division protein divIVA in *Enterococcus faecalis* pathogenesis, biofilm and drug resistance: a future perspective by in silico approaches, Microb. Pathog. 125 (2018) 361–365.
[27] W. Ludwig, M. Weizenegger, D. Betzl, E. Leidel, T. Lenz, A. Ludvigsen, D. Möllenhoff, P. Wenzig, K.H. Schleifer, Complete nucleotide sequences of seven eubacterial genes coding for the elongation factor Tu: functional, structural and phylogenetic evaluations, Arch. Microbiol. 153 (3) (1990) 241–247.
[28] M. Sprinzl, Elongation factor Tu: a regulatory GTPase with an integrated effector, Trends Biochem. Sci. 19 (1994) 245–250.
[29] G.R. Willie, N. Richman, W.P. Godtfredsen, J.W. Bodley, Some characteristics of and structural requirements for the interaction of 24,25-dihydrofusidic acid with ribosome-elongation factor g complexes, Biochemistry 14 (1975) 1713–1718.
[30] D. Sharma, B. Kumar, M. Lata, B. Joshi, K. Venkatesan, S. Shukla, D. Bisht, Comparative proteomic analysis of aminoglycosides resistant and susceptible Mycobacterium tuberculosis clinical isolates for exploring potential drug targets, PLoS One 10 (10) (2015) e0139414.
[31] D. Zgur-Bertok, DNA damage repair and bacterial pathogens, PLoS Pathog. 9 (11) (2013) e1003711.
[32] K.S. Winther, K. Gerdes, Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA, Proc. Natl. Acad. Sci. 108 (18) (2011) 7403–7407.
[33] J. Robson, J.L. McKenzie, R. Cursons, G.M. Cook, V.L. Arcus, The vapBC operon from *Mycobacterium smegmatis* is an autoregulated toxin–antitoxin module that controls growth via inhibition of translation, J. Mol. Biol. 390 (3) (2009) 353–367.
[34] Y. Zhang, L. Zhu, J. Zhang, M. Inouye, Characterization of ChpBK, an mRNA interferase from Escherichia coli, J. Biol. Chem. 280 (2005) 26080–26088.
[35] V. Rastogi, P.S. Nirwan, S. Jain, A. Kapil, Nosocomial outbreak of septicaemia in neonatal intensive care unit due to extended spectrum β-lactamase producing *Klebsiella pneumoniae* showing multiple mechanisms of drug resistance, Indian J. Med. Microbiol. 28 (4) (2010) 380.
[36] J.M. Blair, M.A. Webber, A.J. Baylay, D.O. Ogbolu, L.J. Piddock, Molecular mechanisms of antibiotic resistance, Nat. Rev. Microbiol. 13 (1) (2015) 42.
[37] P.Z. Gatzeva-Topalova, A.P. May, M.C. Sousa, Crystal structure and mechanism of the *Escherichia coli* ArnA (PmrI) transformylase domain. An enzyme for lipid A modification with 4-amino-4-deoxy-L-arabinose and polymyxin resistance, Biochemistry 44 (14) (2005) 5328–5338.
[38] A.J. Brinkworth, C.H. Hammer, L.R. Olano, S.D. Kobayashi, L. Chen, B.N. Kreiswirth, F.R. De Leo, Identification of outer membrane and exoproteins of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*, PLoS One 10 (4) (2015) e0123219.
[39] J.H. Chen, L.K. Siu, C.P. Fung, J.C. Lin, K.M. Yeh, T.L. Chen, ... F.Y. Chang, Contribution of outer membrane protein K36 to antimicrobial resistance and virulence in Klebsiella pneumoniae, J. Antimicrob. Chemother. 65 (5) (2010) 986–990.
[40] A. Levitzki, D.E. Koshland Jr., Cytidine triphosphate synthetase. Covalent intermediates and mechanisms of action, Biochemistry 10 (18) (1971) 3365–3371.
[41] X. Lin, L. Kang, H. Li, X. Peng, Fluctuation of multiple metabolic pathways is required for Escherichia coli in response to chlortetracycline stress, Mol. BioSyst. 10 (4) (2014) 901–908.
[42] A. Doménech-Sánchez, V. Javier Benedí, L. Martínez-Martínez, S. Alberti, Evaluation of differential gene expression in susceptible and resistant clinical isolates of Klebsiella pneumoniae by DNA microarray analysis, Clin. Microbiol. Infect. 12 (9) (2006) 936–940.
[43] M.K. Alam, A. Alhhazmi, J.F. DeCoteau, Y. Luo, C.R. Geyer, RecA inhibitors potentiate antibiotic activity and block evolution of antibiotic resistance, Cell Chem. Biol. 23 (3) (2016) 381–391.
[44] L. Zamorano, T.M. Reeve, L. Deng, C. Juan, B. Moyá, G. Cabot, ... A. Oliver, NagZ inactivation prevents and reverts β-lactam resistance, driven by AmpD and PBP 4 mutations, in *Pseudomonas aeruginosa*, Antimicrob. Agents Chemother. 54 (9) (2010) 3557–3563.
[45] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.

Briefings in Bioinformatics, 00(00), 2020, 1–11

doi: 10.1093/bib/bbaa195
Case Study

Discerning novel drug targets for treating *Mycobacterium avium* ss. *paratuberculosis*-associated autoimmune disorders: an *in silico* approach

Anjali Garg, Neelja Singhal and Manish Kumar

Corresponding author: Manish Kumar, Department of Biophysics, University of Delhi South Campus, New Delhi-110021, India.
E-mail: manish@south.du.ac.in; Neelja Singhal, Department of Biophysics, University of Delhi South Campus, New Delhi-110021, India.
E-mail: neelja30@gmail.com

Abstract

Mycobacterium avium subspecies *paratuberculosis* (MAP) exhibits 'molecular mimicry' with the human host resulting in several autoimmune diseases such as multiple sclerosis, type 1 diabetes mellitus (T1DM), Hashimoto's thyroiditis, Crohn's disease (CD), etc. The conventional therapy for autoimmune diseases includes immunosuppressants or immunomodulators that treat the symptoms rather than the etiology and/or causative mechanism(s). Eliminating MAP–the etiopathological agent might be a better strategy to treat MAP-associated autoimmune diseases. In this case study, we conducted a systematic *in silico* analysis to identify the metabolic chokepoints of MAP's mimicry proteins and their interacting partners. The probable inhibitors of chokepoint proteins were identified using DrugBank. DrugBank molecules were stringently screened and molecular interactions were analyzed by molecular docking and 'off-target' binding. Thus, we identified 18 metabolic chokepoints of MAP mimicry proteins and 13 DrugBank molecules that could inhibit three chokepoint proteins *viz.* katG, rpoB and narH. On the basis of molecular interaction between drug and target proteins finally eight DrugBank molecules, viz. DB00609, DB00951, DB00615, DB01220, DB08638, DB08226, DB08266 and DB07349 were selected and are proposed for treatment of three MAP-associated autoimmune diseases namely, T1DM, CD and multiple sclerosis. Because these molecules are either approved by the Food and Drug Administration or these are experimental drugs that can be easily incorporated in clinical studies or tested *in vitro*. The proposed strategy may be used to repurpose drugs to treat autoimmune diseases induced by other pathogens.

Key words: Molecular mimicry; autoimmunity; mimicry interaction partners; metabolic pathways; chokepoints; drug repurposing

Introduction

Mycobacterium avium subspecies *paratuberculosis* (MAP) belongs to the *Mycobacterium avium* complex, which contains mycobacterial species like *M. avium* and *M. intracellulare* [1]. It is an obligate intracellular bacterial pathogen that can cause chronic granulomatous gastroenteritis in ruminants, called Johne's disease [2, 3]. Humans might become infected with MAP through fecal–oral, waterborne, foodborne and zoonotic routes [4–6]. MAP exhibits 'epitope mimicry' with the host peptide(s)/protein(s), which elicits host autoreactive T- or B-cells leading to tissue and/or organ damage and ultimately autoimmune diseases [7]. It is widely believed that autoimmune diseases develop in genetically

Anjali Garg is currently working for PhD in Bioinformatics at the Department of Biophysics, University of Delhi South Campus, New Delhi, India. Her research interest includes functional interactions between proteins in metabolic and signaling pathways and the usage of this information to develop new drug targets.
Neelja Singhal is a senior researcher at the Department of Biophysics, University of Delhi South Campus, New Delhi. Her area of interest includes antimicrobial resistance and microbial pathogenesis.
Manish Kumar is an assistant professor at the Department of Biophysics, University of Delhi South Campus, New Delhi, India. His main areas of research are the development of bioinformatics prediction and annotation pipelines; and usage of computational methods to study microbial pathogenesis and antimicrobial resistance.
Submitted: 13 April 2020; **Received (in revised form):** 24 June 2020

predisposed individuals because of some environmental triggers [8, 9]. Host genetic factors like polymorphisms in the CARD15 (NOD2) [2, 3, 10], SLC11a1 (NRAMP1) [9, 11, 12], LRRK2 [13, 14], PTPN2/22 [15] and VDR [16] genes have been associated with MAP and various diseases like Crohn's disease (CD) [2, 12], Blau syndrome [2], multiple sclerosis (MS) [2, 17], Hashimoto's thyroiditis [18–20], Parkinson's disease [13, 21], rheumatoid arthritis [11, 15, 22] and type 1 diabetes mellitus (T1DM) [16, 23]. Several studies have suggested MAP as an environmental trigger for multiple human autoimmune diseases such as MS, T1DM, Hashimoto's thyroiditis, sarcoidosis, CD and Blau syndrome [6, 18, 24–27]. MAP was reportedly also detected from the body samples of patients suffering from autoimmune T1DM and MS [8, 9, 28, 29].

The conventional therapy for autoimmune diseases has been the usage of immunosuppressants or immunomodulators, which treat symptoms rather than the etiology and/or the causative mechanism(s). Even though 60–70% of the patients initially respond to immunosuppression, in many cases the patients show subsequent clinical remission or relapse of the autoimmune disease [28, 30]. Studies have indicated that immunosuppressants conventionally used for treating MAP-associated autoimmune diseases actually inhibited the growth of MAP *in vitro* [31, 32]. Interestingly, the drug regimen consisting of clarithromycin, rifabutin and clofazimine (antimycobacterial drugs) was reportedly effective in the treatment of MAP-associated autoimmune diseases such as CD [33] and MS (https://clinicaltrials.gov/ct2/show/NCT01717664). This suggests that eliminating the etiopathological agent itself might be a better strategy for treating MAP-associated autoimmune diseases.

In the present study, we have performed a systematic *in silico* analysis of metabolic chokepoints of MAP mimicry proteins and their interacting protein partners to identify suitable drug targets that can eliminate the pathogen itself and, thereby treat various MAP-associated autoimmune diseases. Metabolic pathway/metabolic network analysis and identification of metabolic chokepoints are widely used *in silico* methods to identify drug targets in pathogen genomes. By definition, a chokepoint enzyme either consumes a unique substrate or produces a unique product in the pathogen metabolic network [34]. Inhibition of chokepoint enzymes may disrupt crucial metabolic processes in the pathogen, so chokepoints that are essential to the pathogen represent good potential drug targets [35–39]. In our study, information regarding the interacting proteins of MAP involved in molecular mimicry with the host proteins was retrieved from the STRING database. All the interacting partners of MAP mimicry proteins that showed homology with the human proteins were removed, and only nonhomologous MAP proteins were included in further analysis. Then we determined the chokepoint(s) of the metabolic pathways, followed by further confirmation of the essentiality of the identified chokepoints in the survival of MAP. Finally, the potential drugs that can block the chokepoint(s) were identified using the DrugBank database and confirmed by molecular docking (Figure 1). Additionally, the 'off-target' binding potential of each proposed drug was also assessed.

Materials and methods

Retrieval of experimentally validated MAP mimicry proteins involved in autoimmune diseases

The information about MAP proteins involved in molecular mimicry and autoimmune diseases was retrieved from the database—miPepBase [40]. The miPepBase is a database of experimentally validated mimicry proteins, which was earlier developed in our laboratory. It contains extensive manually curated information about all the mimicry proteins/peptides and the autoimmune diseases reported, till date. Here, we used '*Mycobacterium avium* subsp. *paratuberculosis*' as the keyword to access the data.

Protein–protein interaction studies

The STRING database was used to retrieve the information about the interacting partners of MAP mimicry proteins (IPMMP). STRING contains information about protein–protein interactions (PPI) that were established by experimental studies and, also by different methods of genome analysis such as domain fusion, phylogenetic profiling and gene neighborhood [41]. Each PPI of STRING is assigned with a confidence score (0.15 = low confidence, ≥0.4 = medium confidence and default threshold, ≥0.7 = high confidence and ≥ 0.9 = highest confidence). The STRING confidence score is a combined score of eight methods namely neighborhood on the chromosome, gene fusion, phylogenetic co-occurrence, homology, coexpression, experimentally determined expression, database annotated and automated text mining. If the combined STRING score is >0.5 then the chances of false-positive interactions in the second shell are quite less [42]. Hence, in the present work, we have used a high confidence STRING combined score threshold of ≥0.7 to reduce the false-positive and negative PPIs.

Removal of human homologs of MAP proteins

To avoid nonspecific binding of a ligand to host proteins, MAP proteins that were homologous to human proteins were removed from the list of interacting proteins. In this study, if an IPMMP showed ≥30% identity and ≥80% coverage with a human protein, it was considered as homologous to the human protein and referred as homologous interacting partners of MAP mimicry proteins (HIPMMP), whereas remaining proteins were considered as nonhomologous interacting partners of MAP mimicry proteins (nHIPMMP).

Determining the chokepoint(s) of the metabolic pathways of MAP

Each nHIPMMP was mapped in their corresponding metabolic networks in the KEGG database [43]. KEGG is a database resource that cross-integrates genomic, chemical and systemic functional information of an organism. It is widely used as a reference knowledge base for the integration and interpretation of large-scale datasets generated by genome sequencing and other high-throughput experimental technologies. In the present work, each nHIPMMP was manually surveyed to assess their capability to be a potential chokepoint protein.

Validation of the essentiality of chokepoint proteins

The validation of the essentiality of chokepoint proteins in MAP metabolic pathways was performed using two parameters. First, we assessed the presence of homologs of the chokepoint proteins in all the mycobacterial reference proteomes present in UniProtKB (total 44 mycobacterial proteomes, as in December 2019) [44]. If a chokepoint protein exhibited ≥50% identity over 80% of the sequence length in at least 10 mycobacterial

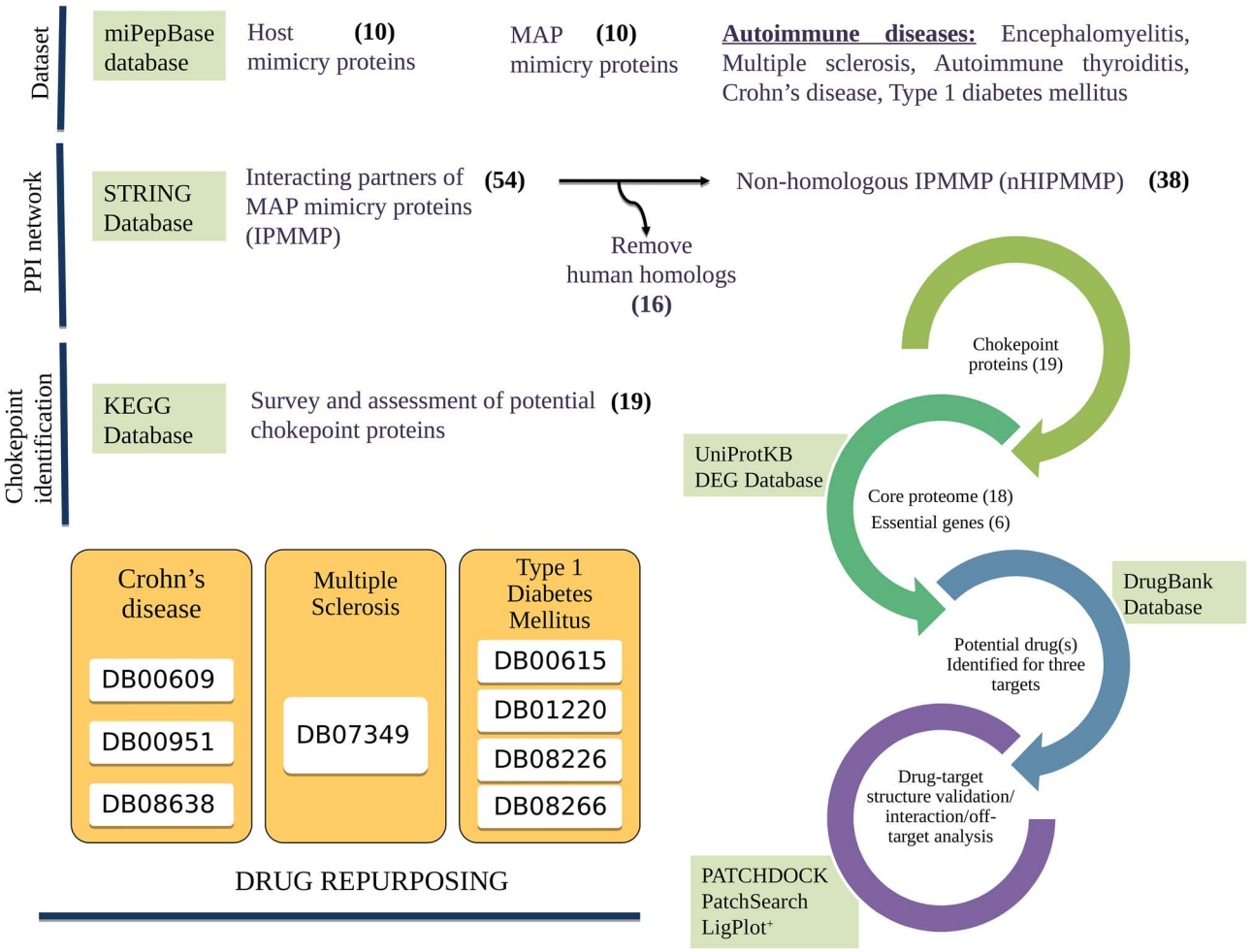

Figure 1. The workflow adopted for the identification of novel drug targets for treating *M. avium* subsp. *paratuberculosis*-associated autoimmune disorders.

proteomes, it was considered as the part of the core proteome. Second, a chokepoint protein was considered as an essential protein if it showed similarity with other mycobacterial essential genes enlisted in the database of essential genes (DEG) [45]. The chokepoint protein(s) was considered an essential gene of MAP, if the alignment identity and sequence coverage with a protein in DEG was more than equal to 50% and 80%, respectively.

Identification and validation of drug molecules against the chokepoint proteins

The potential drugs that can block the chokepoint(s) were searched in the DrugBank, the most popular and well-curated and regularly updated database of drug molecules. The DrugBank version 5.1.5 was used in this study, which contains more than 13 000 drug entries, including approved small-molecule drugs, biologics (proteins, peptides, vaccines and allergenics), nutraceuticals and experimental (discovery phase) drugs [46]. BLAST^{+} search was used to find the homologs of chokepoint under four target categories namely, drug targets, drug enzymes, drug carriers and drug transporters. A drug target protein of DrugBank was considered as a homolog of chokepoint if it showed $\geq$50% identity over 80% of the sequence length. The drug molecules that showed the best hit of DrugBank target proteins were selected as potential binders of chokepoint proteins.

Further, drug molecules were optimized according to Lipinski's rule of five scales. The properties required by this rule are: molecular weight $\leq$ 500, number of rotatable bonds $\leq$ 10, hydrogen bond donors $\leq$ 5, hydrogen bond acceptors $\leq$ 10 and logP $\leq$ 5. In addition to Lipinski's rule of five, half-life $\geq$ 60 min and toxicity information of potential drug(s) were also considered. Only those drugs that qualified at least five of the seven parameters were included, whereas dietary supplements, micronutrients and vitamins were excluded from the list of potential drug molecules.

Analysis of drug–target interactions

To assess the binding potential of selected drug candidates with their target, PatchDock was used. PatchDock is a molecular docking algorithm that gives geometry shape complementarity score, area, atomic contact energy and 3D transformation outputs [47]. The 3D structures of the MAP chokepoint protein were not found in Protein Data Bank (PDB) hence they were modeled using Swiss-model. The quality assessment of the homology models was carried out using several structure assessment parameters namely, Qualitative Model Energy Analysis (QMEAN) score [48], MolProbity score [49] and Ramachandran plot [50]. Further, we

also calculated the root-mean-square deviation (RMSD) between the modeled and the template structure. The structures of the potential drugs were downloaded from DrugBank. The protein–ligand complex with the highest docked score was selected and Ligplot^{+} was used for the analysis of binding interactions [51].

SiteEngine was used to recognize the functional binding sites of the protein models and compare it with a similar ligand-binding pocket. SiteEngine maps the 4 Å region around the ligand and uses it to search for similar structural patterns on the surface of other proteins [52]. In the present work, we used SiteEngine to compare the constituent amino acids of ligand-binding sites of the modeled and experimentally determined structure of proteins. The information of proteins, whose structure was solved and are known to bind the same ligand, was obtained from the DrugBank.

Results and discussion

Identification of interacting protein partners of MAP mimicry proteins

The keyword search '*Mycobacterium avium* subsp. *paratuberculosis*' in miPepBase displayed 14 entries/events related to mimicry. In these 14 events, 10 distinct proteins of MAP were found that exhibited molecular mimicry with the host proteins. These proteins cross-reacted with 10 different types of host proteins resulting in 5 different types of autoimmune diseases namely, encephalomyelitis, multiple sclerosis, autoimmune thyroiditis, CD and T1DM (Table S1). We found interacting protein partners for only eight of these 10 MAP proteins in the STRING database (Table S2). The STRING database failed to provide information about interacting proteins of two MAP mimicry proteins (UniProt accession number: Q53467 and Q73SP6), hence these two proteins were removed from further analysis.

STRING analysis revealed that eight MAP mimicry proteins had 54 interacting protein partners (IPMMP) at a high confidence score. The individual interaction score and the final confidence score of each PPI are in Table S3. Of these, 16 IPMMPs were found to be homologous to human proteins, hence they were excluded from further analysis (Table 1).

Identification and validation of the metabolic chokepoints of MAP

The 38 nonhuman homologs of interacting partners of MAP mimicry proteins were mapped on 17 metabolic pathways of MAP (Table 1). The pathways were analyzed manually to find possible chokepoint reaction(s). We found that these 38 proteins were a part of the 19 chokepoints of the MAP metabolic pathways (Table 1 and Figure S1).

Among the 19 chokepoint proteins, it was observed that 18 proteins were part of the core mycobacterial proteome (Table S4), and 6 of the 19 chokepoint proteins shared a close homology with *M. tuberculosis* essential genes as per the DEG database. Thus, total 18 chokepoint proteins identified by us in this study qualified one or the other parameter of essentiality and were included in further analysis (Table 2).

Drug molecules for metabolic chokepoints

DrugBank search revealed 13 potential drug molecules against 3 of the 18 chokepoint proteins *viz.* katG, rpoB and narH (Table 2). After benchmarking the drug molecules on the seven drug-like parameters, we were finally left with eight probable drug candidates (Table S5). Of these eight probable drug candidates, we noted that four molecules, *viz.* DB00609, DB00951, DB00615 and DB01220 were Food and Drug Administration (FDA)-approved drugs and the remaining four molecules, viz. DB08638, DB08226, DB08266 and DB07349 were experimentally validated drugs (Table 3). The details of metabolic chokepoints of MAP and their corresponding DrugBank molecules are as follows:

(i) katG: It was discerned as the chokepoint protein of the metabolic processes associated with MAP mimicry protein, alkylohydroperoxidase C or ahpC (UniProt Id: Q73ZL3). The ahpC protein of MAP mimics the human cytoskeleton-associated protein 5 (colonic and hepatic tumor overexpressed gene protein) resulting in CD [53]. katG is a bifunctional enzyme that shows both catalase and peroxidase activity [54]. Several studies have reported that katG protects *M. tuberculosis* from toxic reactive oxygen species, shows peroxynitritase activity and helps in survival within the host macrophages [55–57]. We observed MAP-associated katG can be targeted by three DrugBank molecules *viz.* DB00609 (Ethionamide), DB00951 (Isoniazid) and DB08638. Of these, DB00951 (Isoniazid) has proved useful in the treatment of *M. tuberculosis* and *M. avium* intracellulare infections [58]. Ethionamide (DB00609) is a FDA-approved drug, which is used in combination with other antituberculosis drugs in second-line treatment of multidrug-resistant active tuberculosis. It has been shown to be effective against *M. bovis* and *M. smegmatis* also [59]. DB08638 (1-hydroperoxy-L-tryptophan) is an experimentally validated drug molecule that was reported to bind with the katG protein of *Burkholderia pseudomallei* (https://www.drugbank.ca/drugs/DB08638).

(ii) rpoB: It was identified as the chokepoint protein of metabolic processes associated with MAP mimicry protein hsp65 (UniProt Id: P42384) and its interacting partners. The hsp65 of MAP shows mimicry with human glutamate decarboxylase 2, which results in T1DM [60]. The rpoB gene encodes the β-subunit of bacterial RNA polymerase. It was found to be the drug target of four DrugBank molecules *viz.* DB00615 (Rifabutin), DB01220 (Rifaximin), DB08226 (Myxopyronin B) and DB08266. Rifabutin (DB00615) and Rifaximin (DB01220) are FDA-approved rifamycin antibiotics. Rifabutin is used for treating *M. avium* complex infections, whereas Rifaximin is used to treat traveler's diarrhea, irritable bowel syndrome and hepatic encephalopathy [61]. Myxopyronin B (DB08226) and DB08266 are experimental drugs, which were shown to target rpoB of *Thermus thermophilus* (https://www.drugbank.ca/drugs/DB08266).

(iii) narH: The MAP mimicry protein narH (UniProt Id: Q73WP1) mimics the human myelin oligodendrocyte glycoprotein, which results in MS [62]. Here, we found DrugBank molecule DB07349 as a potential inhibitor of narH. The narH gene encodes the beta chain of the enzyme nitrate reductase, which helps the bacteria during anaerobic growth. The nitrate reductase enzyme complex has been studied extensively in *Escherichia coli* where it helps in using nitrate as an electron acceptor during anaerobic growth [63]. DrugBank search revealed that the drug molecule DB07349 was already known to effectively target the narH protein of *E. coli* (https://www.drugbank.ca/bio_entities/BE0003816). DB07349 belongs to the class phosphatidylglycerols and is an experimental drug.

Table 1. The list of MAP mimicry proteins, interacting partners of MAP mimicry proteins (IPMMP), name of the human homolog of IPMMP (if present), the nonhuman homolog of IPMMP (nHIPMMP), KEGG pathway ID to which non-human homolog of IPMMP belongs (KEGG ID is in parenthesis) and chokepoint proteins

S. No.	MAP mimicry proteins	IPMMP	HIPMMP	nHIPMMP	Pathway in which nHIPMMPs are involved	Chokepoint protein(s)
1	P42384	clpP2, dnaK, groEL1, groEL2, groES, grpE, guaA, gyrB, htpG, MAP_2278c, rpoB	clpP2, dnaK, groEL1, groEL2, groES, guaA, htpG, MAP_2278c	grpE, gyrB, rpoB	RNA polymerase (mpa03020)	rpoB
2	Q73T54	MAP_3865c, MAP_3866c, MAP_3867c	MAP_3865c	MAP_3866c, MAP_3867c	NA	NA
3	Q73WG6	fum, MAP_2692, MAP_2694	fum	MAP_2692, MAP_2694	Glycolysis / Gluconeogenesis (mpa00010) Methane metabolism (mpa00680) Pentose phosphate pathway (mpa00030) Fructose and mannose metabolism (mpa00051)	NA
4	Q73WP1	MAP_0807c, MAP_2102c, MAP_2722c, MAP_3707c, narG, narH, narI, narJ, narU, nirB, nirD	MAP_0807c	MAP_2102c, MAP_2722c, MAP_3707c, narG, narH, narI, narJ, narU, nirB, nirD	Nitrogen metabolism (mpa00910) Two-component system (mpa02020)	MAP_2102c, MAP_3707c, narG, narH, narI, narU, nirB, nirD, nirJ
5	Q73ZL3	ahpC, ahpD, ahpE, catB, katG, oxyR, sodA, sodC, sucB, tpx, trxB2	ahpC, ahpE, sodA, sucB	ahpD, catB, katG, oxyR, sodC, tpx, trxB2	Selenocompound metabolism (mpa00450) Phenylalanine metabolism (mpa00360) Drug metabolism - other enzymes (mpa00983) Glyoxylate and dicarboxylate metabolism (mpa00630) Tryptophan metabolism (mpa00380)	katG, catB
6	Q740V8	MAP_1234, MAP_1235, MAP_3252	NA	MAP_1234, MAP_1235, MAP_3252	Lipopolysaccharide biosynthesis (mpa00540)	MAP_3252
7	Q741P6	apt, MAP_1042, MAP_1045, relA, secD, secE, secF, secY, yidC	apt	MAP_1042, MAP_1045, relA, secD, secE, secF, secY, yidC	Quorum sensing (mpa02024) Protein export (mpa03060) Bacterial secretion system (mpa03070) Purine metabolism (mpa00230)	MAP_1042, MAP_1045, secD, secE, secF, secY
8	Q745A5	MAP_0106c, MAP_2148, MAP_2752	NA	MAP_0106c, MAP_2148, MAP_2752	NA	NA

Table 2. The list of chokepoint proteins and drugs identified against them

MAP mimicry protein	Chokepoint proteins	Chokepoint proteins that were part of		Drug target and candidate as per the DrugBank database	Drugs that qualified at least 5 of 7 drug-like properties
		Essential gene database	Core proteome		
P42384	rpoB	rpoB	rpoB	rpoB: DB04788; DB08226; DB08266; DB00615; DB01045; DB01220; DB04934; DB11753	DB00615 (Rifabutin), DB01220 (Rifaximin), DB08226 (Myxopyronin B), DB08266 (Methyl [(1E,5R)-5-{3-[(2E,4E)-2,5-dimethyl-2,4-octadienoyl]-2,4-dioxo-3,4-dihydro-2H-pyran-6-yl}hexylidene]carbamate)
Q73WP1	MAP_2102c, MAP_3707c, narG, narH, narI, narU, nirB, nirD, nirJ	NA	MAP_2102c, MAP_3707c, narG, narH, narI, narU, nirB, nirD, nirJ	narH: DB04464; DB07349	DB07349 (1S)-2-{[{[(2S)-2,3-dihydroxypropyl]oxy}(hydroxy)phosphoryl]oxy}-1-[(pentanoyloxy)methyl)ethyl octanoate
Q73ZL3	katG, catB	katG	katG, catB	katG: DB00609; DB00951; DB08638	DB00609 (Ethionamide), DB00951 (Isoniazid), DB08638 (1-hydroperoxy-L-tryptophan)
Q740V8	MAP_3252	NA	NA	NA	NA
Q741P6	MAP_1042, MAP_1045, secD, secE, secF, secY	secD, secE, secF, secY	MAP_1042, MAP_1045, secD, secE, secF, secY	NA	NA

Table 3. Details of drugs proposed against MAP associated autoimmune disorder

Autoimmune Disease	MAP mimicry protein	Metabolic Chokepoint	Drug Name	DrugBank ID	Drug Group
Crohn's Diseases	Q73ZL3 (ahpC)	katG	Ethionamide	DB00609	FDA approved
			Isoniazid	DB00951	FDA approved
			1-hydroperoxy-L-tryptophan	DB08638	Experimentally validated
Multiple Sclerosis	Q73WP1 (narH)	narH	(1S)-2-{[{[(2S)-2,3-dihydroxypropyl]oxy}(hydroxy)phosphoryl]oxy}-1-[(pentanoyloxy)methyl)ethyl octanoatehyl)ethyl octanoate	DB07349	Experimentally validated
Type 1 diabetes mellitus	P42384 (hsp65)	rpoB	Rifabutin	DB00615	FDA approved
			Rifaximin	DB01220	FDA approved
			Myxopyronin B	DB08226	Experimentally validated
			Methyl [(1E,5R)-5-{3-[(2E,4E)-2,5-dimethyl-2,4-octadienoyl]-2,4-dioxo-3,4-dihydro-2H-pyran-6-yl}hexylidene] carbamate	DB08266	Experimentally validated

Table 4. Quality assessment scores of *in silico* protein models

Target protein	Amino acids in the allowed regions of Ramachandran plot (%)	QMEAN score	MolProbity score	RMSD value
katG	97.25	-0.06	1.09	0.118
narH	94.61	-2.64	1.14	0.187
rpoB	97.88	0.65	0.78	0.051

Quality assessment of the modeled metabolic chokepoint protein structures

To evaluate the quality of the modeled proteins, we evaluated several structural features of the modeled structures. Analysis of the Ramachandran plot showed more than 94% residues of the modeled protein structures were present in the favored regions of the Ramachandran plot (Table 4), which is more than the required 90% threshold for a good quality protein structure [64]. The QMEAN server (https://swissmodel.expasy.org/qmean) provides a comprehensive quality score (Z-score) that measures the 'degree of nativeness' of the structural features of a modeled protein structure. The score also provides the likelihood that a given model is of comparable quality to experimental structures [65]. The QMEAN-score of the modeled structure of katG, rpoB and narH were −0.06, −0.65 and −2.64, respectively, comparable to that of high-resolution experimental structures (Table 4). The MolProbity score provides a single quantifiable measure to assess the quality of a biomolecular structure. The score is calculated by a comprehensive all-atom contact analysis to find the steric problems within the query biomolecule [66]. The low MolProbity scores also indicated the good quality of the modeled structure (Table 4). Together, the three assessment methods confirmed that the modeled structures were of good quality. Further, when the protein models were superimposed with the corresponding templates and the RMSD of two structures were calculated, we found a very low RMSD value (0.118 for katG, 0.187 for narH and 0.051 for rpoB). The structure validation scores and superimposed structures of target–template are shown in Figure S2.

SiteEngine webserver that was used to validate the drug binding site revealed a high similarity score and low RMSD values between the drug binding site of modeled protein and the binding site of the same drug on already reported target PDB structures (Table S6). Also, the amino acids present at the drug target interfaces of the modeled and experimentally determined protein structures were similar. These observations suggest that the proposed drug molecules bind to the appropriate binding sites in the protein models of katG, narH and rpoB.

Protein–ligand interactions between the MAP proteins and DrugBank molecules

Binding analysis of katG with the identified DrugBank molecules (Table 5) revealed a better binding affinity for DB00951 (Isoniazid) (−139.56 kcal/mol) than for DB08638 (−114.46 kcal/mol) or DB00609 (Ethionamide) (−101.17 kcal/mol). Also, the number of residues bound by hydrogen and hydrophobic bonding in the katG-DB00951/DB08638 protein–ligand complex was more than the katG-DB00609 complex. In the case of rpoB, DB00615 (Rifabutin) showed a significantly higher binding energy (−489.49 kcal/mol) and, more hydrogen and hydrophobic binding residues in the protein–ligand complex than the other three DrugBank molecules, DB01220 (rifaximin), DB08226 (myxopyroninB) and DB08266. In the case of narH, the only identified drug molecule DB07349 showed a good binding affinity of −11.62 kcal/mol and strong polar and hydrophobic interactions between the protein and ligand. The interactions between each drug and target are shown in Figure S3.

Prediction of off-target binding

The targets of the proposed drug molecules are MAP proteins. Although MAP proteins that have a human homolog were omitted from this study, yet the presence of structurally conserved binding sites at the surfaces of human proteins might lead to the 'off-target' binding of the proposed drugs. Thus to remove this possibility, we used the PatchSearch webserver that searches for the potential 'off-target' binding sites in a set of user-supplied protein structures. Further, it also estimates the binding affinity [67].

The PatchSearch 'off-target' search results showed that the DrugBank molecules DB00609, DB00951 and DB08638 targeted against katG had a very poor binding affinity for human proteins (Table S7). The binding affinity of katG and the three DrugBank molecules ranged from −1.854 to −0.002, −0.256 to 0 and − 0.381 to 0 kcal/mol, respectively. Similarly, the proposed DrugBank molecule DB07349 (target narH) also showed a negligible affinity for human proteins (range, −3.141 to −0.006 kcal/mol). The four drug molecules namely, DB00615, DB01220, DB08226 and DB08266 (target rpoB), also showed a very low binding affinity for human proteins (−0.446 to 0, −1.935 to 0, −10.25 to 0 and 0 kcal/mol, respectively. Interestingly, two human proteins (PDB ID: 6D0M and 2YGN) showed a higher affinity for Myxopyronin B (DB08226). But, because of a nonsignificant higher RMSD value at the binding site (>1.5), it has a very low chance of 'off-target' binding.

To summarize, in this study we have performed a systematic *in silico* analysis of MAP mimicry proteins and their interacting partners to identify chokepoints of their respective metabolic processes. Drug molecules, which qualified several stringent parameters, were shortlisted as appropriate inhibitors of the chokepoints. Molecular interactions between the drug target(s) and drug molecule(s) were further confirmed by molecular docking and off-target binding potential of the proposed drugs. Finally, we were left with eight DrugBank molecules that might prove useful for treating three MAP-associated autoimmune diseases, namely T1DM, CD and MS. Interestingly, all the drug molecules identified in our study, except Rifabutin are novel drug molecules that have not been tested for treating MAP-associated T1DM, CD and MS. Moreover, the drug molecules identified during our analysis are either FDA-approved drugs or experimental drugs with proven efficacy. Hence, these can be easily incorporated in clinical studies or tested *in vitro* for assessing their suitability in treating MAP-associated autoimmune diseases. Thus, instead of proposing new chemotherapeutics, our study helps in repurposing the known drugs for treating MAP induced autoimmune T1DM, CD and MS. We would like to

Table 5. Interaction pattern of the proposed drug and target proteins

Selected target	Drug molecule	PatchDock score (kcal/mol)	Nature of interaction	Amino acid on binding sites
katG	DB00609 (Ethionamide)	−101.17	Hydrophobic interaction	Leu387, Pro241, Ser324, Arg110, Lys283, Trp330, His279
			Polar H interaction	Nill
	DB00951 (Isoniazid)	−139.56	Hydrophobic interaction	Ile109, Gly278, Trp113, His279, Gly282, Gly106, Phe281, Try330
			Polar H interaction	Lys283
	DB08638 (1-hydroperoxy-L-tryptophan)	−114.46	Hydrophobic interaction	Gly106, Gly278, His279, Trp113, Asp143, Arg110, Lys283, Phe281
			Polar H interaction	Ser324
narH	(1S)-2-{[{[(2S)-2,3-dihydroxypropyl]oxy}(hydroxy)phosphoryl]oxy}-1-[(pentanoyloxy)methyl]ethyl octanoate	−11.62	Hydrophobic interaction	Asn187, Ala192, Trp66, Leu74, Asp69, Gly80
			Polar H interaction	Arg81, Arg73, Arg75
rpoB	DB00615 (Rifabutin)	−489.49	Hydrophobic interaction	Lys 298, Asn292, Glu391, Leu87, Arg276, Arg299, Pro89, Glu91, Glu86, Thr288, Asp92, Arg395, Try272, Glu279, Ser94, Thr282, Glu284, Ser285, Thr288
			Polar H interaction	Pro280
	DB01220 (Rifaximin)	−45.14	Hydrophobic interaction	Arg276, Arg395, Leu293, Gly203, Arg384, Ala204, Asn381, Gln382, Glu423, Ser388, Arg392, Leu289
			Polar H interaction	Arg299, Arg389, Val385,
	DB08226 (Myxopyronin B)	−25.07	Hydrophobic interaction	Ile90, Leu289, Arg395, Glu391, Glu91, Pro89, Asp221, Gly203, Ala204, Trp205, Arg373, Arg175, Val174, Arg384, Arg222
			Polar H interaction	Arg299, Arg276
	DB08266 (Methyl [(1E,5R)-5-{3-[(2E,4E)-2,5-dimethyl-2,4-octadienoyl]-2,4-dioxo-3,4-dihydro-2H-pyran-6-yl}hexylidene] carbamate)	19.09	Hydrophobic interaction	Arg373, Glu377, Arg384, Ser201, Asn381, Val385, Arg395, Glu391, Leu289, Glu91, Thr288
			Polar H interaction	Ser388, Arg299, Arg276

add if other relevant databases like ChEMBL [68], PubChem [69], ChemBank [70] and ZINC database [71] were also included in the study, we might have discerned more drug molecules against these or other MAP-associated autoimmune diseases.

Key points

- A novel method to treat autoimmune diseases associated with *Mycobacterium avium* subspecies *paratuberculosis* (MAP).
- Our work found eight known drug molecules that may be evaluated to treat MAP-associated autoimmunity.
- The proposed schema may be used to repurpose drugs to treat autoimmune diseases induced by other pathogens.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Author contributions

NS and MK conceived and designed the study. AG performed the work, prepared the illustrations and wrote initial draft manuscript. All authors contributed to the critical analysis and revision. All authors analyzed the results and wrote the final versions of manuscript.

Acknowledgments

Authors thank Prof. B. K. Thelma (Department of Genetics, University of Delhi South Campus, New Delhi) for critical inputs to address the reviewer's comments. The insightful comments from the anonymous reviewers are also greatly appreciated. We thank the University of Delhi to provide excellent infrastructure to carry out the work.

Funding

The work was carried out using the resources funded by the Science and Engineering Research Board under Fast Track Proposals for Young Scientists Scheme [Grant No: SR/FT/LS-84/2010] and UGC Major Research Project [Grant No: 41-38/2012(SR)]. AG is supported by ICMR-JRF scheme [Grant Number: 3/1/3 J.R.F.-2016/LS/HRD-(32262)]. NS is supported by CSIR Senior Research Associateship (Scientists' Pool Scheme) [Grant Number: 13(9089-A)/2019-Pool].

References

1. Shin SJ, Lee BS, Koh W-J, *et al*. Efficient differentiation of *Mycobacterium avium* complex species and subspecies by use of five-target multiplex PCR. *J Clin Microbiol* 2010;**48**:4057–62.
2. Sechi LA, Dow CT. *Mycobacterium avium* ss. *Paratuberculosis zoonosis* - the hundred year war - beyond Crohn's disease. Front. *Immunology* 2015;**6**:96.
3. Dow CT, Ellingson JLE. Detection of *Mycobacterium avium* ss. paratuberculosis in Blau syndrome tissues. *Autoimmune Dis* 2010;**2011**:127692.
4. Wynne JW, Bull TJ, Seemann T, *et al*. Exploring the zoonotic potential of *Mycobacterium avium* subspecies paratuberculosis through comparative genomics. *PLoS One* 2011;**6**:e22171.
5. Whiley H, Keegan A, Giglio S, *et al*. *Mycobacterium avium* complex–the role of potable water in disease transmission. *J Appl Microbiol* 2012;**113**:223–32.
6. Robertson RE, Cerf O, Condron RJ, *et al*. Review of the controversy over whether or not *Mycobacterium avium* subsp. paratuberculosis poses a food safety risk with pasteurised dairy products. *Int Dairy J* 2017;**73**:10–8.
7. Bitti MLM, Masala S, Capasso F, *et al*. *Mycobacterium avium* subsp. paratuberculosis in an Italian cohort of type 1 diabetes pediatric patients. *Clin Dev Immunol* 2012;**2012**:785262.
8. Miller FW. Environmental agents and autoimmune diseases. *Adv Exp Med Biol* 2011;**711**:61–81.
9. Dow CT. M. Paratuberculosis heat shock protein 65 and human diseases: bridging infection and autoimmunity. *Autoimmune Dis* 2012;**2012**:150824.
10. Sechi LA, Gazouli M, Ikonomopoulos J, *et al*. *Mycobacterium avium* subsp. paratuberculosis, genetic susceptibility to Crohn's disease, and Sardinians: the way ahead. *J Clin Microbiol* 2005;**43**:5275–7.
11. Wang H, Yuan F-F, Dai Z-W, *et al*. Association between rheumatoid arthritis and genetic variants of natural resistance-associated macrophage protein 1 gene: a meta-analysis. *Int J Rheum Dis* 2018;**21**:1651–8.
12. Sechi L-A, Gazouli M, Sieswerda L-E, *et al*. Relationship between Crohn's disease, infection with *Mycobacterium avium* subspecies paratuberculosis and SLC11A1 gene polymorphisms in Sardinian patients. *World J Gastroenterol* 2006;**12**:7161–4.
13. Dow CT. M. Paratuberculosis and Parkinson's disease – is this a trigger. *Med Hypotheses* 2014;**83**:709–12.
14. Härtlova A, Herbst S, Peltier J, *et al*. LRRK2 is a negative regulator of *Mycobacterium tuberculosis* phagosome maturation in macrophages. *EMBO J* 2018;**37**.
15. Sharp RC, Beg SA, Naser SA. Polymorphisms in protein tyrosine phosphatase non-receptor type 2 and 22 (PTPN2/22) are linked to hyper-proliferative T-cells and susceptibility to mycobacteria in rheumatoid arthritis. *Front Cell Infect Microbiol* 2018;**8**.
16. Dow CT. Paratuberculosis and type I diabetes: is this the trigger? Med. *Hypotheses* 2006;**67**:782–5.
17. Cossu D, Masala S, Cocco E, *et al*. Association of *Mycobacterium avium* subsp. paratuberculosis and SLC11A1 polymorphisms in Sardinian multiple sclerosis patients. *J Infect Dev Ctries* 2013;**7**:203–7.
18. D'Amore M, Lisi S, Sisto M, *et al*. Molecular identification of *Mycobacterium avium* subspecies paratuberculosis in an Italian patient with Hashimoto's thyroiditis and Melkersson–Rosenthal syndrome. *J Med Microbiol* 2010;**59**:137–9.
19. Gong L, Liu B, Wang J, *et al*. Novel missense mutation in PTPN22 in a Chinese pedigree with Hashimoto's thyroiditis. *BMC Endocr Disord* 2018;**18**.
20. Sisto M, Cucci L, D'Amore M, *et al*. Proposing a relationship between *Mycobacterium avium* subspecies paratuberculosis infection and Hashimoto's thyroiditis. *Scand J Infect Dis* 2010;**42**:787–90.
21. Arru G, Caggiu E, Paulus K, *et al*. Is there a role for *Mycobacterium avium* subspecies paratuberculosis in Parkinson's disease? *J Neuroimmunol* 2016;**293**:86–90.
22. Bo M, Erre GL, Niegowska M, *et al*. Interferon regulatory factor 5 is a potential target of autoimmune response triggered by Epstein-barr virus and *Mycobacterium avium*

subsp. paratuberculosis in rheumatoid arthritis: investigating a mechanism of molecular mimicry. *Clin Exp Rheumatol* 2018;**36**:376–81.
23. Paccagnini D, Sieswerda L, Rosu V, *et al*. Linking chronic infection and autoimmune diseases: *Mycobacterium avium* subspecies paratuberculosis, SLC11A1 polymorphisms and type-1 diabetes mellitus. *PLoS One* 2009;**4**:e7109.
24. Sechi LA, Scanu AM, Molicotti P, *et al*. Detection and isolation of *Mycobacterium avium* subspecies paratuberculosis from intestinal mucosal biopsies of patients with and without Crohn's disease in Sardinia. *Am J Gastroenterol* 2005;**100**:1529–36.
25. Naser SA, Ghobrial G, Romero C, *et al*. Culture of *Mycobacterium avium* subspecies paratuberculosis from the blood of patients with Crohn's disease. *Lancet* 2004;**364**:1039–44.
26. Mameli G, Cocco E, Frau J, *et al*. Epstein Barr virus and *Mycobacterium avium* subsp. paratuberculosis peptides are recognized in sera and cerebrospinal fluid of MS patients. *Sci Rep* 2016;**6**:22401.
27. Quayle AJ, Wilson KB, Li SG, *et al*. Peptide recognition, T cell receptor usage and HLA restriction elements of human heat-shock protein (hsp) 60 and mycobacterial 65-kDa hsp-reactive T cell clones from rheumatoid synovial fluid. *Eur J Immunol* 1992;**22**:1315–22.
28. Chandrashekara S. The treatment strategies of autoimmune disease may need a different approach from conventional protocol: a review. *Indian J Pharm* 2012;**44**:665.
29. Mameli G, Cossu D, Cocco E, et al. Epstein–Barr virus and *Mycobacterium avium* subsp. paratuberculosis peptides are cross recognized by anti-myelin basic protein antibodies in multiple sclerosis patients. *J Neuroimmunol* 2014; **270**:51–5
30. Van der Kooij SM, de Vries-Bouwstra JK, Goekoop-Ruiterman YPM, *et al*. Limited efficacy of conventional DMARDs after initial methotrexate failure in patients with recent onset rheumatoid arthritis treated according to the disease activity score. *Ann Rheum Dis* 2007;**66**:1356–62.
31. Greenstein RJ, Su L, Brown ST. The thioamides methimazole and thiourea inhibit growth of *M. avium* subspecies paratuberculosis in culture. *PLoS One* 2010;**5**:e11099.
32. Greenstein RJ, Su L, Juste RA, *et al*. On the action of cyclosporine a, Rapamycin and Tacrolimus on M. avium including subspecies paratuberculosis. 2008;**3**:PLoS One, e2496.
33. Chamberlin W, Borody TJ, Campbell J. Primary treatment of Crohn's disease: combined antibiotics taking center stage. *Expert Rev Clin Immunol* 2011;**7**:751–60.
34. Yeh I, Hanekamp T, Tsoka S, *et al*. Computational analysis of plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res* 2004;**14**:917–24.
35. Siwo GH, Tan A, Button-Simons KA, *et al*. Predicting functional and regulatory divergence of a drug resistance transporter gene in the human malaria parasite. *BMC Genomics* 2015;**16**:115.
36. Chung BK-S, Dick T, Lee D-Y. In silico analyses for the discovery of tuberculosis drug targets. *J Antimicrob Chemother* 2013;**68**:2701–9.
37. Sharma A, Pan A. Identification of potential drug targets in Yersinia pestis using metabolic pathway analysis: MurE ligase as a case study. *Eur J Med Chem* 2012;**57**:185–95.
38. Duffield M, Cooper I, McAlister E, *et al*. Predicting conserved essential genes in bacteria: in silico identification of putative drug targets. *Mol Biosyst* 2010;**6**:2482–9.
39. Garg A, Kumari B, Singhal N, *et al*. Using molecular-mimicry-inducing pathways of pathogens as novel drug targets. *Drug Discov Today* 2019;**24**:1943–52.
40. Garg A, Kumari B, Kumar R, *et al*. miPepBase: a database of experimentally verified peptides involved in molecular mimicry. *Front Microbiol* 2017;**8**:2053.
41. Mering C. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;**31**:258–61.
42. Szklarczyk D, Morris JH, Cook H, *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017;**45**:D362–8.
43. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.
44. Apweiler R, Bairoch A, Wu CH, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**:D115–9.
45. Zhang R, Ou H-Y, Zhang C-T. DEG: a database of essential genes. *Nucleic Acids Res* 2004;**32**:D271–2.
46. Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
47. Schneidman-Duhovny D, Inbar Y, Nussinov R, *et al*. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 2005;**33**:W363–7.
48. Benkert P, Tosatto SCE, Schomburg D. QMEAN: a comprehensive scoring function for model quality assessment. *Proteins* 2008;**71**:261–77.
49. Chen VB, Arendall WB, III, Headd JJ, *et al*. MolProbity : all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 2010;**66**:12–21.
50. Ho BK, Brasseur R. The Ramachandran plots of glycine and pre-proline. *BMC Struct Biol* 2005;**5**:14.
51. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. 'Protein engineering. *Design and Selection* 1995;**8**:127–34.
52. Shulman-Peleg A, Nussinov R, Wolfson HJ. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res* 2005;**33**:W337–41.
53. Polymeros D, Bogdanos DP, Day R, *et al*. Does cross-reactivity between mycobacterium avium paratuberculosis and human intestinal antigens characterize Crohn's disease? *Gastroenterology* 2006;**131**:85–96.
54. Singh R, Wiseman B, Deemagarn T, *et al*. Comparative study of catalase-peroxidases (KatGs). *Arch Biochem Biophys* 2008;**471**:207–14.
55. Wengenack NL, Jensen MP, Rusnak F, *et al*. Mycobacterium tuberculosis KatG is a peroxynitritase. *Biochem Biophys Res Commun* 1999;**256**:485–7.
56. Sherman DR, Mdluli K, Hickey MJ, *et al*. Compensatory ahpC gene expression in isoniazid-resistant mycobacterium tuberculosis. *Science* 1996;**272**:1641–3.
57. Ng VH, Cox JS, Sousa AO, *et al*. Role of KatG catalase-peroxidase in mycobacterial pathogenesis: countering the phagocyte oxidative burst. *Mol Microbiol* 2004;**52**:1291–302.
58. Society RC of TBT, Research Committee of the British Thoracic Society. First randomised trial of treatments for pulmonary disease caused by M avium intracellulare, M malmoense, and M xenopi in HIV negative patients: rifampicin, ethambutol and isoniazid versus rifampicin and ethambutol. *Thorax* 2001;**56**:167–72.
59. Rastogi N, Labrousse V, Goh KS. In vitro activities of fourteen antimicrobial agents against drug susceptible and

resistant clinical isolates of mycobacterium tuberculosis and comparative intracellular activities against the virulent H37Rv strain in human macrophages. *Curr Microbiol* 1996;**33**: 167–75.
60. Naser SA, Thanigachalam S, Dow CT, *et al*. Exploring the role of *Mycobacterium avium* subspecies paratuberculosis in the pathogenesis of type 1 diabetes mellitus: a pilot study. *Gut Pathog* 2013;**5**:14.
61. Rothstein DM. Rifamycins, alone and in combination. *Cold Spring Harb Perspect Med* 2016;**6**:a027011.
62. Cossu D, Mameli G, Masala S, *et al*. Evaluation of the humoral response against mycobacterial peptides, homologous to MOG35–55, in multiple sclerosis patients. *J Neurol Sci* 2014;**347**:78–81.
63. Ceccaldi P, Rendon J, Léger C, *et al*. Reductive activation of E. coli respiratory nitrate reductase. *Biochim Biophys Acta* 1847;**2015**:1055–63.
64. Pereira GRC, Da Silva ANR, Do Nascimento SS, *et al*. In silico analysis and molecular dynamics simulation of human superoxide dismutase 3 (SOD3) genetic variants. *J Cell Biochem* 2019;**120**:3583–98.
65. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 2011;**27**:343–50.
66. Davis IW, Leaver-Fay A, Chen VB, *et al*. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 2007;**35**:W375–83.
67. Rey J, Rasolohery I, Tufféry P, *et al*. PatchSearch: a web server for off-target protein identification. *Nucleic Acids Res* 2019;**47**:W365–72.
68. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012;**40**:D1100–7.
69. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem substance and compound databases. *Nucleic Acids Res* 2016;**44**: D1202–13.
70. Seiler KP, George GA, Happ MP, *et al*. Chem Bank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res* 2008;**36**:D351–9.
71. Irwin JJ, Shoichet BK. ZINC–a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;**45**:177–82.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 17 December 2019
doi: 10.3389/fmicb.2019.02865

Down-Regulation of Flagellar, Fimbriae, and Pili Proteins in Carbapenem-Resistant *Klebsiella pneumoniae* (NDM-4) Clinical Isolates: A Novel Linkage to Drug Resistance

Divakar Sharma[1†], Anjali Garg[2], Manish Kumar[2], Faraz Rashid[3] and Asad U. Khan[1]*

[1] Interdisciplinary Biotechnology Unit, Aligarh Muslim University, Aligarh, India, [2] Department of Biophysics, University of Delhi, New Delhi, India, [3] SCIEX Pvt. Ltd., Gurgaon, India

OPEN ACCESS

Edited by:
Raffaele Zarrilli,
University of Naples Federico II, Italy

Reviewed by:
Gokhlesh Kumar,
University of Veterinary Medicine Vienna, Austria
Rita Berisio,
Italian National Research Council (CNR), Italy

***Correspondence:**
Asad U. Khan
asad.k@rediffmail.com

†Present address:
Divakar Sharma,
Central Research Facility, Mass Spectrometry Laboratory, Kusuma School of Biological Sciences, Indian Institute of Technology, New Delhi, India

Specialty section:
This article was submitted to Antimicrobials, Resistance and Chemotherapy, a section of the journal Frontiers in Microbiology

Received: 17 August 2019
Accepted: 27 November 2019
Published: 17 December 2019

Citation:
Sharma D, Garg A, Kumar M, Rashid F and Khan AU (2019) Down-Regulation of Flagellar, Fimbriae, and Pili Proteins in Carbapenem-Resistant Klebsiella pneumoniae (NDM-4) Clinical Isolates: A Novel Linkage to Drug Resistance. Front. Microbiol. 10:2865. doi: 10.3389/fmicb.2019.02865

The emergence and spread of carbapenem-resistant *Klebsiella pneumoniae* infections have worsened the current situation worldwide, in which totally drug-resistant strains (bad bugs) are becoming increasingly prominent. Bacterial biofilms enable bacteria to tolerate higher doses of antibiotics and other stresses, which may lead to the drug resistance. In the present study, we performed proteomics on the carbapenem-resistant NDM-4-producing *K. pneumonia*e clinical isolate under meropenem stress. Liquid chromatography coupled with mass spectrometry (LC–MS/MS) analysis revealed that 69 proteins were down-regulated (≤0.42-fold change) under meropenem exposure. Within the identified down-regulated proteome (69 proteins), we found a group of 13 proteins involved in flagellar, fimbriae, and pili formation and their related functions. Further, systems biology approaches were employed to reveal their networking pathways. We suggest that these down-regulated proteins and their interactive partners cumulatively contribute to the emergence of a biofilm-like state and the survival of bacteria under drug pressure, which could reveal novel mechanisms or pathways involved in drug resistance. These down-regulated proteins and their pathways might be used as targets for the development of novel therapeutics against antimicrobial-resistant (AMR) infections.

Keywords: *Klebsiella pneumoniae* (NDM-4), proteomics, bioinformatics, pathway enrichment, biofilm, carbapenem resistance

INTRODUCTION

Klebsiella pneumoniae is a gram-negative bacteria of the family Enterobacteriaceae. In clinical settings, the emergence and spread of drug-resistant *K. pneumoniae* are worsening the medical situation worldwide. Carbapenems have been considered the last line of defense in the treatment of drug-resistant infections (Paterson, 2000; Paterson and Bonomo, 2005). Interrupted use of

Abbreviations: CLSI, Clinical and Laboratory Standards Institute; ESBLs, extended spectrum beta-lactamases; LB, Luria–Bertani; MIC, minimum inhibitory concentration; STRING, Search Tool for the Retrieval of Interacting Genes/Proteins.

carbapenem during the course of treatment leads to the emergence of carbapenem-resistant infections. Carbapenemases are produced that cleave or hydrolyze the carbapenem drugs and contribute to carbapenem resistance. Carbapenemase over-production and porin deficiency are the two major causes of carbapenem resistance (Ambler et al., 1991; Martínez-Martínez et al., 1999; Jacoby et al., 2004; Loli et al., 2006). Several explanations have been put forward to explain the mechanisms of carbapenem resistance, but our information is as yet incomplete or fragmentary.

Biofilm formation is among the mechanisms known to be responsible for microbial drug resistance. During biofilm formation, bacteria first become sessile and then colonize and grow up from surfaces. The biofilm protects the bacteria from various stresses like altered pH, osmolarity, and nutrient scarcity (Costerton and Lewandowski, 1995; Fux et al., 2005; McCarty et al., 2012) and blocks the entry of drugs to the bacterial communities (Costerton et al., 1999; Stewart and William Costerton, 2001; Sharma et al., 2019c). In the first step of biofilm formation, bacteria lose their motility and become sessile. We assume that decreased expression of proteins related to motility could lead to biofilm formation and thus might contribute to the development of drug resistance. Comparative proteomics addressing the whole-cell proteins of drug-resistant microbes with or without drug pressures have been reported previously (Lata et al., 2015; Khan et al., 2017; Sharma et al., 2018a; Qayyum et al., 2019; Sharma et al., 2019a). However, little information is available regarding the bacterial proteome related to biofilm, and, to the best of our knowledge, no data has yet been reported related to the proteome of drug-resistant microbes, especially carbapenem-resistant *K. pneumonia*, in relation to motility-mediated drug resistance.

In this study, we used comparative proteomics and systems biology-based approaches to investigate the correlation of the decreased expression of motility-related proteins (flagellar, fimbriae, and pili) with biofilm formation, which may lead to the development of drug resistance. Proteomics and systems biology approaches are both among the potential strategies for exploring biological problems such as the mechanisms of drug resistance. In the present study, we used liquid chromatography coupled with mass spectrometry (LC–MS/MS) to determine the expression of the motility-related proteome of a carbapenem-resistant *K. pneumoniae* (NDM-4) clinical isolate under meropenem stress. The results of this study could lead to the exploration of novel therapeutics targets against carbapenem resistance.

MATERIALS AND METHODS

Strain Selection and Drug Susceptibility Testing

An NDM-4-encoding carbapenem-resistant *K. pneumoniae* clinical isolate (AK-97) was selected for this study. This was reported in our earlier study, which showed its presence in the NICU of a northern Indian Hospital (Ahmad et al., 2018). Drug susceptibility testing (DST) against the drug meropenem was carried out via the micro-dilution method according to CLSI guidelines (Wayne, 2014).

Culture Scaling, Drug Induction, and Protein Sample Preparation

A single colony of *K. pneumoniae* was inoculated in LB broth and kept at 37°C at 220 rpm, and a sub-MIC (32 μg/ml) of meropenem was used for induction in a 200 ml culture flask. Bacteria were grown up to the exponential phase (OD_{600} = 0.8), and cells were harvested by centrifugation at 8000 × *g* for 8 min at 4°C. The cells were washed with normal saline and re-suspended in a lysis buffer [50 mM Tris–HCl containing 10 mM $MgCl_2$, 0.1% sodium azide, 1 mM phenyl-methyl-sulfonyl-fluoride (PMSF), and 1 mM ethylene glycol tetra-acetic acid (EGTA); pH 7.4] at a concentration of 1 g wet weight per 5 ml. Cell lysis was performed by intermittent sonication with a sonicator with the power at 35% amplitude (Sonics & Materials Inc., Newtown, CT, United States) for 10 min at 4°C. Further, the homogenate was centrifuged at 12,000 × *g* for 20 min at 4°C, and the supernatant was precipitated overnight at −20°C by adding cold acetone in excess (1:4) (Lata et al., 2015; Sharma and Bisht, 2016; Sharma et al., 2019a). The precipitated protein was collected by centrifugation (12,000 × *g*, 20 min), allowed to air dry, and then suspended in an appropriate volume of protein-dissolving buffer. The protein concentration was estimated using the Bradford (1976) assay. All of the experiments were replicated biologically and technically.

Separation and Identification of the Proteome by nanoLC-TripleTOF 5600 MS

Equal concentrations of protein samples were trypsinized, and digested proteins were analyzed using a TripleTOF 5600 MS (AB Sciex, Foster City, CA, United States) equipped with an Eksigent MicroLC 200 system (Eksigent, Dublin, CA, United States) with an Eksigent C18 reverse-phase column (150 × 0.3 mm, 3 μm, 120 Å) (Sharma et al., 2019a). For protein identification, spectral libraries were generated using information-dependent acquisition (IDA) mode after injecting 2 gm of tryptic digest on the column using an Eksigent NanoLC-Ultra™ 2D Plus system coupled with a SCIEX Triple TOF® 5600 system fitted with a NanoSpray III source. The samples were loaded on the trap (Eksigent Chrom XP 350 μm × 0.5 mm, 3 μm, 120 Å) and washed for 30 min at 3 μl/min. A 120 min gradient in multiple steps (ranging from 5 to 50% acetonitrile in water containing 0.1% formic acid) was set up to elute the peptides from the ChromXP 3-C18 (0.075 × 150 mm, 3 μm, 120 Å) analytical column. Technical replicates of the nanoLC-TripleTOF 5600 MS experiments were performed.

Sequential Window Acquisition of all Theoretical Fragment Ion Spectra (SWATH) Analysis for Label-Free Quantification

For label-free quantification (SWATH analysis), data-dependent analysis (DDA) mode was applied for both samples to generate high-quality spectral ion libraries by operating the mass spectrometer with specific parameters (Sharma and Bisht, 2016).

In the SWATH acquisition method, the Q1 transmission window was set to 12 Da from the mass range 350–1250 Da. A total of 75 windows were acquired independently with an accumulation time of 62 ms, along with three technical replicates for each of the sets. The total cycle time was kept constant at <5 s. Protein Pilot™ v. 5.0 was used to generate the spectral library. For label-free quantification, peak extraction and spectral alignment were performed using PeakView® 2.2 Software with the parameters set as follows: number of peptides, 2; number of transitions, 5; peptide confidence, 95%; XIC width, 30 ppm; XIC extraction window, 3 min. The data were further processed in MarkerView software v. 1.3 (AB Sciex, Foster City, CA, United States) for statistical data interpretation. In MarkerView, the peak area under the curve (AUC) for the selected peptides was normalized by the internal standard protein (beta-galactosidase) spike during SWATH accumulation. The results were extracted as three output files containing the AUC of the ions, the summed intensity of peptides for protein, and the summed intensity of ions for the peptide. All SWATH acquisition data were processed using SWATH Acquisition MicroApp 2.0 in PeakView® Software.

Data Analysis

Data were processed with Protein Pilot Software v. 5.0 (AB Sciex, Foster City, CA, United States) utilizing the Paragon and Progroup Algorithm. The analysis was done using the tools integrated into Protein Pilot at a 1% false discovery rate (FDR) with statistical significance. In brief, the UniProt database searched for the *K. pneumoniae* taxonomy, which was download from the database in July 2018. The download included total combined (reviewed and un-reviewed) entries of 409,060 proteins. We used cRAP analysis to identify the proteins commonly found in proteomics experiments (unavoidable contamination) of protein samples. *E. coli* beta-galactosidase (BGAL_ECOLI-[P00722]) was used as a molecular weight marker to calibrate the system for sample acquisition. The internal standard was used for the normalization of statistical parameters. We exported the label-free quantified data and imported them into MarkerView software V1.3 to obtain statistical data for further interpretation. Triplicate data for each sample were normalized using the internal protein (beta-galactosidase) area, which was initially spiked in the samples. After normalization, principal component analysis (PCA) was performed to check the possible correlated variables within the group. We plotted a volcano curve to determine the statistical significant fold change versus *p*-value for the control and test. Proteins with a significant fold change < 0.42 were considered down-regulated proteins. Peak extraction and spectral alignment were performed using PeakView software (v. 2.2, AB Sciex, Foster City, CA, United States) with the following parameter settings: number of peptides per protein, 5; number of transitions per peptide, 6; selected peptide confidence, 1% FDR; XIC width, 30 ppm; XIC extraction window, 3 min.

Gene Ontology Term Assignment and Analysis

Klebsiella pneumonia subsp. *pneumoniae* (strain ATCC 700721/MGH 78578) was used as a reference strain to carry out the functional studies. The proteins obtained from LC–MS/MS were firstly aligned to the reference strain proteome. Reference strain proteins that showed alignment identity of ≥50% over 80% of the sequence length of the reference strain protein were considered as homologs of the LC–MS/MS identified proteins. The Gene Ontology (GO) terms associated with the reference strain protein were used for the functional annotation. We used the slim version of GO terms, which were obtained from the Gene Ontology Consortium[1] (Camon et al., 2003).

[1]http://www.geneontology.org/

TABLE 1 | Details of the down-regulated proteome (flagella-, fimbriae-, and pili-related proteins) under meropenem stress in *Klebsiella pneumonia* clinical isolates (NDM-4).

S. No.	Protein name	Log fold change vs. *p*-value	Accession number	Protein symbol	Matched organism strain
1	Flagellar motor switch protein FliG	0.42	W1AYD1	FliG	*K. Pneumoniae* IS22
2	Flagellar hook protein FlgE	0.32	W1AUQ1	FlgE	*K. Pneumoniae* IS22
3	Negative regulator of flagellin synthesis FlgM	0.23	W1AWJ3	FlgM	*K. Pneumoniae* IS22
4	Putative fimbriae major subunit StbA	0.23	A6T548	StbA	*K. pneumoniae* subsp. *pneumoniae* (strain ATCC 700721/MGH 78578)
5	Flagellar hook-associated protein 2	0.22	W1B018	.....	*K. Pneumoniae* IS22
6	FimC protein	0.17	W1AP20	FimC	*K. Pneumoniae* IS22
7	Chaperone FimC	0.15	W1BBF2	FimC	*K. Pneumoniae* IS22
8	Flagellar biosynthesis protein FlgN	0.09	W1ATC5	FlgN	*K. Pneumoniae* IS22
9	Flagellar basal body protein	0.08	W1AT19	.....	*K. Pneumoniae* IS22
10	Conjugal transfer protein TraC	0.06	W1B0W2	TraC	*K. Pneumoniae* IS22
11	Fimbrial subunit type 1	0.05	W1B9X4	FimA	*K. Pneumoniae* IS22
12	Chemotaxis regulator-transmits chemoreceptor signals to flagelllar motor components CheY	0.04	W1BDF3	CheY	*K. Pneumoniae* IS22
13	Flagellin	0.01	W1AZS9	.....	*K. Pneumoniae* IS22

TABLE 2 | Functional analysis of down-regulated genes associated with *Klebsiella pneumonia* subsp. *pneumonia* (strain ATCC 700721/MGH 78578).

GO ID	Function	Gene name	No. of genes in which GO term was found
(A) Biological functions			
GO:0005975	Carbohydrate metabolic process	*deoC*, *dhaK*, *dhaL*, *glk*, *gmhB*, *gnd*, *lacZ2*, *malP*, *talB*, *treC*, *uxaC*	11
GO:0006091	Generation of precursor metabolites and energy	*aspA*, *fdhF*, *glk*, *gor*	4
GO:0006259	DNA metabolic process	*ung*, *uvrD*	2
GO:0006399	tRNA metabolic process	*gltX*, *pheS*, *thrS*	3
GO:0006412	Translation	*gltX*, *pheS*, *thrS*	3
GO:0006457	Protein folding	*fimC*	1
GO:0006461	Protein complex assembly	*hscB*	1
GO:0006464	Cellular protein modification process	*ppiD*, *ptsH*	2
GO:0006520	Cellular amino acid metabolic process	*aspA*, *gcvH*, *gcvP*, *gcvT*, *gltX*, *hisD*, *ilvC*, *pheS*, *thrS*	9
GO:0006629	Lipid metabolic process	*dxs*, *glpQ*	2
GO:0006790	Sulfur compound metabolic process	*dxs*, *gor*	2
GO:0006810	Transport	*artI*, *copA*, *pcoC*, *ptsH*	4
GO:0006950	Response to stress	*sodB*, *ung*	2
GO:0007155	Cell adhesion	*fimA*	1
GO:0007165	Signal transduction	*artI*	1
GO:0009056	Catabolic process	*deoC*, *gcvH*, *gcvP*, *gcvT*, *glk*, *glpA*, *treC*, *uxaC*	8
GO:0009058	Biosynthetic process	*dxs*, *hisD*, *hns*, *ilvC*, *nadE*, *pdxY*, *rmlD*, *rnk*, *sul*, *udp*, *uvrD*	11
GO:0034641	Cellular nitrogen compound metabolic process	*cpdB*, *dxs*, *glk*, *gnd*, *gor*, *hisD*, *hns*, *nadE*, *pdxY*, *rmlB*, *rmlD*, *rnk*, *sul*, *talB*, *udp*	15
GO:0034655	Nucleobase-containing compound catabolic process	*cdd*, *cpdB*, *deoC*, *udp*	4
GO:0042592	Homeostatic process	*Gor*	1
GO:0044281	Small molecule metabolic process	*aspA*, *cdd*, *cpdB*, *deoC*, *dhaK*, *dhaL*, *dxs*, *fdhF*, *glk*, *glpA*, *gnd*, *nadE*, *pdxY*, *sul*, *talB*, *udp*, *uxaC*	17
GO:0051186	Cofactor metabolic process	*glk*, *gnd*, *nadE*, *pdxY*, *sul*, *talB*	6
GO:0051276	Chromosome organization	*uvrD*	1
GO:0051604	Protein maturation	*hscB*	1
GO:0055085	Transmembrane transport	*copA*	1
GO:0071554	Cell wall organization or biogenesis	*fimC*	1
(B) Molecular functions			
GO:0003677	DNA binding	*hns*, *rnk*, *uvrD*	3
GO:0003723	RNA binding	*gltX*, *pheS*, *thrS*	3
GO:0004386	Helicase activity	*uvrD*	1
GO:0004871	Signal transducer activity	*artI*	1
GO:0008168	Methyltransferase activity	*gcvT*	1
GO:0016301	Kinase activity	*dhaK*, *dhaL*, *glk*, *pdxY*, *ptsH*, *rnk*	6
GO:0016491	Oxidoreductase activity	*KPN_02441*, *fdhF*, *gcvP*, *glpA*, *gnd*, *gor*, *hisD*, *ilvC*, *nfnB*, *nfsA*, *rmlD*, *sodB*, *ydgJ*	13
GO:0016746	Transferring acyl groups	*Maa*	1
GO:0016757	Transferring glycosyl groups	*malP*, *udp*	2
GO:0016765	Transferring alkyl or aryl (other than methyl) groups	*Sul*	1
GO:0016791	Phosphatase activity	*aphA*, *gmhB*	2
GO:0016798	Acting on glycosyl bonds	*lacZ2*, *rihC*, *treC*, *ung*	4
GO:0016810	Acting on carbon–nitrogen (but not peptide) bonds	*Cdd*	1
GO:0016829	Lyase activity	*acnA*, *aspA*, *deoC*, *rmlB*, *yhbL*	5
GO:0016853	Isomerase activity	*ppiD*, *uxaC*	2
GO:0016874	Ligase activity	*gltX*, *nadE*, *pheS*, *thrS*	4
GO:0016887	ATPase activity	*copA*, *uvrD*	2
GO:0019899	Enzyme binding	*Rnk*	1
GO:0022857	Transmembrane transporter activity	*artI*, *copA*	2
GO:0030234	Enzyme regulator activity	*hscB*	1

(Continued)

TABLE 2 | Continued

GO ID	Function	Gene name	No. of genes in which GO term was found
(B) Molecular functions			
GO:0043167	Ion binding	*KPN_pKPN3p05899*, *aphA*, *cdd*, *copA*, *cpdB*, *dxs*, *fdhF*, *glk*, *glpA*, *gltX*, *gmhB*, *gor*, *hisD*, *ilvC*, *lacZ2*, *malP*, *nadE*, *pcoC*, *pdxY*, *pheS*, *sodB*, *sul*, *thrS*, *uvrD*, *yiiM*	25
(C) Cellular component			
GO:0005622	Intracellular	*artI*, *copA*, *lacZ2*, *ppiD*	1
GO:0005623	Cell	*Hns*	6
GO:0005737	Cytoplasm	*aphA*, *artI*, *fimA*, *fimC*, *gor*, *pcoC*	15
GO:0005886	Plasma membrane	*deoC*, *gcvH*, *glk*, *glpA*, *gltX*, *gmhB*, *maf*, *pheS*, *ptsH*, *talB*, *thrS*, *treC*, *udp*, *ung*, *uvrD*	2

TABLE 3 | List of *Klebsiella pneumonia* sp. proteins mapped on *E. coli* K12 substr.DH10B.

K. pneumoniae IS22 protein entry	Sequence length	*K. pneumoniae* protein name	*E. coli* K12 substr. DH10B protein entry	Sequence length	*E. coli* K12 substr. DH10B gene name	Identity (%)
W1AYD1	331	Flagellar motor switch protein FliG	P0ABZ1	331	fliG	99.698
W1AUQ1	206	Flagellar hook protein FlgE	P75937	402	flgE	91.304
W1AWJ3	97	Negative regulator of flagellin synthesis	P0AEM4	97	flgM	100.000
W1B018	468	Flagellar hook-associated protein 2	P24216	97	fliD	99.786
W1AP20	224	Chaperone FimC	No hit	–	–	–
W1ATC5	138	Flagellar biosynthesis protein FlgN	P43533	468	flgN	100.000
W1AT19	191	Flagellar basal body protein	P75937	138	flgE	95.812
W1BDF3	129	Chemotaxis regulator-transmits chemoreceptor signals to flagelllar motor components CheY	P0AE67	129	cheY	100.000
W1B0W2	533	Conjugal transfer protein traC	No hit	–	–	–
W1AZS9	447	Flagellin	No hit	–	–	–
A6T548	178	Putative fimbriae major subunit StbA	No hit			

Protein–Protein Interaction Network Integration

To find the interaction partner(s) of down-regulated proteins, protein–protein interaction (PPI) information were obtained from the STRING database v10.0[2] and Cytoscape (version 3.6.1) (Shannon et al., 2003; Sharma and Bisht, 2017a,b,c; Sharma et al., 2018b, 2019b; Sharma and Khan, 2018). The PPI information provided by the STRING database has been established by experimental studies or by genomic analyses like domain fusion, phylogenetic profiling high-throughput experiments, co-expression studies, and gene neighborhood analysis. In the present study, only interactions with a score of $\geq$0.4 were used.

RESULTS

Identification of Proteins by LC–MS/MS

In this study, we grew the carbapenem-resistant isolate at meropenem sub-MIC 32 mg/L. Further, we identified the down-regulated proteome of the same bacteria by LC–MS/MS using a SWATH workflow; 1156 proteins were quantified at 1% FDR with statistical significance as per the log fold change vs. *p*-value. Among them, 69 proteins were down-regulated (<0.42 log fold change vs. *p*-value) and are tabulated in the **Supplementary Material** (**Supplementary Table S1**).

In the present work, our main focus was on motility-related proteins. After critical analysis of the 69 down-regulated proteins, we found 13 proteins (around 19%) that belonged to the flagella-, fimbriae-, and pili-related protein functional groups (**Table 1**). These proteins are flagellar motor switch protein FliG, flagellar hook protein FlgE, negative regulator of flagellin synthesis FlgM, putative fimbriae major subunit StbA, flagellar hook-associated protein 2, chaperone FimC protein, flagellar biosynthesis protein FlgN, flagellar basal body protein, conjugal transfer protein TraC, fimbrial subunit type 1, chemotaxis regulator-transmits chemoreceptor signals to flagellar motor components CheY, and flagellin, all of which are involved in motility and its supporting processes.

On the basis of the parameters described in the section "Materials and Methods," we were able to map 67 out of the 69 down-regulated proteins on the proteome of the ATCC 700721/MGH 78578 strain of *K. pneumoniae* (**Supplementary Table S2**). Our GO results for down-regulated genes also show that most down-regulated proteins were involved in

[2]http://www.string-db.org/

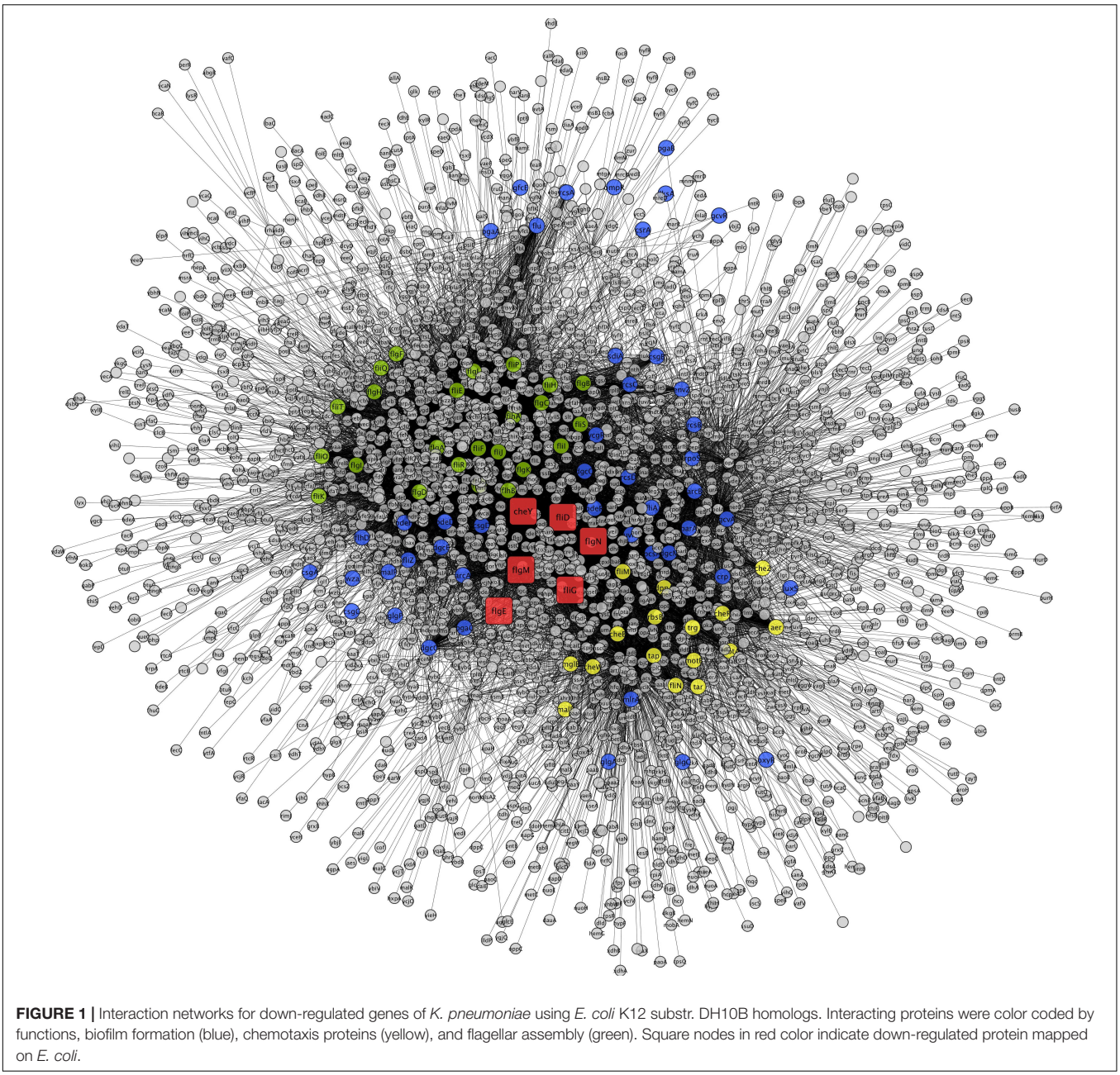

FIGURE 1 | Interaction networks for down-regulated genes of *K. pneumoniae* using *E. coli* K12 substr. DH10B homologs. Interacting proteins were color coded by functions, biofilm formation (blue), chemotaxis proteins (yellow), and flagellar assembly (green). Square nodes in red color indicate down-regulated protein mapped on *E. coli*.

nitrogen and other small molecule metabolism, ion binding, and oxidoreductase activity (**Table 2**).

Protein Network Analysis

To construct the PPI network, the down-regulated proteins with motility-related functions were annotated using *E. coli* K12 strain DH10B homologs, as tabulated in **Table 3**. Among them, few proteins showed no hits in *E. coli*, and the rest of the proteins interacted with other proteins to make an interactome; this was visualized through Cytoscape (version 3.6.1) (**Figure 1**). Interacting proteins were color-coded by their functions: biofilm formation, blue; chemotaxis proteins, yellow; flagellar assembly, green. Square red nodes indicate down-regulated proteins.

DISCUSSION

The development of carbapenem-resistant *K. pneumoniae* has worsened the medical situation around the globe. The emergence of carbapenem resistance is usually due to the over-expression of carbapenemases and loss of porins. Recently, we reported a cluster of over-expressed proteins in carbapenem-resistant *K. pneumoniae* under meropenem pressure. These could be responsible for the drug resistance and belong to various categories such as the protein translational machinery complex, DNA/RNA modifying enzymes or proteins, proteins involved in carbapenem cleavage, modification, and transport, and energy metabolism- and intermediary metabolism-related

proteins (Sharma et al., 2019a). Therefore, we suggested that they could be potential targets for the development of novel therapeutics against this resistance. Biofilm formation is also one of the mechanisms that leads to the development of drug resistance. In the present study, we have found a group flagellar, fimbriae, and pili proteins that are down-regulated under meropenem stress (sub-MIC). We hypothesize that the down-regulation of these proteins under meropenem stress makes the bacteria sessile or non-motile, leading to the emergence of a biofilm-like state that could contribute to carbapenem resistance in *K. pneumoniae* (NDM-4). Earlier microarray analysis of *K. pneumoniae* also reported the down-regulation of genes related to nitrogen metabolism, porin genes, and some membrane-associated proteins in association with the antibiotic resistance phenomenon (Doménech-Sánchez et al., 2006). The significance of the aforementioned pathways in antibiotic-evading mechanisms is also highlighted in several other reports (Yeung et al., 2011; Piek et al., 2014).

A Hub of Flagellar, Fimbriae, and Pili Proteins Could Regulate Resistance

A group of flagellar, fimbriae, and pili proteins involved in the formation of the flagellar machinery complex and the regulation of motility processes were found to be down-regulated in meropenem-induced carbapenem-resistant *K. pneumoniae* clinical strains. These proteins are flagellar motor switch protein FliG, flagellar hook protein FlgE, negative regulator of flagellin synthesis FlgM, putative fimbriae major subunit StbA, flagellar hook-associated protein 2, chaperone FimC protein, flagellar biosynthesis protein FlgN, flagellar basal body protein FlgF, conjugal transfer protein TraC, fimbrial subunit type 1, chemotaxis regulator-transmits chemoreceptor signals to flagellar motor components CheY, and flagellin.

The observed expression down-regulation of flagella-, fimbriae-, and pili-related proteins provides clues to a novel mechanism of drug resistance. Flagellin, flagellar biosynthesis protein FlgN, and FlgM, respectively, cumulatively maintain equilibrium in the biosynthesis of flagella. Flagellin protein is part of the structural component of flagella, and biosynthesis of flagella is favored by flagellar biosynthesis protein FlgN (Bennett et al., 2001). FlgM is a negative regulator that switches off flagellar transport. FlgM can be exported from the cell via a flagellum, and the transport occurs only after the completion of hook formation (Hughes et al., 1993). This unique regulatory mechanism further postpones flagellin synthesis in the cell. Cumulatively, the expression of these proteins is involved in the flagella formation, regulation, and motility of the bacteria. Therefore, we suggest that, under meropenem pressure, down-regulation of these proteins might make the bacteria sessile, which is the first step in biofilm-formation. Therefore, we assume that down-regulation of these proteins could create a biofilm-like state, which ultimately leads to the drug resistance.

Flagella are composed of three different parts: a filament (helical and long), hook (a curved and short structure), and basal body (a complex structure with a central rod and a series of rings). Flagellar motor switch protein (FliG), flagellar hook protein FlgE, flagellar hook-associated protein 2, and flagellar basal body protein FlgF together make up part of the flagellar motor switch complex (FliG, FliN, and FliM), which is involved in bacterial motility, after receipt and transduction of the signal by chemotaxis (Djordjevic and Stock, 1998). This is a complex apparatus that senses the signal from the chemotaxis sensory signaling system and is transduced into motility. Chemotaxis response regulator CheY transmits chemoreceptor signals to flagellar motor components and is believed to be the "on" switch that directly induces tumbles in the swimming pattern (Robinson et al., 2000). Physical interactions of CheY and switch proteins have not been reported. Chemotactic stimuli change the association of the CheY signal protein with the distal $FliM_{NC}$FliN C ring (Dyer et al., 2009; Sarkar et al., 2010). In the present study, down-regulation of the chemotaxis response regulator (CheY) subsequently down-regulates signal transmission to the flagellar motor components, which may act as an "off" switch and make the bacteria sessile or non-motile. Further, it might lead to a biofilm-like state and could contribute to the drug resistance.

Putative fimbriae major subunit StbA, Fimbrial subunit type 1, conjugal transfer protein TraC, and chaperone FimC protein are involved in pillus organization, fimbriae formation, and their associated assemblies (Johnson and Clegg, 2010). FimC protein acts as a periplasmic pilin chaperone that not only protects the bacteria under stress through chaperone-mediated folding but is also involved in pillus formation (Klemm, 1992). In the periplasm, the FimC chaperone binds to the major and minor structural components and protects them from degradation. Conjugal transfer protein TraC, encoded by a gene, *traC*, presents on plasmid and is involved in the conjugation process as well as pili formation (Winans and Walker, 1985; Kado, 1994), leading

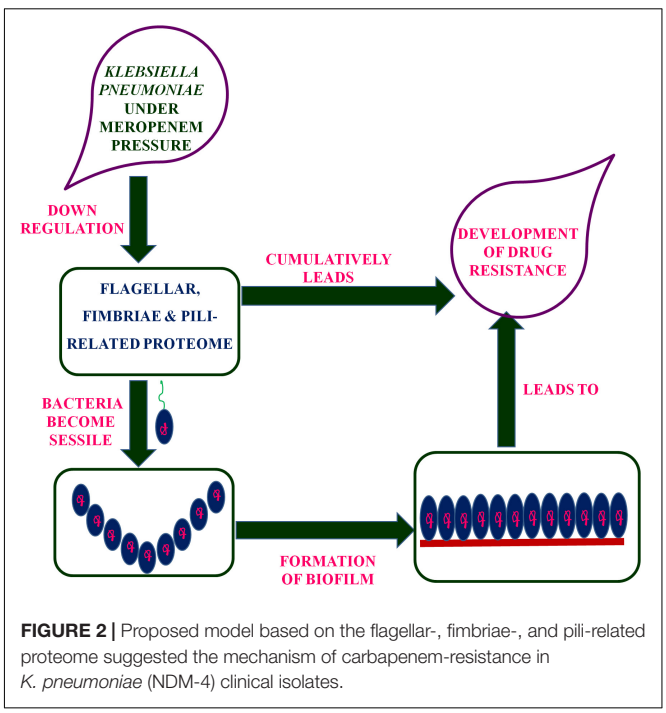

FIGURE 2 | Proposed model based on the flagellar-, fimbriae-, and pili-related proteome suggested the mechanism of carbapenem-resistance in *K. pneumoniae* (NDM-4) clinical isolates.

to the transfer of drug-resistant plasmid to bacteria. These down-regulated proteins (flagellar, fimbriae, and pili proteins) form a hub of proteins in the PPI network, which indicates their important role in flagellar, fimbriae, and pili assemblies, signaling through chemotaxis proteins, and biofilm formation (**Figure 1**). In this study, the down-regulation of fimbriae-, pili-, and conjugative-related proteins leads to the creation of the biofilm-like state, which may contribute to drug resistance. On the basis of the flagella-, fimbriae-, and pili-related proteome, we propose a model (**Figure 2**) that suggests the potential path or mechanism of carbapenem resistance in *K. pneumoniae* (NDM-4) clinical isolates. Overall this group of down-regulated proteins and their interactive protein partners cumulatively make a hub that leads to the formation of a biofilm-like scenario and could contribute to meropenem resistance.

CONCLUSION

In brief, the present study focused on the down-regulated proteome of carbapenem-resistant *K. pneumoniae* clinical isolate (NDM-4) under meropenem pressure through proteomics and systems biology approaches. A group of down-regulated proteins was identified that belongs to the flagellar, fimbriae, and pili proteins. Therefore, we suggest that these proteins and their interactive protein partners cumulatively lead to the bacteria becoming sessile, which further creates a biofilm-like state and could contribute to the survival of bacteria under meropenem pressure, which might reveal a novel mechanism of drug resistance. Further research on these motility-related protein targets and their pathways may lead to the development of novel therapeutics against the worsening situation of drug resistance.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

DS designed the concept, and experimented and wrote the manuscript. AG and MK carried out the systems biology analysis. FR provided support in the LC–MS experiments and analysis. AK designed and guided the study and finalized the manuscript. All authors approved the final manuscript.

FUNDING

Science and Engineering Research Board (SERB) is gratefully acknowledged for providing fellowship and funds to DS (SERB-N-PDF/2016/001622) to work at IBU-AMU Aligarh. The authors also acknowledge SCIEX Pvt. Ltd., Gurgaon, India, for data acquisition and proteomics facility support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.02865/full#supplementary-material

REFERENCES

Ahmad, N., Khalid, S., Ali, S. M., and Khan, A. U. (2018). Occurrence of blaNDM variants among Enterobacteriaceae from a neonatal intensive care unit in a northern India hospital. *Front Microbiol.* 9:407. doi: 10.3389/fmicb.2018.00407

Ambler, R. P., Coulson, A. F., Frère, J. M., Ghuysen, J. M., Joris, B., Forsman, M., et al. (1991). A standard numbering scheme for the class A beta-lactamases. *Biochem. J.* 276(Pt 1), 269–270. doi: 10.1042/bj2760269

Bennett, J. C., Thomas, J., Fraser, G. M., and Hughes, C. (2001). Substrate complexes and domain organization of the *Salmonella flagellar* export chaperones FlgN and FliT. *Mol. Microbiol.* 39, 781–791. doi: 10.1046/j.1365-2958.2001.02268.x

Bradford, M. M. (1976). A rapid and sensitive method for the quantification of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72, 248–254. doi: 10.1006/abio.1976.9999

Camon, E., Barrell, D., Lee, V., Dimmer, E., and Apweiler, R. (2003). The Gene Ontology Annotation (GOA) database-an integrated resource of GO annotations to the UniProt Knowledgebase. *Silico Biol.* 4, 5–6.

Costerton, J. W., and Lewandowski, Z. (1995). Microbial biofilms. *Annu. Rev. Microbiol.* 49, 711–745.

Costerton, J. W., Stewart, P. S., and Greenberg, E. P. (1999). Bacterial biofilms: a common cause of persistent infections. *Science* 1999, 1318–1322. doi: 10.1126/science.284.5418.1318

Djordjevic, S., and Stock, A. M. (1998). Structural analysis of bacterial chemotaxis proteins: com-ponents of a dynamic signaling system. *J. Struct. Biol.* 124, 189–200. doi: 10.1006/jsbi.1998.4034

Doménech-Sánchez, A., Javier, Benedi V, Martínez-Martínez, L., and Alberti, S. (2006). Evaluation of differential gene expression in susceptible and resistant clinical isolates of *Klebsiella pneumoniae* by DNA microarray analysis. *Clin. Microbiol. Infect.* 12, 936–940. doi: 10.1111/j.1469-0691..01470.x

Dyer, C. M., Vartanian, A. S., Zhou, H., and Dahlquist, F. W. (2009). A molecular mechanism of bacterial flagellar motor switching. *J. Mol. Biol.* 388, 71–84. doi: 10.1016/j.jmb.2009.02.004

Fux, C. A., Costerton, J. W., Stewart, P. S., and Stoodley, P. (2005). Survival strategies of infectious biofilms. *Trends Microbiol.* 13, 34–40. doi: 10.1016/j.tim.2004.11.010

Hughes, K. T., Gillen, K. L., Semon, M. J., and Karlinsey, J. E. (1993). Sensing structural intermediates in bacterial flagellar assembly by export of a negative regulator. *Science* 262, 1277–1280. doi: 10.1126/science.8235660

Jacoby, G. A., Mills, D. M., and Chow, N. (2004). Role of beta-lactamases and porins in resistance to Ertapenem and other beta-lactams in *Klebsiella pneumoniae*. *Antimicrob. Agents Chemother.* 48, 3203–3206. doi: 10.1128/aac.48.8.3203-3206.2004

Johnson, J. G., and Clegg, S. (2010). Role of MrkJ, a phosphodiesterase, in type 3 fimbrial expression and biofilm formation in *Klebsiella pneumoniae*. *J. Bacteriol.* 192, 3944–3950. doi: 10.1128/JB.00304-10

Kado, C. I. (1994). Promiscuous DNA transfer system of *Agrobacterium tumefaciens*: role of the virB operon in sex pilus assembly and synthesis. *Mol. Microbiol.* 12, 17–22. doi: 10.1111/j.1365-2958.1994.tb00990.x

Khan, A., Sharma, D., Faheem, M., Bisht, D., and Khan, A. U. (2017). Proteomic analysis of a carbapenem-resistant *Klebsiella pneumoniae* strain in response to meropenem stress. *J. Glob. Antimicrob. Resist.* 8, 172–178. doi: 10.1016/j.jgar.2016.12.010

Klemm, P. (1992). FimC, a chaperone-like periplasmic protein of *Escherichia coli* involved in biogenesis of type 1 fimbriae. *Res. Microbiol.* 143, 831–838. doi: 10.1016/0923-2508(92)90070-5

Lata, M., Sharma, D., Deo, N., Tiwari, P. K., Bisht, D., and Venkatesan, K. (2015). Proteomic analysis of ofloxacin-mono resistant *Mycobacterium tuberculosis* isolates. *J. Proteomics* 127, 114–121. doi: 10.1016/j.jprot.2015.07.031

Loli, A., Tzouvelekis, L. S., Tzelepi, E., Carattoli, A., Vatopoulos, A. C., Tassios, P. T., et al. (2006). Sources of diversity of carbapenem resistance levels in *Klebsiella pneumoniae* carrying blaVIM-1. *J. Antimicrob. Chemother.* 58, 669–672. doi: 10.1093/jac/dkl302

Martínez-Martínez, L., Pascual, A., Hernández-Allés, S., Alvarez-Díaz, D., Suárez, A. I., Tran, J., et al. (1999). Roles of β-lactamases and porins in activities of carbapenems and cephalosporins against *Klebsiella pneumoniae*. *Antimicrob. Agents Chemother.* 43, 1669–1673. doi: 10.1128/aac.43.7.1669

McCarty, S. M., Cochrane, C. A., Clegg, P. D., and Percival, S. L. (2012). The role of endogenous and exogenous enzymes in chronic wounds: a focus on the implications of aberrant levels of both host and bacterial proteases in wound healing. *Wound Repair Regen.* 20, 125–136. doi: 10.1111/j.1524-475X.2012.00763.x

Paterson, D. L. (2000). Recommendation for treatment of severe infections caused by Enterobacteriaceae producing extended-spectrum beta-lactamases(ESBLs). *Clin. Microbiol. Infect.* 6, 460–463. doi: 10.1046/j.1469-0691.2000.00107.x

Paterson, D. L., and Bonomo, R. A. (2005). Extended-spectrum beta-lactamases: a clinical update. *Clin. Microbiol. Rev.* 18, 657–686. doi: 10.1128/cmr.18.4.657-686.2005

Piek, S., Wang, Z., Ganguly, J., Lakey, A. M., Bartley, S. N., Mowlaboccus, S., et al. (2014). The role of oxidoreductases in determining the function of the neisserial lipid A phosphoethanolamine transferase required for resistance to polymyxin. *PLoS One* 9:e106513. doi: 10.1371/journal.pone.0106513

Qayyum, S., Sharma, D., Bisht, D., and Khan, A. U. (2019). Identification of factors involved in *Enterococcus faecalis* biofilm under quercetin stress. *Microb. Pathog.* 126, 205–211. doi: 10.1016/j.micpath.2018.11.013

Robinson, V. L., Buckler, D. R., and Stock, A. M. (2000). A tale of two components: a novel kinase and a regulatory switch. *Nat. Struct. Biol.* 7, 626–633.

Sarkar, M. K., Paul, K., and Blair, D. (2010). Chemotaxis signaling protein CheY binds to the rotor protein FliN to control the direction of flagellar rotation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9370–9375. doi: 10.1073/pnas.1000935107

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharma, D., and Bisht, D. (2016). An efficient and rapid method for enrichment of lipophilic proteins from *Mycobacterium tuberculosis* H37Rv for two dimensional gel electrophoresis. *Electrophoresis* 37, 1187–1190. doi: 10.1002/elps.201600025

Sharma, D., and Bisht, D. (2017a). M.tuberculosis hypothetical proteins and proteins of unknown function: hope for exploring novel resistance mechanisms as well as future target of drug resistance. *Front. Microbiol.* 8:465. doi: 10.3389/fmicb.2017.00465

Sharma, D., and Bisht, D. (2017b). Role of bacterioferritin & ferritin in *M.tuberculosis* pathogenesis and drug resistance: a future perspective by interactomic approach. *Front. Cell. Infect. Microbiol.* 7:240. doi: 10.3389/fcimb.2017.00240

Sharma, D., and Bisht, D. (2017c). Secretory proteome analysis of streptomycin resistant *Mycobacterium tuberculosis* clinical isolates. *SLAS Discov.* 22, 1229–1238. doi: 10.1177/2472555217698428

Sharma, D., Bisht, D., and Khan, A. U. (2018a). Potential alternative strategy against drug resistant tuberculosis: a proteomics prospect. *Proteomes* 6:26. doi: 10.3390/proteomes6020026

Sharma, D., Singh, R., Deo, N., and Bisht, D. (2018b). Interactome analysis of Rv0148 to predict potential targets and their pathways linked to aminoglycosides drug resistance: an insilico approach. *Microb. Pathog.* 121, 179–183. doi: 10.1016/j.micpath.2018.05.034

Sharma, D., Garg, A., Kumar, M., and Khan, A. U. (2019a). Proteome profiling of carbapenem-resistant *K. pneumoniae* clinical isolate (NDM-4): exploring the mechanism of resistance and potential drug targets. *J. Proteomics* 200, 102–110. doi: 10.1016/j.jprot.2019.04.003

Sharma, D., Lata, M., Faheem, M., Khan, A. U., Joshi, B., Venkatesan, K., et al. (2019b). Role of *M.tuberculosis* protein Rv2005c in the aminoglycosides resistance. *Microb. Pathog.* 132, 150–155. doi: 10.1016/j.micpath.2019.05.001

Sharma, D., Misba, L., and Khan, A. U. (2019c). Antibiotics versus Biofilm: an emerging battleground in microbial communities. *Antimicrob. Resist. Infect. Control* 8:76. doi: 10.1186/s13756-019-0533-3

Sharma, D., and Khan, A. U. (2018). Role of cell division protein divIVA in *Enterococcus faecalis* pathogenesis, biofilm and drug resistance: a future perspective by in silico approaches. *Microb. Pathog.* 125, 361–365. doi: 10.1016/j.micpath.2018.10.001

Stewart, P. S., and William Costerton, J. (2001). Antibiotic resistance of bacteria in biofilms. *Lancet* 358, 135–138. doi: 10.1016/s0140-6736(01)05321-1

Wayne, P. A. (2014). Performance standards for antimicrobial susceptibility testing: 24 informational supplement. *CLSI* 100, S24.

Winans, S. C., and Walker, G. C. (1985). Conjugal transfer system of the IncN plasmid pKM101. *J. Bacteriol.* 161, 402–410.

Yeung, A. T., Bains, M., and Hancock, R. E. (2011). The sensor kinase CbrA is a global regulator that modulates metabolism, virulence, and antibiotic resistance in *Pseudomonas aeruginosa*. *J. Bacteriol.* 15, 918–931. doi: 10.1128/JB.00911-10

Conflict of Interest: FR was employed by SCIEX Pvt. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 26 May 2020
doi: 10.3389/fmicb.2020.00915

Biophysical and Biochemical Characterization of Nascent Polypeptide-Associated Complex of *Picrophilus torridus* and Elucidation of Its Interacting Partners

Neelja Singhal[1†], Archana Sharma[1†], Shobha Kumari[1], Anjali Garg[1], Ruchica Rai[1], Nirpendra Singh[2], Manish Kumar[1] and Manisha Goel[1]*

[1] Department of Biophysics, University of Delhi, New Delhi, India, [2] Regional Centre for Biotechnology, NCR-Biotech Science Cluster, Faridabad, India

OPEN ACCESS

Edited by:
Jerry Eichler, Ben-Gurion University of the Negev, Israel

Reviewed by:
Bruno Franzetti, GDR3635 Biodiversité, Origine, Processus Cellulaires Fondamentaux, Biotechnologies (Archaea), France
Masafumi Yohda, Tokyo University of Agriculture and Technology, Japan

***Correspondence:**
Manisha Goel
manishagoel@south.du.ac.in

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Biology of Archaea, a section of the journal Frontiers in Microbiology

Received: *20 December 2019*
Accepted: *17 April 2020*
Published: *26 May 2020*

Citation:
Singhal N, Sharma A, Kumari S, Garg A, Rai R, Singh N, Kumar M and Goel M (2020) Biophysical and Biochemical Characterization of Nascent Polypeptide-Associated Complex of Picrophilus torridus and Elucidation of Its Interacting Partners. Front. Microbiol. 11:915. doi: 10.3389/fmicb.2020.00915

Nascent polypeptide-associated complex (NAC) is a ribosome-associated molecular chaperone which is present only in archaea and eukaryotes. The primary function of NAC is to shield the newly synthesized polypeptide chains from inappropriate interactions with the cytosolic factors. Besides that, NAC has been implicated in diverse biological functions, which suggest that it might be a multifunctional protein. An elaborate study on NAC can provide useful information on protein folding in extreme conditions in which many archaea grow. Thus, in the present study, we have studied the biophysical and the biochemical characteristics of NAC of *Picrophilus torridus*, an extreme thermoacidophilic archaeon. The study of protein–protein interactions and binding partners of a protein provides useful insights into the new/unreported roles of a protein. Thus, in this study, we have identified the binding partners of NAC in *P. torridus*. The NAC protein of *P. torridus* was cloned, expressed, and purified, and its binding partners were isolated by a pull down assay followed by identification with liquid chromatography–mass spectrometry. To the best of our knowledge, this is the first report on the biophysical and the biochemical characterization of NAC from *P. toridus* and the identification of its interacting partners.

Keywords: chaperone, circular dichroism, liquid chromatography–mass spectrometry, STRING, interactome

INTRODUCTION

Nascent polypeptide-associated complex (NAC) is a ribosome-associated molecular chaperone which is present near the peptide exit site of the translating ribosomes (Wiedmann et al., 1994). The major function associated with the NAC complex is thought to be to shield the newly synthesized polypeptide chains from inappropriate interactions with the cytosolic factors. Since most of the translating ribosomes likely associated with NAC, it was considered an abundant cellular protein, expressed in equimolar quantities relative to ribosomes (Raue et al., 2007; Hoffmann et al., 2010). Ribosome-associated chaperones vary across the three domains of life. While the prokaryotes employ the bacterial trigger factors in preventing the inappropriate folding of newly synthesized

polypeptide chains (Rospert et al., 2002; Kaiser et al., 2006) in eukaryotes and archaea, this function is performed by NAC (Koplin et al., 2010; del Alamo et al., 2011). In eukaryotes, NAC is reported to be a heterodimeric protein composed of alpha- and beta-subunits, while in archaea, it was reportedly a homodimer composed of only alpha-subunits (Spreter et al., 2005).

Several recent studies indicate that, besides its primary function as a molecular chaperone, NAC is involved in many other biological functions. It has been reported to be involved in the translation and the subcellular targeting of nascent polypeptides (Wickner, 1995; Powers and Walter, 1996) and in the prevention of mistargeting of ribosome nascent chain complex (Arsenovic et al., 2012; Gamerdinger et al., 2015). Besides that, NAC has been functionally implicated in the folding of the polypeptide chain, in ribosome biogenesis, in modulating protein secretion, in regulating apoptosis, as a transcription factor, etc. (Creagh et al., 2009; Kogan and Gvozdev, 2014). Despite being implicated in diverse cellular functions, its *in vivo* role and knowledge regarding its mechanism of action are still fragmentary. An elaborate study on NAC is expected to provide further insights into protein folding in extreme conditions in which many archaea grow. Thus, in the present study, we have performed biophysical and biochemical characterization of NAC of *Picrophilus torridus* to understand its role, if any, in the thermoacidophilic adaptation of this archaeon. *P. torridus* is an interesting organism which has some unique characteristics like small genome size (1.55 Mbp), adaptability to thrive in extremely low pH (0–1) and moderately high temperatures (50–60°C), and with an intracellular pH of about 4.0 (Schleper et al., 1995).

The study of protein–protein interactions (PPIs) adds crucially to our understanding of protein function(s) and helps in the characterization of the various pathways in which cellular proteins might be involved. The identification of the binding partners of a protein provides useful insights into the new/unreported roles of a protein. Also, an analysis of the interaction network aids in identifying the mechanism(s) underlying various related biological processes. Thus, in this study we have identified the binding partners of NAC in *P. torridus*. The NAC protein of *P. torridus* was cloned, expressed, and purified, and its binding partners were isolated by a pull down assay followed by identification using mass spectrometry. To the best of our knowledge, this is the first report on the biophysical and the biochemical characterization of NAC of *P. torridus* and the identification of its interacting partners.

MATERIALS AND METHODS

Cloning and Expression of *P. torridus* Nascent Polypeptide-Associated Chaperone

The gene encoding PtNAC (327 bases) (accession no. AE017261.1) was synthesized at a commercial facility (Genscript, United States) in pUC57 plasmid with flanking *Nhe*I and *Sal*I restriction sites. The plasmid with the insert was digested with restriction enzymes *Nhe*I and *Sal*I (30-µl reaction mixture: 3.0 µl 1X buffer, 25 µl plasmid DNA with insert, 1.0 µl *Nhe*I, and 1.0 µl *Sal*I) to obtain the free insert. The insert obtained was purified from the gel using a gel extraction kit and ligated with similarly digested pET28a(+) vector in 10 µl reaction mixture [ligation reaction in 10 µl reaction mixture: 1.0 µl ligase buffer, 1.0 µl insert, 7.5 µl digested pET28a(+) plasmid, and 0.5 µl ligase enzyme]. *Escherichia coli* DH5α cells were transformed with the pET28a(+)-PtNAC construct. Colony PCR and double-restriction digestion with enzymes *Nhe*I and *Sal*I were used for the screening of the transformants.

The pET28a(+)-PtNAC construct was isolated from the transformed *E. coli* DH5α cells and further transformed in the expression vector *E. coli* BL21 (DE3). The transformants were grown overnight in 50 ml Luria–Bertani broth, containing 50 µg ml^{-1} kanamycin, at 37°C and 200 rpm and inoculated (2%, v/v) again in 1 L of the same medium. The culture was kept for incubation at 37°C and 200 rpm until A_{600} of 0.6 was achieved when the protein expression was induced by adding 1 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG). The induced cells were harvested after 4 h by centrifuging at 8,000 × *g*, resuspended in 70 ml of lysis buffer (1X phosphate-buffered saline, PBS; pH 7.2), and sonicated (Sonics, Vibra^{TM}cell, Connecticut, United States) intermittently for 30 min to release the intracellular protein. The cell debris was removed by centrifugation at 10,000 × *g* (30 min at 4°C). Column chromatography was performed on an AKTA Prime Plus (GE Healthcare) system. The clear supernatant was filtered through a 0.22-µm membrane (mdi: SY25KG-S) and applied to 10 ml Co^{2+}-NTA beads (G Biosciences) packed in a XK16 column (Pharmacia Biotech) pre-equilibrated with 1X PBS, pH 7.2. The column was washed with five column volumes of equilibration buffer and then with wash buffer W1 (1X PBS with 20 mM imidazole), and the recombinant PtNAC was eluted using 1X PBS containing 200 mM imidazole. To remove imidazole, the eluted PtNAC protein was dialyzed in 1X PBS buffer (pH 7.2) overnight. The purity and the identity of the expressed protein was determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and matrix-assisted laser desorption/ionization (MALDI-TOF) analysis, respectively.

Chaperone Assay

The chaperone activity of PtNAC was evaluated by its capability to prevent the thermal aggregation of the substrate protein, bovine carbonic anhydrase (BCA II) (Rajaraman et al., 1996; Tomar et al., 2013). A cuvette with the sample (2 ml) was inserted in the pre-heated sample chamber (65°C) of a spectrofluorimeter containing a Peltier-controlled stirred cell, and scattering was monitored with time. The excitation and emission monochromators were set at 400 nm. BCA II (0.75 µM) heated at 65°C in 50 mM Tris-HCl buffer (pH 7.5) was used as the substrate control. Keeping the BCA II concentration (0.75 µM) constant, PtNAC was added in increasing concentration ratios, i.e., 1:0.5, 1:1, 1:2, and 1:4. For this experiment, purified NAC was dialyzed overnight in 50 mM Tris-HCl buffer (pH 7.5) and then used for estimating the chaperone activity.

Biophysical Characterization of PtNAC Using Circular Dichroism Spectroscopy

Circular dichroism (CD) experiments were conducted in a Jasco J-815 spectropolarimeter with a Peltier-type temperature controller (Jasco CDF-426 S/15). The far-UV CD spectra of PtNAC were recorded in the wavelength range 260–190 nm using a quartz cuvette with 0.1 cm path length. The data points were collected using step resolution 0.1 nm, time constant of 2 s, and scan speed of 100 nm/min, with a spectral band width of 2.0 nm. Three accumulations per sample were recorded. The software BestSel was used for predicting the percentage of α-helices and β-sheets[1].

Effect of pH and Temperature

The effect of pH on the secondary structure of PtNAC was studied by incubating NAC protein (0.5 mg/ml) in a buffer of different pH, ranging from 2 to 10 (20 mM glycine HCl, pH 2.0 and 3.0; 20 mM acetate buffer, pH 4.0 and 5.0; 20 mM phosphate buffer, pH 7.0 and 8.0; and 20 mM glycine NaOH buffer, pH 9.0 and 10). The effect of temperature on the secondary structure of PtNAC was studied by incubating the protein (0.5 mg ml^{-1}) for 2 min at different temperatures (20–80°C), after which changes in the structural conformation were recorded. For this, the protein was dialyzed overnight in 20 mM Tris HCl buffer (pH 7.5).

Effect of Denaturants on the Secondary Structure of NAC

A fixed volume of PtNAC (0.5 mg ml^{-1}) was incubated for 3 h with different concentrations of denaturant, guanidine hydrochloride (Gdn-HCl) (1–6 M), and urea (1–6 M), and changes in the structural conformation of PtNAC were recorded.

Conditions of *P. torridus* Culture and Preparation of Cell Lysate

P. torridus (DSM 9790) was purchased from DSMZ-German Collection of Microorganisms and Cell Cultures, GmbH Leibniz Institute, Germany. Archaea were grown in 500 ml of culture medium (Arora et al., 2014) at 55°C in a shaking incubator (100 rpm). The archaeal cells were harvested in the late log growth phase by centrifuging at 10,000 × *g* for 10 min at 4°C and washed with 10 mM phosphate-buffered saline (pH 7.4). The cells were resuspended in cell lysis buffer (50 mM Tris–HCl, 10 mM $MgCl_2$, 1 mM EDTA; pH 7.4) and lysed by intermittent sonication on ice for 10 min. The cell lysate was cleared by centrifuging at 14,000 × *g* for 20 min (4°C).

Isolation of Interacting Proteins of PtNAC by Pull Down Assay

His6-PtNAC was mixed with Co^{2+}-NTA-Agarose beads (the beads were washed and equilibrated with 1X PBS, pH 7.2) and incubated at 4°C for 1 h to allow it to bind with the beads. These beads were then transferred into a glass column and washed with five column volumes of the equilibration buffer, followed by washing with 1X PBS containing 50 mM imidazole, and the samples were collected to remove non-specific proteins. The beads were then washed extensively with 1X PBS to remove imidazole. Following this, 10 ml of *P. torridus* cell lysate was added to the column, which was incubated overnight at 4°C on a rocker, allowing the NAC interacting proteins to bind with the PtNAC protein associated with the beads. On the next day, the column was washed with 1X PBS buffer to remove non-interacting proteins, and the elution of the bound proteins was done in 1X PBS containing 200 mM imidazole. The samples were subjected to SDS-PAGE analysis. The interacting proteins were identified using liquid chromatography–mass spectrometry (LC–MS).

Sample Processing for LC–MS Analysis

The methods for sample preparation for LC–MS analysis have been described previously (Kumar et al., 2018). Briefly, the proteins in the eluate were incubated with 10% trichloroacetic acid overnight at 4°C. The resulting protein precipitate was washed with 2% sodium acetate in ethanol, air-dried, and resuspended in 200 μl of 8 M urea buffer (UB), loaded in a 3-kDa filter unit (Amicon-Millipore), and centrifuged at 14,000 × *g* for 15 min. Then, 100 μl of 0.05 M iodoacetamide prepared in UB was added, mixed at 600 rpm for 1 min, again kept for incubation for 20 min, followed by centrifugation at 14,000 × *g* for 10 min, and followed by twice washing with 100 μl of UB and centrifugation at 14,000 × *g* for 15 min. Then, two washings were done with 100 μl of 0.05 M ammonium bicarbonate (ABC), and centrifugation was done at 14,000 × *g* for 10 min. After this, 40 μl of ABC and trypsin (Promega V511A) solution (enzyme/protein ratio, 1:100) was added, mixed at 600 rpm for 1 min, and kept for 16–18 h of incubation in a water bath at 37°C. The digested peptides were eluted at 14,000 × *g* for 10 min, followed by recovering the remaining uneluted peptides with another 20–30 μl of ABC. The total eluted sample was acidified with 0.1% formic acid and concentrated by speed vac to attain a final volume of 10 μl. LC–MS/MS analysis was performed in a AB SCIEX Triple TOF 5600. The MASCOT and PARAGON search engines were used for peptide identification. Proteins were selected for further study based on a 5% false-discovery rate cutoff and a minimum of two peptides per protein.

RESULTS

Cloning, Expression, and Purification of NAC

The positive clones encoding PtNAC were confirmed by double-digestion method (**Figure 1A**). The protein was found to be overexpressed in *E. coli* BL21 (DE3) (**Figure 1B**). The protein was purified using Co^{2+}-NTA His6-affinity chromatography, followed by size exclusion chromatography on a HiPrepTM S-200 HR column (GE Healthcare). The molecular mass of the recombinant NAC protein after

[1]http://bestsel.elte.hu/

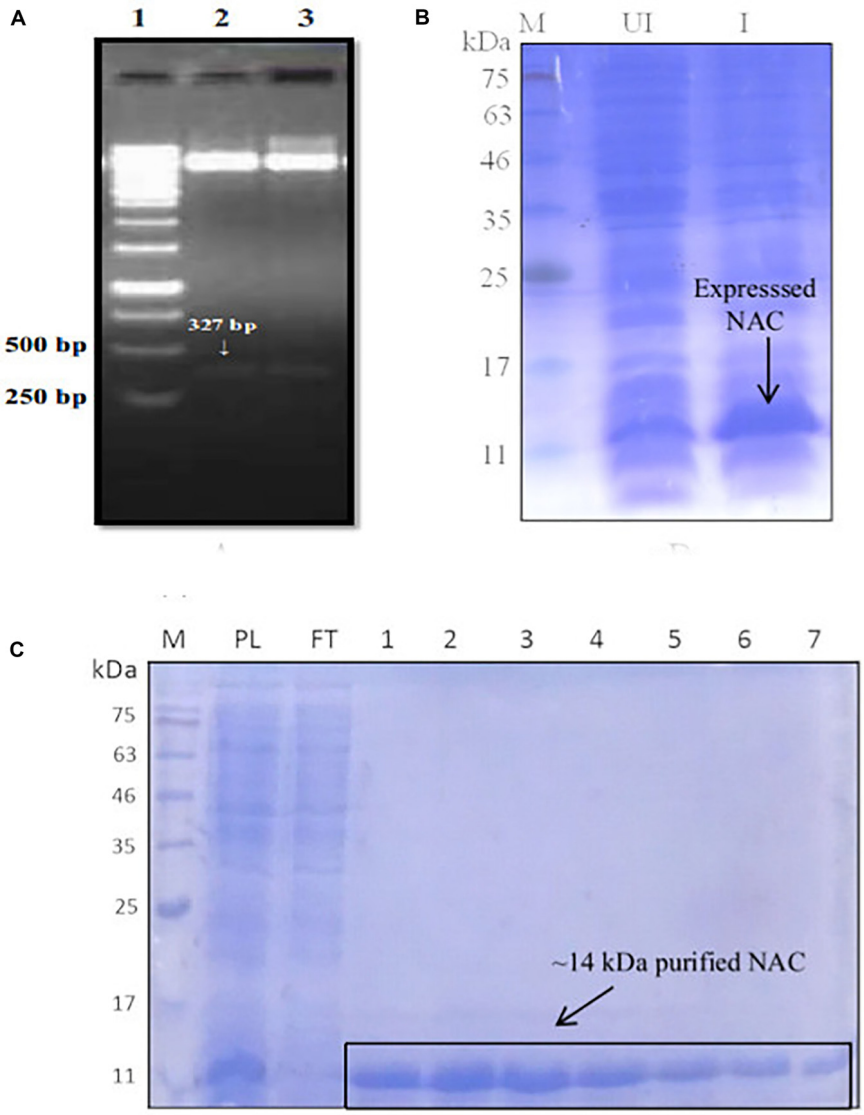

FIGURE 1 | (A) Confirmation of positive clones of PtNAC by double-digestion method. Lanes: 1—marker (Thermo Scientific O' GeneRuler 1-kb DNA ladder); 2, 3—double-digested pET28a(+) with the fallout. **(B)** Expression of PtNAC in *E. coli* BL21(DE3). Lanes: 1—SDS-PAGE marker, 2—uninduced: total cellular extract from *E. coli* BL21(DE3) before isopropyl β-D-1-thiogalactopyranoside (IPTG) induction, and 3—induced: total cellular extract after IPTG induction shown by arrow. **(C)** Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) analysis of recombinant NAC purified in pET28a(+) systems (samples were resolved on 15% polyacrylamide gel and stained with Coomassie Brilliant Blue R-250). Lanes: M—SDS protein marker, PL—preload, FT—flow through, 1–7 purified recombinant nascent polypeptide-associated complex eluents obtained from Co^{2+}-NTA metal affinity column using imidazole (200 mM). BlueRAY prestained protein marker was used as the SDS-PAGE marker.

electrophoresis on 15% SDS-PAGE was observed to be ∼14 kDa (**Figure 1C**). The MALDI-TOF analysis further confirmed that the recombinant, overexpressed protein was the NAC protein of *P. torridus*. The MALDI-TOF analysis revealed two peaks corresponding to the molecular weights of 14.5 and 29.0 kDa, suggesting that the PtNAC protein purified by affinity chromatography was a mixture of dimer and monomer population (**Supplementary Figure S1**). However, only a single peak was observed in the elution on the gel filtration chromatography, which corresponds to the dimer population when compared to the protein standards run on this column in our laboratory (**Supplementary Figure S2**).

Chaperone Activity

There was a slight reduction in the thermal aggregation of BCAII in the presence of an equimolar concentration of NAC (BCAII/NAC, 1:1) compared to the sample containing BCAII only, suggesting that the PtNAC protein might have a chaperone-like activity *in vitro* (**Figure 2**). The higher concentrations of NAC were accompanied by a greater aggregation (BCAII/NAC, 1:2 and 1:4). However, the effect of the chaperone-like

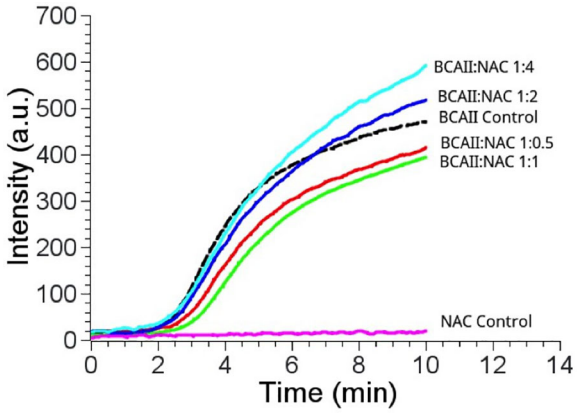

FIGURE 2 | Chaperone activity of PtNAC/BCAII control is substrate control (black); BCAII/NAC in different ratios; 1:0.5 (red), 1:1 (green), 1:2 (blue), and 1:4 (cyan). The best prevention of aggregation of BCAII in the presence of NAC was observed at equimolar BCAII/NAC concentration.

activity on cell viability under *in vivo* stress conditions remains to be studied.

Effect of Temperature and pH on PtNAC Structure

The CD spectrum showed the predominance of β-sheets in the secondary structure of PtNAC at pH 7.5 and 20°C (**Figure 3A**). The changes in the secondary structure of PtNAC were studied at different temperatures (20–80°C), and it was observed that there was no significant change in the secondary structure with the increase in temperature up to 80°C, indicating that PtNAC was a thermostable protein (**Figure 3B**). The secondary structure of PtNAC was stable over a broad range of pH between 2 and 10 (**Figure 3C**).

CD Spectra of the Native and Denatured PtNAC

Guanidine hydrochloride (Gdn-HCl), a strong chaotrophic agent, was used to determine the effect of denaturants on the secondary structure of PtNAC. BestSel analysis predicted the presence of 21.9% α-helices and 28.7% antiparallel β-sheet in the native protein (**Supplementary Figure S3**). However, in the presence of 6 M Gdn-HCl, the α-helices content decreased significantly to 12.2%, while the antiparallel β-sheet content showed a smaller decrease to 24.9%. A progressive shift in the spectral wavelength and a decrease in negative ellipticity were observed with an increase in the concentration of Gdn-HCl up to 6 M (**Figure 4**, the change in secondary structure at 222 nm is shown in the inset). However, when PtNAC was incubated with increasing concentrations of urea, the decrease in negative ellipticity and the shift in spectral wavelength were not very significant (**Figure 5**, the change in secondary structure at 222 nm is shown in the inset).

PtNAC-Binding Proteins

Protein interactions play an important role in cellular organization, and information about the functional partners of a protein can help in understanding its role in biological processes. Thus, an attempt was made to identify the interacting protein partners of PtNAC. PtNAC-binding proteins were isolated by pull down assay, separated on an SDS-PAGE (**Figure 6**) and identified by LC–MS. The LC–MS analysis identified 19 proteins as the binding partners of PtNAC (**Table 1**). These were elongation factor 1-alpha (Q6L202), protein secretion chaperonin CsaA (Q6L208), glutaredoxin (Q6L248)-related protein, thermosome subunit (Q6KZS2), 50S ribosomal protein L12 (Q6L1X7), thermosome subunit (Q6L132), hypothetical aldo-keto reductase (Q6L0H6), succinyl-CoA synthetase beta chain (Q6L0B4), DNA-binding protein (Q6L048), Rieske iron–sulfur protein (Q6KZY4), diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase (Q6KZV0), glutamate dehydrogenase (Q6KZF2), pyruvate ferredoxin oxidoreductase, alpha chain (Q6KZA7), translation initiation factor 5A (Q6L150), serine/threonine protein kinase (Q6L2G5), homoserine dehydrogenase (Q6KZ50), and uncharacterized proteins (Q6L1Y4, Q6L168, and Q6L0Y6).

DISCUSSION

The present study aimed to unravel the biophysical and the biochemical characteristics of PtNAC and discern its binding partners. The homologs of PtNAC have been reported in all archaeal species, except *Nanarchaeum equitans, Sulfolobus solfataricus* P2, *Sulfolobus islandicus* REY15A, and *Sulfolobus islandicus* M.14.25 (Rani et al., 2016). To date, the crystal structure has been reported for NAC from only one archaeal species *Methanothermobacter marburgensis*, where it was found to be a homodimer of alpha-subunits (Spreter et al., 2005). The results of the CD spectroscopy of PtNAC indicated that ß-sheet is the predominant form of secondary structure present in PtNAC, similar to that reported in the crystal structure of NAC of *M. marburgensis* (Spreter et al., 2005).

Earlier studies had proposed that the NAC protein has a chaperone-like activity as it binds to ribosomes and interacts with nascent polypeptides (Bukau et al., 2000; Hartl and Hayer-Hartl, 2002; Wegrzyn and Deuerling, 2005). Since PtNAC mildly prevented the aggregation of heat-denatured BCAII, it might also have a chaperone-like activity. However, the effect of a chaperone-like activity on cell viability under *in vivo* stress conditions remains to be studied. The CD spectroscopic studies also revealed that the secondary structure of PtNAC was stable over a wide range of temperature and pH, which fortifies its role in aiding the adaptability of this thermoacidophilic archaea in extreme growth conditions, possibly by contributing to the prevention of aggregation and helping in the proper folding of other cellular proteins. The denaturants Gdn-HCl and urea showed different effects on the secondary structure conformations of PtNAC. This might be attributed to the difference in the nature of both denaturants. Gdn-HCl is a charged denaturant, while

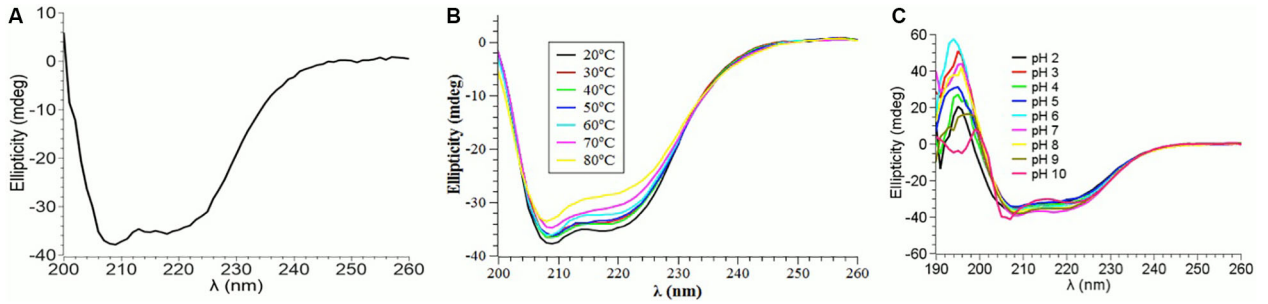

FIGURE 3 | Changes in the secondary structure of PtNAC investigated using circular dichroism spectroscopy: **(A)** secondary structure estimation at pH 7.5 and 20°C, **(B)** effect of different temperatures ranging from 20 to 80°C, and **(C)** effect of different pH values ranging from 2 to 10.

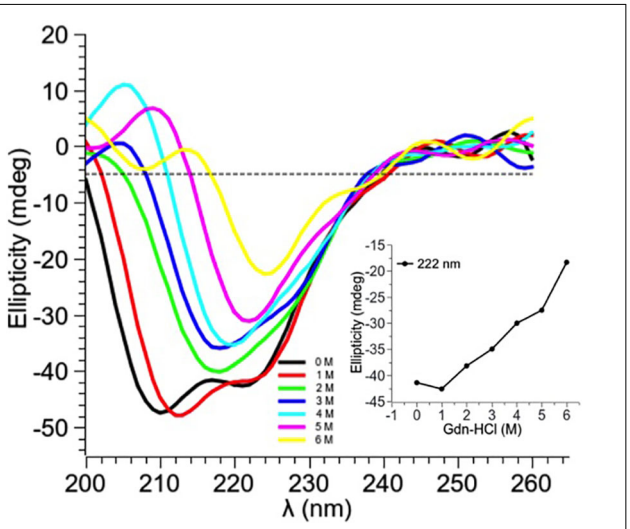

FIGURE 4 | Effect of increasing concentrations of Gdn-HCl (1–6 M) on PtNAC: Inset: ellipticity at 222 nm versus Gdn-HCl concentration. On increasing Gdn-HCl concentration from 1 to 2 M, a gradual change in the secondary structure was observed. A decrease in the negative ellipticity and a right shift in the spectral wavelength were observed with the progressive increase in Gdn-HCl concentration up to 6 M.

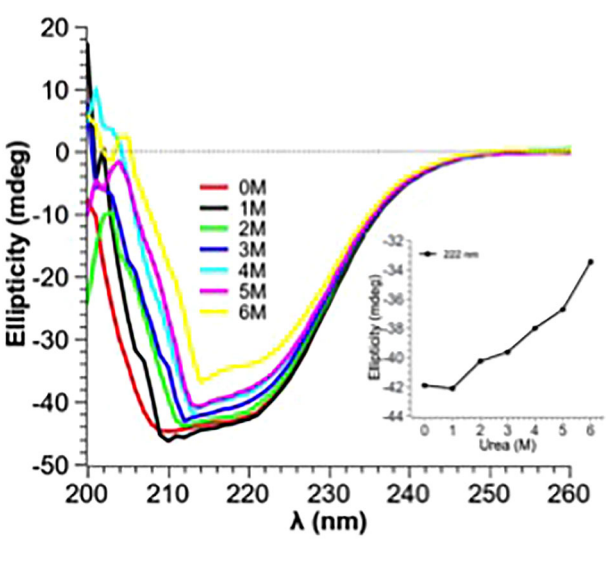

FIGURE 5 | Effect of increasing concentrations of urea (1–6 M) on PtNAC: Inset: ellipticity at 222 nm versus urea concentration. A gradual change in the secondary structure was observed when the concentration of urea was increased from 1 to 6 M.

urea is neutral by nature. Gdn-HCl, due to its charged nature, hides the favorable electrostatic interactions that might stabilize the native state of protein, affecting the protein structure, while urea specifically binds to the amide units and therefore denatures proteins by minimizing the hydrophobic effect.

To understand the participation of NAC in metabolic processes and pathways, the present study was designed to isolate the binding proteins of *P. torridus* NAC by affinity separation followed by LC–MS analysis. The results of the pull down assay indicated that the NAC of *P. torridus* interacted with a diverse set of proteins, suggesting that it might be a multifaceted protein. The Kyoto Encyclopedia of Genes and Genomes-based metabolic pathway analysis of the NAC-binding proteins revealed that many interacting protein partners of NAC were multi-functional. A functional categorization of these proteins revealed that two of them were associated with amino acid metabolism (Q6KZ50 and Q6KZF2), two with carbohydrate metabolism (Q6L0B4 and Q6KZA7), three with energy metabolism (Q6L0B4, Q6KZA7, and Q6KZF2), four with global and overview maps (Q6L0B4, Q6KZA7, Q6KZF2, and Q6KZ50), one with metabolism of other amino acids (Q6KZF2), and one with translation (Q6L1X7) (**Table 2**). Interestingly, four of the interacting protein partners of PtNAC were archaeal chaperones. Two of them (Q6L132 and Q6KZS2) were subunits of thermosome that belong to the family of Group II archaeal chaperonins, one was CsaA (Q6L208) and the other was a glutaredoxin-related protein (Q6L248), a member of the thioredoxin superfamily of archaeal chaperones (**Table 2**).

The NAC-binding partner Q6L202 was identified as elongation factor 1-alpha. Elongation factor 1-alpha promotes the GTP-dependent binding of aminoacyl-tRNA to the A-site of ribosomes during the biosynthesis of proteins. The NAC-binding protein Q6L208 was recognized as CsaA. CsaA is one of the

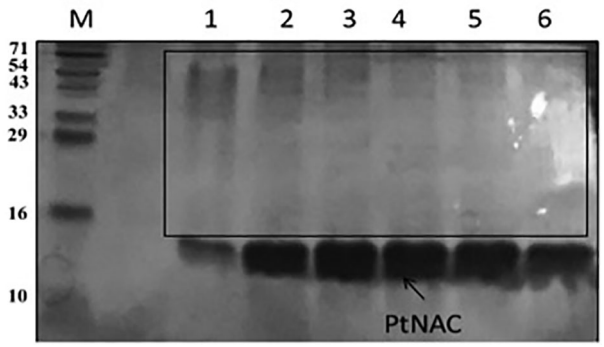

FIGURE 6 | Sodium dodecyl sulfate-polyacrylamide gel electrophoresis protein elution profile of PtNAC with 200 mM imidazole after binding with *P. torridus* cell lysate: lane M—three-color prestained protein ladder from PUReGeNe; lanes 1–6 show the eluents with interacting proteins (boxed).

few chaperones which are present in bacteria and archaea but is absent in eukaryotes. CsaA prevents the aggregation of unfolded proteins and helps in the translocation of proteins across the cytoplasmic membrane (Sharma et al., 2018). The binding partner Q6L248 was identified as a glutaredoxin-related protein. Glutaredoxin is a member of the thioredoxin superfamily of proteins which is crucial for maintaining a reduced intracellular redox state. It also helps in protein folding and has been shown to have a chaperone-like activity (Berndt et al., 2008; Rani et al., 2016).

The NAC-binding proteins Q6L132 and Q6KZS2 were identified as thermosome subunits. Thermosomes belong to the family of Group II chaperonins which are ubiquitously present in archaea and impart extreme thermal stability (Phipps et al., 1991; Bigotti and Clarke, 2008). The Group II chaperonins comprise up to 40% of the total cellular protein and are abundantly produced by the cells exposed to heat shock (Trent et al., 1991). The binding partner Q6L1X7 was identified as 50S ribosomal protein L12. The 50S ribosomal protein L12 binds to the 23S rRNA and is an important constituent of the secondary structure of the ribosome. The NAC-binding protein Q6L0H6 was identified as a hypothetical aldo-keto reductase. Aldo-keto reductases are a superfamily of enzymes which are involved in the reduction of aldehydes and ketones. The binding protein Q6L0B4 was identified as the beta chain of succinyl-CoA synthase. It is a mitochondrial matrix enzyme composed of two subunits, alpha and beta. Succinyl-CoA synthase is an enzyme in the Krebs cycle that converts succinyl-CoA to succinate and free coenzyme A and converts ADP or GDP to ATP or GTP, respectively (Johnson et al., 1998).

The PtNAC-binding partner Q6KZY4 was identified as Rieske iron–sulfur protein. Proteins containing Rieske-type [2Fe-2S] clusters are associated with essential functions in all the three domains of life. The Rieske proteins occur as subunits in the cytochrome bc1 and cytochrome b6f complexes of prokaryotes and eukaryotes or form components of archaeal electron transport systems (Schmidt and Shaw, 2005). The families encoding for Rieske iron–sulfur proteins are more common in bacteria and archaea than in eukaryotes. Recent studies suggest that Rieske proteins are functionally versatile like other redox proteins, and the use of multiple Rieske proteins in electron transfer reactions helps in microbial adaptation to changing environmental conditions (Schneider and Schmidt, 2005). The binding protein Q6KZF2 was identified as glutamate dehydrogenase. Glutamate dehydrogenase is one of the most widely studied enzymes for understanding the thermostability of hyperthermophiles (Vetriani et al., 1998;

TABLE 1 | Details of PtNAC interacting proteins identified using the pull down assay.

S. No.	Gene accession number	Protein accession number	Protein name
1	PTO0415	Q6L202	Elongation factor 1-alpha
2	PTO0409	Q6L208	Protein secretion chaperonin CsaA
3	PTO0369	Q6L248	Glutaredoxin related protein
4	PTO1195	Q6KZS2	Thermosome subunit
5	PTO0433	Q6L1Y4	Uncharacterized protein
6	PTO0440	Q6L1X7	50S ribosomal protein L12
7	PTO0699	Q6L168	Uncharacterized protein
8	PTO0735	Q6L132	Thermosome subunit
9	PTO0941	Q6L0H6	Hypothetical aldo-keto reductase
10	PTO1003	Q6L0B4	Succinyl-CoA synthetase beta chain
11	PTO1069	Q6L048	DNA-binding protein
12	PTO1133	Q6KZY4	Rieske iron–sulfur
13	PTO1167	Q6KZV0	CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase
14	PTO1315	Q6KZF2	Glutamate dehydrogenase
15	PTO1360	Q6KZA7	Pyruvate ferredoxin oxidoreductase, alpha chain
16	PTO0717	Q6L150	Translation initiation factor 5A
17	PTO0781	Q6L0Y6	Uncharacterized protein
18	PTO0252	Q6L2G5	Serine/threonine protein kinase
19	PTO1417	Q6KZ50	Homoserine dehydrogenase OS

TABLE 2 | Annotation of Kyoto Encyclopedia of Genes and Genomes pathways of the interacting protein partners of *P. torridus* NAC.

Pathway	KEGG ID	Protein accession number	Pathway maps
Arginine biosynthesis	pto00220	Q6KZF2	Amino acid metabolism
Lysine biosynthesis	pto00300	Q6KZ50	"
Cysteine and methionine metabolism	pto00270	Q6KZ50	"
Alanine aspartate and glutamate metabolism	pto00250	Q6KZF2	"
Glycine serine and threonine metabolism	pto00260	Q6KZ50	"
Biosynthesis of secondary metabolites	pto01110	Q6L0B4, Q6KZA7, Q6KZ50	Biosynthesis of other secondary metabolites
Citrate cycle	pto00020	Q6L0B4, Q6KZA7	Carbohydrate metabolism
Propanoate metabolism	pto00640	Q6L0B4	"
Glycolysis/gluconeogenesis	pto00010	Q6KZA7	"
Butanoate metabolism	pto00650	Q6KZA7	"
C5-branched dibasic acid metabolism	pto00660	Q6L0B4	"
Pyruvate metabolism	pto00620	Q6KZA7	"
Carbon fixation pathways in prokaryotes	pto00720	Q6L0B4, Q6KZA7	Energy metabolism
Nitrogen metabolism	pto00910	Q6KZF2	"
Microbial metabolism in diverse environments	pto01120	Q6L0B4, Q6KZF2, Q6KZA7, Q6KZ50	Global and overview maps
Carbon metabolism	pto01200	Q6L0B4, Q6KZF2, Q6KZA7	"
Biosynthesis of antibiotics	pto01130	Q6L0B4, Q6KZA7, Q6KZ50	"
Biosynthesis of amino acids	pto01230	Q6KZ50	"
D-Glutamine and D-glutamate metabolism	pto00471	Q6KZF2	Metabolism of other amino acids
Ribosome	pto03010	Q6L1X7	Translation

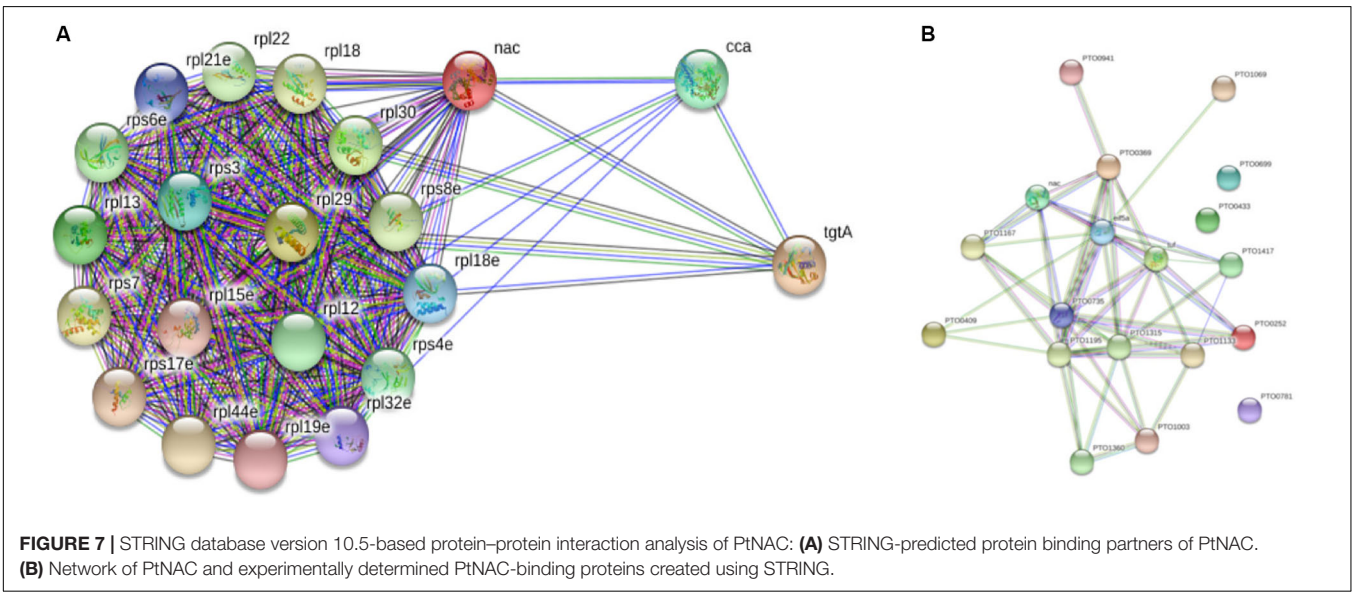

FIGURE 7 | STRING database version 10.5-based protein–protein interaction analysis of PtNAC: **(A)** STRING-predicted protein binding partners of PtNAC. **(B)** Network of PtNAC and experimentally determined PtNAC-binding proteins created using STRING.

Vieille and Zeikus, 2001). The binding protein Q6KZA7 was identified as pyruvate ferredoxin oxidoreductase alpha chain. In enzymology, a pyruvate ferredoxin oxidoreductase/pyruvate synthase is an enzyme that catalyzes the interconversion of pyruvate and acetyl-CoA.

The PtNAC-binding protein Q6KZ50 was identified as reversed homoserine dehydrogenase. Homoserine dehydrogenase is a key enzyme in the aspartate pathway involved in the NAD (P)-dependent reduction of aspartate beta-semialdehyde into homoserine. Homoserine is an intermediate in the biosynthesis of three amino acids—threonine, methionine, and isoleucine—in plants and microorganisms. The binding protein Q6L2G5 was identified as a serine/threonine protein kinase. Serine/threonine protein kinases are involved in protein phosphorylation, one of the most important post-translational modifications that regulate almost every cellular process, including signal transduction. The binding protein Q6L150 was identified as translation initiation factor 5A. The translation initiation factor 5A promotes the formation of the first peptide bond at the onset of protein synthesis. The binding protein Q6KZV0 was identified as CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase. The CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase participates in the

metabolism of glycerophospholipids/phosphoglycerides. Glycerophospholipids are present abundantly in the cell membranes where they serve as an anchor for proteins in cell membranes and also participate in cell signaling.

When String database version 10.5 (Szklarczyk et al., 2017) was used to discern the interacting protein partners of PtNAC, it was observed that, of the 19 interacting protein partners predicted by the STRING database, only one protein, 50S ribosomal protein L12 (rpl12), was present in the experimental interactome (**Figure 7A**). This is, however, not surprising because to date no experimental interactome studies have been conducted for the NAC proteins of archaea. Also, because NAC proteins are absent in prokaryotes, the NAC interaction networks in STRING might have been built on the basis of information available about eukaryotic NAC proteins. The STRING database suggests the maximum interacting partners of PtNAC to be ribosomal proteins, as predicted by other studies as well (Pech et al., 2010); however, only a single ribosomal protein was identified in our study. It is possible that the inclusion of EDTA in the binding buffers could have possibly led to the dissociation of ribosomes and affected their binding to PtNAC. Since the binding of most proteins *in vivo* is transient, the interacting partners identified in the study are also expected to be sensitive to conditions such as the buffers used for facilitating binding and subsequent separation from the column. More elaborate studies (e.g., similar pull down studies at different pH and buffer concentrations) may be required to make the predictions more robust. However, the final validation of any such predicted interaction lies in seeking the effect of such interactions on the physiology of the organism, such as growth and reproduction. However, this preliminary study does provide an opportunity to investigate the role of NAC protein in pathways or cellular functions hitherto unassigned to this protein. We therefore made an attempt to find if the interacting partners predicted in this study can be predicted to be part of a common pathway or have been shown to interact with each other in any previous studies.

Next, a network analysis of PtNAC and its interacting protein partners experimentally identified in our study was performed using String database version 10.5 (**Figure 7B**). Interestingly, the STRING analysis revealed a strong interaction network among the experimentally identified PtNAC-binding proteins. Sixteen of the 19 experimentally identified proteins showed mutual connections with each other on the PPI network map. The three PtNAC interacting proteins which did not integrate in the interaction network were uncharacterized proteins (protein accession/gene accession—Q6L1Y4/PTO0433, Q6L168/PTO0699, and Q6L0Y6/PTO0781). Thus, the results of the STRING analysis corroborated our experimental findings and suggested novel functional possibilities for the NAC protein.

CONCLUSION

The present work is the first study of the interactome analysis of NAC of any archaeal species and, to the best of our knowledge, the first report on the biophysical and the biochemical characterization of PtNAC. Although this is a preliminary study, our results do provide an opportunity to investigate the role of NAC in pathways or cellular functions hitherto unassigned to this protein.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

MG conceptualized the study. NS, AS, SK, RR, and NS worked on the methodology. AG, MK, and NS were responsible for the software. NS, AS, SK, RR, and NS validated the study. NS contributed to the formal analysis. MG helped with the resources. NS and AS prepared the original draft. NS, AS, and MG reviewed and edited the manuscript.

FUNDING

This research was funded by the Indian Council of Medical Research (ICMR) (BIC/12(14)/2015).

ACKNOWLEDGMENTS

AS is thankful to the Department of Biotechnology, Government of India, for the research grant (BT/Bio-CARe/01/9935/2013-14, dated 30.09.2014) received under BioCARe Scheme that supported this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00915/full#supplementary-material

REFERENCES

Arora, J., Goswami, K., and Saha, S. (2014). Characterization of the replication initiator Orc1/Cdc6 from the archaeon *Picrophilus torridus*. *J. Bacteriol.* 196, 276–286. doi: 10.1128/JB.01020-13

Arsenovic, P. T., Maldonado, A. T., Colleluori, V. D., and Bloss, T. A. (2012). Depletion of the C. *elegans* NAC engages the unfolded protein response, resulting in increased chaperone expression and apoptosis. *PLoS One* 7:e44038. doi: 10.1371/journal.pone.0044038

Berndt, C., Lillig, C. H., and Holmgren, A. (2008). Thioredoxins and glutaredoxins as facilitators of protein folding. *Biochim. Biophys. Acta* 1783, 641–650. doi: 10.1016/j.bbamcr.2008.02.003

Bigotti, M. G., and Clarke, A. R. (2008). Chaperonins: the hunt for the Group II mechanism. *Arch. Biochem. Biophys.* 474, 331–339. doi: 10.1016/j.abb.2008.03.015

Bukau, B., Deuerling, E., Pfund, C., and Craig, E. A. (2000). Getting newly synthesized proteins into shape. *Cell* 101, 119–122.

Creagh, E. M., Brumatti, G., Sheridan, C., Duriez, P. J., Taylor, R. C., Cullen, S. P., et al. (2009). Bicaudal is a conserved substrate for Drosophila and mammalian caspases and is essential for cell survival. *PLoS One* 4:e5055. doi: 10.1371/journal.pone.0005055

del Alamo, M., Hogan, D. J., Pechmann, S., Albanese, V., Brown, P. O., and Frydman, J. (2011). Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with ribosome-nascent chain complexes. *PLoS Biol.* 9:e1001100. doi: 10.1371/journal.pbio.1001100

Gamerdinger, M., Hanebuth, M. A., Frickey, T., and Deuerling, E. (2015). The principle of antagonism ensures protein targeting specificity at the endoplasmic reticulum. *Science* 348, 201–207. doi: 10.1126/science.aaa5335

Hartl, F. U., and Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* 95, 1852–1858.

Hoffmann, A., Bukau, B., and Kramer, G. (2010). Structure and function of the molecular chaperone Trigger Factor. *Biochim. Biophys. Acta* 1803, 650–661. doi: 10.1016/j.bbamcr.2010.01.017

Johnson, J. D., Mehus, J. G., Tews, K., Milavetz, B. L., and Lambeth, D. O. (1998). Genetic evidence for the expression of ATP- and GTP-specific succinyl-CoA synthetases in multicellular eucaryotes. *J. Biol. Chem.* 273, 27580–27586.

Kaiser, C. M., Chang, H. C., Agashe, V. R., Lakshmipathy, S. K., Etchells, S. A., Hayer-Hart, M., et al. (2006). Real-time observation of trigger factor function on translating ribosomes. *Nature* 444, 455–460.

Kogan, G. L., and Gvozdev, V. A. (2014). Multifunctional protein complex NAC (nascent polypeptide associated complex). *Mol. Biol.* 48, 223–231.

Koplin, A., Preissler, S., Ilina, Y., Koch, M., Scior, A., Erhardt, M., et al. (2010). Dual function for chaperones SSB-RAC and the NAC nascent polypeptide-associated complex on ribosomes. *J. Cell Biol.* 189, 57–68. doi: 10.1083/jcb.200910074

Kumar, R., Singh, N., Abdin, M. Z., Patel, A. H., and Medigeshi, G. R. (2018). Dengue virus capsid interacts with DDX3X-a potential mechanism for suppression of antiviral functions in dengue infection. *Front. Cell Infect. Microbiol.* 7:542. doi: 10.3389/fcimb.2017.00542

Pech, M., Spreter, T., Beckmann, R., and Beatrix, B. (2010). Dual binding mode of the nascent polypeptide-associated complex reveals a novel universal adapter site on the ribosome. *J. Biol. Chem.* 285, 19679–19687. doi: 10.1074/jbc.M109.092536

Phipps, B. M., Hoffmann, A., Stetter, K. O., and Baumeister, W. A. (1991). Novel ATPase complex selectively accumulated upon heat shock is a major cellular component of thermophilic archaebacteria. *EMBO J.* 10, 1711–1722.

Powers, T., and Walter, P. (1996). The nascent polypeptide-associated complex modulates interactions between the signal recognition particle and the ribosome. *Curr. Biol.* 6, 331–338.

Rajaraman, K., Raman, B., and Rao, C. M. (1996). Molten-globule state of carbonic anhydrase binds to the chaperone-like alpha-crystallin. *J. Biol. Chem.* 271, 27595–27600.

Rani, S., Srivastava, A., Kumar, M., Goel, M., and Lund, P. (2016). CrAgDb-a database of annotated chaperone repertoire in archaeal genomes. *FEMS Microbiol. Lett.* 364:fnw030. doi: 10.1093/femsle/fnx030

Raue, U., Oellerer, S., and Rospert, S. (2007). Association of protein biogenesis factors at the yeast ribosomal tunnel exit is affected by the translational status and nascent polypeptide sequence. *J. Biol. Chem.* 282, 7809–7816.

Rospert, S., Dubaquié, Y., and Gautschi, M. (2002). Nascent-polypeptide-associated complex. *Cell Mol. Life Sci.* 59, 1632–1639.

Schleper, C., Puehler, G., Holz, I., Gambacorta, A., Janekovic, D., Santarius, U., et al. (1995). Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. *J. Bacteriol.* 177, 7050–7059.

Schmidt, C. L., and Shaw, L. A. (2005). A comprehensive phylogenetic analysis of Rieske and Rieske-type iron-sulfur proteins. *J. Bioenerg. Biomembr.* 33, 9–26.

Schneider, D., and Schmidt, C. L. (2005). Multiple Rieske proteins in prokaryotes: where and why? *Biochim. Biophys. Acta* 1710, 1–12.

Sharma, A., Rani, S., and Goel, M. (2018). Navigating the structure-function-evolutionary relationship of CsaA chaperone in archaea. *Crit. Rev. Microbiol.* 44, 274–289. doi: 10.1080/1040841X.2017.1357535

Spreter, T., Pech, M., and Beatrix, B. (2005). The crystal structure of archaeal nascent polypeptide-associated complex (NAC) reveals a unique fold and the presence of a ubiquitin-associated domain. *J. Biol. Chem.* 280, 15849–15854.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Tomar, R., Garg, D., Mishra, R., Thakur, A., and Kundu, B. (2013). N-terminal domain of *Pyrococcus furiosus* L-asparaginase functions as a non-specific, stable, molecular chaperone. *FEBS J.* 280, 2688–2699. doi: 10.1111/febs.12271

Trent, J. D., Nimmesgern, E., Wall, J. S., Hartl, F. U., and Horwich, A. L. (1991). A molecular chaperone from a thermophilic archaebacterium is related to the eukaryotic protein t-complex polypeptide-1. *Nature* 354, 490–493.

Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S., Stillman, T. J., Britton, K. L., et al. (1998). Protein thermostability above 100 degrees C: a key role for ionic interactions. *Proc. Natl. Acad. Sci. U.S.A.* 95, 12300–12305.

Vieille, C., and Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 65, 1–43.

Wegrzyn, R. D., and Deuerling, E. (2005). Molecular guardians for newborn proteins: ribosome-associated chaperones and their role in protein folding. *Cell Mol. Life Sci.* 62, 2727–2738.

Wickner, W. (1995). The nascent-polypeptide-associated complex: having a "NAC" for fidelity in translocation. *Proc. Natl. Acad. Sci. U.S.A.* 92, 9433–9434.

Wiedmann, B., Sakai, H., Davis, T. A., and Wiedmann, M. (1994). A protein complex required for signal-sequence-specific sorting and translocation. *Nature* 11, 434–440.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Emerging Role of HSP70 in Human Diseases

Anjali Garg, Bandana Kumari, and Manish Kumar

Abstract HSP70 are prominent stress proteins, which also act like molecular chaperones. The synthesis of HSP70 increases when the cell is exposed to any form of stress physical, biological or chemical. Under stress conditions, HSP70 recognize and bind to the unstable protein substrates and protect them from denaturation and aggregation. Besides, HSP70 are also essential during normal growth where they assist in folding of nascent proteins, degradation of misfolded and truncated proteins and, in subcellular localizations of proteins and vesicles. Since HSP70 are involved in a plethora of cellular activities, their role been implicated with several pathological diseases primarily related to apoptosis, carcinogenesis, amyloidogenesis. Here, we summarize the current knowledge on the HSP70 and their relevance in diseases such as cancer, diabetes, seizures and many more. Further, the relevance of HSP70 to serve as biomarkers and/or therapeutics in human diseases is also discussed.

Keywords Chaperone · Heat shock proteins · Human disease · Protein aggregation · Protein refolding · Stress

Abbreviations

AFLD	Alcoholic fatty liver diseases
AIF	Apoptosis-inducing factor
ATP	Adenosine triphosphate
DISC	Death inducing signaling complex
HSP	Heat shock protein
iNOS	Inducible nitric oxide synthase
MMP	Matrix metalloproteinase

Author contributed equally Anjali Garg and Bandana Kumari

A. Garg · B. Kumari · M. Kumar (✉)
Department of Biophysics, University of Delhi South Campus, New Delhi, India
e-mail: manish@south.du.ac.in

A. A. A. Asea, P. Kaur (eds.), *HSP70 in Human Diseases and Disorders*, Heat Shock Proteins 14, https://doi.org/10.1007/978-3-319-89551-2_15

NAFLD	Nonalcoholic fatty liver diseases
NBD	Nucleotide binding domain
NEF	Nucleotide exchange factor
SBD	Substrate binding domain
TLR	Toll-like receptor

Introduction

Prokaryotes and eukaryotes are exposed to various environmental stresses. To counteract their effects, these have evolved a wide array of molecular and physiological processes. Stress proteins, also named as heat shock proteins, are one of the responses against the deleterious effects of many abiotic and biotic stresses including extreme temperatures, radiations, heavy metals, drought, hypoxia, ischemia and assaults of bacterial, viral and parasitic origin (Whitley *et al.* 1999). Heat shock proteins (HSP) are a class of evolutionarily conserved, functionally related cellular proteins which primarily act as chaperons (Verghese *et al.* 2012; Tóth *et al.* 2015). HSP are ubiquitous in nature and present in cytoplasm under normal conditions, but they are transferred to the nucleus and their expression is increased when cells are exposed to high temperature or shock (Xu *et al.* 2012). In 1962, Ferrucio Ritossa serendipitously discovered the response to heat shock in the form of heat-induced chromosomal puffing in salivary gland chromosomes of *Drosophila buscii* (Ritossa 1962). Later, Tissieres *et al.* (1974) observed that exposure to heat shock led to increased synthesis of a new kind of proteins in different tissues of *Drosophila melanogaster,* which were highly similar. However, it was observed that the concentrations of other proteins declined during heat shock (Tissieres *et al.* 1974). Based on initial studies in different organisms, HSP were considered to be upregulated in response to heat only, and were therefore named so. Later, it was found that in addition to temperature, several other stress factors are also responsible for higher expression of HSP (Kalmar and Greensmith 2009; Tóth *et al.* 2015). The most important function of HSP is to protect cells from stress by maintaining homeostasis and by assisting the folding of denatured proteins under stressed conditions (Hartl and Hayer-Hartl 2002). During heat-treatment, expression of cellular proteins is highly suppressed, while the expression of HSP mRNA is highly increased. HSP are essential to prevent the conformational changes in other proteins, prevent aggregation of misfolded proteins, refolding of misfolded proteins, support proteasomal removal of peptides that cannot be refolded and membrane protection. In addition, they are important for growth and development and have anti-apoptotic functions.

Types of Heat Shock Proteins

Initially, HSP were classified on the basis of their molecular weight into following groups:

(a) Small heat shock protein (sHSP) family
(b) HSP40 (J-proteins)
(c) Chaperonin (HSP60/GroEL) family
(d) 70-kDa heat shock protein (HSP70/DnaK) family
(e) HSP90 family
(f) HSP100/ClpB family.

It is pertinent to mention here that a few reports have also included ubiquitin (8.5 kDa) as one of the HSP class in eukaryotes (Vierling 1997). There is a distinct pattern of ATP usage in HSP. For example, high molecular weight HSP (hHSP; 27–110 kDa), i.e., HSP60, HSP70, HSP90 and HSP100 are ATP dependent while smaller HSP (sHSP; 15–42 kDa) are ATP independent. Expression of hHSP at euthermic or stress temperatures show distinct set of functions such as protein folding and translocation, cytoprotection, regulation of nuclear hormone receptors as well as regulation of apoptosis, whereas sHSP are mostly tissue specific, and play an important role as chaperone for protein folding as well as strong anti-apoptotic effectors. In 2009, Kampinga *et al.* proposed a new classification system for human HSP families and categorized them into (Kampinga *et al.* 2009):

(a) Small heat shock protein (HSPB)
(b) DNAJ (HSP40)
(c) HSPA (HSP70)
(d) HSPC (HSP90)
(e) HSPH (HSP110)
(f) Chaperonin family HSPD/E (HSP60/HSP10)
(g) CCT (TRiC)

In addition to the above HSP classes, human exclusively contain HSP33 (De Maio 1999). The detailed comparison between the different families of HSP is presented in the Table 1.

In the present chapter, we would be focusing primarily on the HSP of the HSP70 family, which are well characterized, ubiquitous and highly conserved ATP-dependent chaperones. In comparison to other HSP, HSP70 are the most prominent response to the heat stress, toxic chemicals and heavy metals. In stressed cells, HSP70 are mostly localized in the nucleus and nucleolus. Under stress, HSP70 are over-expressed, refolds denatured proteins and induces tolerance. In addition to helping the cell to survive in stress, HSP70 have several functions in unstressed conditions also, e.g., folding of nascent peptides, intracellular protein transport and apoptosis. It has also been shown that expression level profile of HSP70 vary for healthy and diseased conditions (Radons 2016).

Table 1 The heat shock protein families (Kaminga *et al.* 2009)

Characters	Family						
	Small heat shock protein (HSPB)	DNAJ (HSP40)	HSPA (HSP70)	HSPC (HSP90)	HSPH (HSP110)	HSPD/E (HSP60/HSP10)	CCT (TRiC)
Number of member in human	11 (mammals)	~50	13	5	4	–	–
Characteristics	HSPB have conserved α-crystalline C-terminal domain of 100 amino acids (de Jong *et al.* 1998)	HSP40 contains a conserved N-terminal J-domain that stimulates ATPase activity (Qiu *et al.* 2006)	HSPA contains N-terminal regulatory ATPase domain and C-terminal substrate binding domain (Gragerov *et al.* 1994; Zhu *et al.* 1996)	HSPC contain three regions: the N-terminal region, central region and C-terminal (Csermely *et al.* 1998)	The HSPH are homologous to HSPA. HSPH have longer linker region between N- and C-terminal domains (Kampinga *et al.* 2009)	HSPD has three domains i.e. apical domain, equatorial domain and intermediate domain these play important role in binding of substrate and co-chaperone, ATP binding and act as hinge prompting conformational change respectively (Fink 1999) and it is heat inducible protein	CCT is not upregulated during heat shock

Functions		Folding, refolding, translocation, responsible for stimulation of HSPA ATPase activity	Folding of newly synthesized proteins, protein transport across intracellular membrane, DNA repair	Suppress aggregation of unfolded proteins, disaggregates loose protein aggregates, and enhances refolding of partially denatured proteins and cellular signaling	Folding, nucleotides exchange factor and removal of ADP after ATP hydrolysis	Folding of nascent and misfolded proteins in ATP-dependent manner	Folding of newly synthesized cytosolic proteins, preventing protein aggregation
	HSPB proteins bind to unfolded, partially denatured and damaged proteins, and inhibiting their irreversible aggregation, retaining them in a refolding competent state, it provides protection aganist both apoptosis and oxidative stress						
Subcellular location	Cytoplasm, cytoskeleton and nucleus	Cytosol, nucleus, endoplasmic reticulum, mitochondria, endosomes and ribosomes	Mitochondria (HSPA9), endoplasmic reticulum (HSPA5)	Cytosol, endoplasmic reticulum and mitochondria	Cytosol and endoplasmic reticulum	Mitochondria and chloroplast	Cytosol
Comments	Present in both prokaryotes and eukaryotes that expressed under stress conditions	–	–	–	–	Represented by GroEL in prokaryotes and HSP60 in mitochondria	Present in eukaryotes

HSP70: Structure and Mechanism

There are three distinct regions of HSP70: (a) Conserved N-terminal ATPase domain or nucleotide binding domain (NBD) of ~40 kDa; composed of four subdomains (IA, IB, IIA, IIB) surrounding the ATP-binding pocket, (b) Substrate binding domain (SBD) of ~18-kDa and (c) Variable C-terminal of ~10-kDa. Each of the three domains has different functions. The SBD binds to the substrate proteins and their association is regulated by NBD. The variable C-terminal acts as a "lid" of HSP70 and helps to hold the substrates at SBD (Zhu *et al.* 1996). Human HSP70 (HSPA) is a dimer of N-terminal ATPase domain (45 kDa) (Flaherty *et al.* 1990) and a C-terminal peptide binding domain (25 kDa) (Zhu *et al.* 1996). A small linker domain separates the N-terminal domain and the C-terminal domain. In order to help in protein folding, HSPA repeatedly binds and release the unfolded protein. The binding occurs at hydrophobic regions, since they are exposed in unfolded proteins. This cyclic process is dependent on the ATPase activity of HSP70, which is assisted by co-chaperones J-proteins and nucleotide exchange factor (Miot *et al.* 2011). J-proteins induce the hydrolysis of ATP that is required for binding of solvent-exposed hydrophobic amino acids of substrate proteins whereas nucleotide exchange factor is associated with ATP–ADP exchange which release ADP from HSP70 and ultimately the substrate.

Functions of HSP70

The HSP70 family of proteins is housekeeping proteins and is highly conserved across all living domains. The major responsibilities of HSP70 are folding of nascent proteins in normal cells and refolding of denatured proteins under shock condition. Apart from this, HSP70 are also involved in multiple biological functions including import and translocation of proteins and vesicles into organelles across membranes, growth, apoptosis, proteolytic degradation of unstable proteins by targeting the proteins to lysosomes or proteasomes and the degradation of unwanted proteins. The functions of HSP70 family members highly depend on their cellular localization, and on the basis of their localization they are broadly classified in two types, intracellular and extracellular HSP70. The intracellular residing HSP70 protect cells against lethal damage induced by stress, and support folding and transport of newly synthesized polypeptides and aberrant proteins as well as assembly of multi-protein complexes. Further, extracellular HSP70 are considered as molecules with immunomodulatory functions, which act either as cross-presenters of immunogenic peptides via MHC antigen or in a peptide-free version as chaperokines or stimulators of innate immune responses. The major cellular functions of HSP and their molecular mechanism are as follows:

1. *Unfolding/refolding*: HSP70 family members under normal physiological conditions act as molecular chaperones. In response to the stress-induced damages,

intracellular HSP70 bind to the exposed hydrophobic amino acids of non-native conformation of proteins, thus protecting them against denaturation or aggregation until the cell attain the favorable condition (for reviews, see (Boston *et al.* 1996; Hartl 1996)). In conjunction with other chaperones i.e., dimeric HSP40 and co-chaperones i.e., nucleotide exchange factor (NEF), HSP70 recognizes stable misfolded polypeptides and convert them into native proteins by repeated cycle of binding, ATP-dependent unfolding, and spontaneous refolding. Improper unfolding/refolding phenomenon might lead to attachment of substrate to "holdases", including small HSP and HSP90, which maintain the substrate in a non-aggregated folding-component state and pass it to the HSP70 unfoldase machinery for refolding. Additionally, HSP70/HSP110 heterodimer converts protein aggregates into natively unfolded substrates and form NEFs by acting reciprocally on each other and, cooperatively, they efficiently disassemble stable protein aggregates.

2. *Anti-apoptotic Activity*: HSP70 are potent anti-apoptotic proteins and block apoptosis at many different levels. HSP70 block the mitochondrial translocation and activation of Bax, inhibiting mitochondrial membrane permeabilization and release of pro-apoptotic factors, and also inhibit assembly of death inducing signaling complex (DISC) (Gurbuxani *et al.* 2003; Lanneau *et al.* 2008).
3. *Repairing*: HSP70–1 in nucleus, assist the repairing machines of ssDNA by binding to poly (ADP-ribose) polymerase 1 (PARP-1), thus mediating their assembly and initiating their functions.
4. *Tumorigenic*: In cancer cells, a constitutive high-level expression of cytosolic HSP70 is observed frequently. Here, they provide resistance to stress-induced apoptosis, assist in suppressing default senescence, and are correlated with the development of metastasis and drug resistance. HSP70 stabilize the lysosomal membranes and affect autophagy, leading to the survival of cancerous cells. Cell senescence is initiated when HSP70-1 undergoes down-regulation via p53-dependent and p53-independent pathways (Yaglom *et al.* 2007). Another role of extracellular HSP70-1 in tumor invasion and metastasis comes from its ability to increase the MMP-9 expression by activating NF-κB and activating protein-1 (AP1) (Lee *et al.* 2006). However, cytosolic HSP70 have negative impact on cancer patients. Extracellular HSP70 are associated with cancer immunity and thus can be used as drug.
5. *Immunomodulation*: HSP70 act as stimulators of the adaptive immune response through their ability to bind antigenic peptide during intracellular antigen processing. The extracellular HSP70 may act as a danger signal to the innate immune system and is also relevant for the establishment of cancerous and autoimmune diseases. HSP70 exert anti-inflammatory properties, by modulation of cytokine production of dendritic cells that provide a link between innate and adaptive immune response.

HSP70 Superfamily in Mammals

In mammalian cells, there are four major isoforms of HSP70 localized in different organelles: the constitutively expressed heat shock cognate 70 (HSC70/HSPA8/HSP73) in the cytoplasm and nucleus, the stress-induced HSP70 (or HSP72/HSPA1A) in cytoplasm, the glucose-regulated BiP (or Grp78/HSPA5) in endoplasmic reticulum (ER) and mtHSP70 (Grp75/mortalin/HSPA9/mito-HSP70) in mitochondrion. Despite the difference in expression pattern of HSP70 and HSC70, their major functions are same i.e., to avoid protein aggregation; folding and assembly of nascent polypeptides, to refold misfolded or aggregated proteins, to enhance the ubiquitination and the degradation of misfolded protein. These proteins are also involved in translocation of protein through intracellular membrane and show interaction with signal transduction proteins. BiP is a major regulator of ER stress, which binds to the proteins transported to the ER and assist in the formation of quaternary structure. The mtHSP70 is mainly involved in protein transportation from mitochondria.

In humans, HSP70 has 13 members, which share several structural and functional features. For example, HSPA1 (HSP70) (reviewed in (Kampinga *et al.* 2009)) is induced by high temperature, has subfamilies called HSPA1A (HSP70-1) and HSPA1B (HSP70-2), which differ by only two amino acids. The gene sequence of another member, HSPA6 (HSP70B') is 77% similar to the HSPA1 gene. The expression of these proteins is transcriptionally controlled by Heat Shock Factors (HSF) which includes four members: HSF1, HSF2, HSF3 and HSF4. HSF have distinctive and overlapping functions and have tissue-specific patterns of expression. Among all HSF, HSF1 is the prime transcriptional regulator and is required for transactivation of HSP genes and maintenance of thermo-tolerance. During stress, HSF1 is induced and binds to the promoter of HSP70 to enhance its transcription.

Role of HSP70 in Human Diseases

As discussed earlier, the HSP70 chaperones are mainly involved in folding of translated proteins, intracellular localization and prevention of aggregation. Therefore, improper functioning of HSP lead to several diseases related to defects in protein folding or trafficking. The implications of malfunctioning of HSP70 in some of the major human diseases are discussed below:

HSP70 in Cancer

HSP70 act as an important factor in development of different types of cancers and can be used as potential tumor biomarker. Usually HSP70 is overexpressed on the cell surface of tumors. Because of their chaperonin activity as well as cell signaling

regulation activity, HSP70 are involved in tumor cell proliferation, differentiation, invasion, metastasis and death. However, in some cancers (renal and cervix), survival is not correlated with the Hsp70 levels. The sequential increase in the level of HSP70 has a potential prognostic value in patients with chronic hepatitis, liver cirrhosis and liver carcinomas intrahepatic cholangiocarcinoma (IH-ChCa) and metastatic tumors (Yang *et al.* 2010). Hence, change in the expression level of HSP70 could be used as a biomarker and prognosis in cancers like colon cancer, breast cancer, melanoma, bladder cancer, cholangiocarcinoma and squamous cell carcinoma of the head and neck (SCCHN). The clinical outcome of radiotherapy can also be monitored by ascertaining the levels of HSP70 in SCCHN patients. HSP70 are considered as favorable target for treating several cancers. The association of HSP70 and Bag3 (nucleotide exchange factor of HSP70) changes the activity of certain transcription factors (NF-κB, FoxM1, Hif1α), the translation regulator (HuR) and the cell-cycle regulators (p21, survivin) (Colvin *et al.* 2014). One of the ways to check the proliferation of tumors is to induce senescence; in some cases it is done by down-regulating HSP72 via p53-dependentand p53-independent pathways. HSP70 are also associated to base pair excision system, therefore inhibition of HSP70-based DNA repair in cancer cell might be important in chemotherapeutic regimens. Furthermore, the combination of the two chaperones HSP70 and HSP90 along with conventional anti-cancer drugs is a favorable therapeutic selection for patients suffering with advanced bladder cancer.

HSP70 in Apoptosis

Besides their role as molecular chaperones, HSP70 are also anti-apoptotic proteins. HSP70 inhibit the apoptosis at multiple points in intrinsic as well as extrinsic pathways. HSP70 interact with stress-induced kinases and inhibit their functions in apoptosis. In the intrinsic pathway, HSP70 inhibit the disruption of the mitochondrial membrane potential and help to prevent the release of pro-apoptotic factors such as cytochrome c and apoptosis-inducing factor (AIF) (Gurbuxani *et al.* 2003). In extrinsic pathway, it responds to the apoptotic stimulus by inhibiting the assembly of DISC (Lanneau *et al.* 2008). HSP70 also provide protection against hypoxia/reoxygenation-induced apoptosis and maintaining intestinal epithelial cells, with the increase in expression level of BCL-2.

HSP70 in Diabetes Mellitus

The expression of HSP70 is reported to high in type I diabetes mellitus (TIDM) and type II diabetes mellitus (TIIDM) (Nakhjavani *et al.* 2010). Generally, the pancreas regulates the levels of HSP70 where they protect the susceptible beta cells from exocrine pancreatic damage and from the stress associated with insulin hypersecretion. Recently, it has been shown that if the expression of HSP decreases in TIIDM patients, their wound healing process is impaired (Singh *et al.* 2015). It has also

been proposed that one of the most effective and feasible strategy to improve the glucose tolerance in hyperglycemic (i.e. high blood sugar) condition is to increase the HSP70 level, potentially by targeting hyperglycemia-related deficits in HSF1 (Kavanagh *et al.* 2011). HSP70 have a direct correlation with several molecules and can be used as an indicator for variety of diseases. For example, (i) increased serum HSP70 and hemoglobin A1c (HbA1c) levels in women indicates gestational diabetes mellitus, (ii) In patients with high C-reactive protein (CRP) and in case of hunger inhibiting hormone such as leptin, higher levels of HSP70 and asymmetric dimethyl arginine (ADMA) were reported. Furthermore, a relationship between chronic inflammation with diabetes mellitus and diabetes mellitus-associated albuminuria can be postulated from the higher levels of HSP70 observed in diabetic patients with albuminuria (i.e. presence of albumin in the urine).

HSP70 in Obesity, Non-alcoholic Fatty Liver Disease, Alcoholic Fatty Liver Disease and Hepatic Steatosis

The nonalcoholic fatty liver diseases (NAFLD) and hepatic steatosis (HS) induces the risk of type 2 diabetes (T2D) and cardio-cerebrovascular diseases (Qu *et al.* 2015a); moreover, obesity, NAFLD, alcoholic fatty liver disease (AFLD) and HS increase the inflammation. In case of NAFLD, there is a decrease in the expression level of HSF-1 of liver and adipose tissue, which affects the HSP70-dependent anti-inflammation. The HSP70 inhibition in NAFLD patients occurs in kupffer cells. Obese patients with NAFLD also have a lower HSP70 serum concentration (Di Naso *et al.* 2015). Additionally, in comparison to patients with mild alcoholic fatty liver disease (AFLD) or alcohol consuming individuals without AFLD, the lower level of HSP70 was found in AFLD patients (Qu *et al.* 2015b). In case of hepatocellular injury in AFLD patients, HSP70 shows increased positive immunoreactivity and could be used as a sensitive marker.

HSP70 in Chronic Glomerulonephritis

The expression of HSP70 in urine is also higher in case of higher chronic glomerunephritis (CGN) activity and transient creatinine as compared to inactive nephritis, active CGN and preserved renal function, and persistent proteinuria and chronic renal failure (Chebotareva *et al.* 2014).

HSP70 in Stroke and Seizure-Related Pathological Events

One of the vital functions of HSP70 is to prevent the occurrence of apoptosis in brain. The neuro-protective effect is achieved via anti-apoptotic mechanism in association with the overexpression of HSP70 (Zhao *et al.* 2014). Extracellular HSP70 facilitates the production of cytotoxic levels of tumor necrosis factor alpha via

TLR4/MyD88 signaling cascade, which results in increased neuronal death (Dvoriantchikova *et al.* 2014). In case of seizure related pathologic events also, HSP70 has a potential value as a sensitive and specific biomarker.

HSP70 in Helicobacter pylori Infection

Helicobacter pylori (*H. pylori*) are important causative agents of gastritis, peptic ulcer diseases, and mucosa associated lymphoid tissue (MALT) lymphoma and gastric cancer. It has been reported that HSP70 level changes significantly during *H. pylori* infection, viz. H. pylori-associated chronic gastritis, ulcerative colitis, and glutamine-treated patients (Leri *et al.* 1996). Studies also suggested that during *H. pylori* infection, HSP70 expression level decreases. This might involve the initiation of HSP70 expression for cytoprotection against *H. pylori* infection to prevent the expression of inducible nitric oxide synthase (iNOS) (Yeo *et al.* 2004). Pierzchalski *et al.* suggested that HSP70 protects cytoplasmic and nuclear proteins from the damaging effects of bacterial products by delaying the apoptosis of monocytes (Pierzchalski *et al.* 2014).

HSP70 in Atherosclerosis

Atherosclerosis is an inflammatory disease that affects a large human population. HSP70 concentration changes during the progression of atherosclerosis, and thus it is an effective biomarker to monitor atherosclerosis. However, there are contradictory examples over the correlation of HSP70 in atherosclerosis disease: Dulin *et al.* measured a significantly lower concentration of extracellular HSP70 in atherosclerosis patients (Dulin *et al.* 2010) while other group has reported that in patients suffering from carotid artery disease and chronic lower limb ischemia, the concentration of serum HSP70 changed depending on the severity of atherosclerosis (Krepuska *et al.* 2011).

Conclusions

HSP70 are expressed in normal cellular conditions where they regulate protein homeostasis facilitating protein folding and degradation. However, their expression is increased manifold when the cell is exposed to stress. HSP70 protect the cell from stress-induced protein unfolding and other adversities. Thus, their expression has important implications on progression of several human diseases. Therefore, HSP70 have promising role as bio-molecular marker in diagnosis of several diseases and as potential drug targets.

Acknowledgements The work was supported by grants from Indian Council of Medical Research, India to AG (ICMR JRF: 3/1/3/JRF-2016/LS/HRD-3 (32262) and BK (ICMR SRF: BIC/11(33)/2014). Authors also acknowledge efforts of Dr. Neelja Singhal for critical reading of manuscript.

References

Boston, R. S., Viitanen, P. V., & Vierling, E. (1996). Molecular chaperones and protein folding in plants. *Plant Molecular Biology, 32*, 191–222.

Chebotareva, N. V., Neprintseva, N. V., Bobkova, I. N., & Kozlovskaia, L. V. (2014). Investigation of 70-kDa heat shock protein in the serum and urine of patients with chronic glomerulonephritis. *Terapevticheskiĭ Arkhiv, 86*, 18–23.

Colvin, T. A., Gabai, V. L., Gong, J., Calderwood, S. K., Li, H., Gummuluru, S., Matchuk, O. N., Smirnova, S. G., Orlova, N. V., Zamulaeva, I. A., et al. (2014). Hsp70-Bag3 interactions regulate cancer-related signaling networks. *Cancer Research, 74*, 4731–4740.

Csermely, P., Schnaider, T., Soti, C., Prohaszka, Z., & Nardai, G. (1998). The 90-kDa molecular chaperone family: Structure, function, and clinical applications. A comprehensive review. *Pharmacology & Therapeutics, 79*, 129–168.

De Jong, W. W., Caspers, G. J., & Leunissen, J. A. (1998). Genealogy of the alpha-crystallin–small heat-shock protein superfamily. *International Journal of Biological Macromolecules, 22*, 151–162.

De Maio, A. (1999). Heat shock proteins: Facts, thoughts, and dreams. *Shock, 11*, 1–12.

Di Naso, F. C., Porto, R. R., Fillmann, H. S., Maggioni, L., Padoin, A. V., Ramos, R. J., Mottin, C. C., Bittencourt, A., Marroni, N. A., de Bittencourt, P. I., & Jr. (2015). Obesity depresses the anti-inflammatory HSP70 pathway, contributing to NAFLD progression. *Obesity (Silver Spring), 23*, 120–129.

Dulin, E., Garcia-Barreno, P., & Guisasola, M. C. (2010). Extracellular heat shock protein 70 (HSPA1A) and classical vascular risk factors in a general population. *Cell Stress & Chaperones, 15*, 929–937.

Dvoriantchikova, G., Santos, A. R., Saeed, A. M., Dvoriantchikova, X., & Ivanov, D. (2014). Putative role of protein kinase C in neurotoxic inflammation mediated by extracellular heat shock protein 70 after ischemia-reperfusion. *Journal of Neuroinflammation, 11*, 81.

Fink, A. L. (1999). Chaperone-mediated protein folding. *Physiological Reviews, 79*, 425–449.

Flaherty, K. M., DeLuca-Flaherty, C., & McKay, D. B. (1990). Three-dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein. *Nature, 346*, 623–628.

Gragerov, A., Zeng, L., Zhao, X., Burkholder, W., & Gottesman, M. E. (1994). Specificity of DnaK-peptide binding. *Journal of Molecular Biology, 235*, 848–854.

Gurbuxani, S., Schmitt, E., Cande, C., Parcellier, A., Hammann, A., Daugas, E., Kouranti, I., Spahr, C., Pance, A., Kroemer, G., et al. (2003). Heat shock protein 70 binding inhibits the nuclear import of apoptosis-inducing factor. *Oncogene, 22*, 6669–6678.

Hartl, F. U. (1996). Molecular chaperones in cellular protein folding. *Nature, 381*, 571–579.

Hartl, F. U., & Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: From nascent chain to folded protein. *Science, 295*, 1852–1858.

Kalmar, B., & Greensmith, L. (2009). Induction of heat shock proteins for protection against oxidative stress. *Advanced Drug Delivery Reviews, 61*, 310–318.

Kampinga, H. H., Hageman, J., Vos, M. J., Kubota, H., Tanguay, R. M., Bruford, E. A., Cheetham, M. E., Chen, B., & Hightower, L. E. (2009). Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress & Chaperones, 14*, 105–111.

Kavanagh, K., Flynn, D. M., Jenkins, K. A., Zhang, L., & Wagner, J. D. (2011). Restoring HSP70 deficiencies improves glucose tolerance in diabetic monkeys. *American Journal of Physiology. Endocrinology and Metabolism, 300*, E894–E901.

Krepuska, M., Szeberin, Z., Sótonyi, P., Sarkadi, H., Fehérvári, M., Apor, A., Rimely, E., Prohászka, Z., & Acsády, G. (2011). Serum level of soluble HSP70 is associated with vascular calcification. *Cell Stress & Chaperones, 16*, 257–265.

Lanneau, D., Brunet, M., Frisan, E., Solary, E., Fontenay, M., & Garrido, C. (2008). Heat shock proteins: Essential proteins for apoptosis regulation. *Journal of Cellular and Molecular Medicine, 12*, 743–761.

Lee, K. J., Kim, Y. M., Kim, D. Y., Jeoung, D., Han, K., Lee, S. T., Lee, Y. S., Park, K. H., Park, J. H., Kim, D. J., et al. (2006). Release of heat shock protein 70 (Hsp70) and the effects of extracellular Hsp70 on matric metalloproteinase-9 expression in human monocytic U937 cells. *Experimental & Molecular Medicine, 38*, 364–374.

Leri, O., Teichner, A., Sinopoli, M. T., Abbolito, M. R., Pustorino, R., Nicosia, R., & Paparo Barbaro, S. (1996). Heat-shock-proteins-antibodies in patients with Helicobacter pylori associated chronic gastritis. *Rivista Europea per le Scienze Mediche e Farmacologiche, 18*, 45–47.

Miot, M., Reidy, M., Doyle, S. M., Hoskins, J. R., Johnston, D. M., Genest, O., Vitery, M. C., Masison, D. C., & Wickner, S. (2011). Species-specific collaboration of heat shock proteins (Hsp) 70 and 100 in thermotolerance and protein disaggregation. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 6915–6920.

Nakhjavani, M., Morteza, A., Khajeali, L., Esteghamati, A., Khalilzadeh, O., Asgarani, F., & Outeiro, T. F. (2010). Increased serum HSP70 levels are associated with the duration of diabetes. *Cell Stress & Chaperones, 15*, 959–964.

Pierzchalski, P., Jastrzebska, M., Link-Lenczowski, P., Leja-Szpak, A., Bonior, J., Jaworek, J., Okon, K., & Wojcik, P. (2014). The dynamics of heat shock system activation in Monomac-6 cells upon Helicobacter pylori infection. *Journal of Physiology and Pharmacology, 65*, 791–800.

Qiu, X. B., Shao, Y. M., Miao, S., & Wang, L. (2006). The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. *Cellular and Molecular Life Sciences, 63*, 2560–2570.

Qu, B., Jia, Y., Liu, Y., Wang, H., Ren, G., & Wang, H. (2015a). The detection and role of heat shock protein 70 in various nondisease conditions and disease conditions: A literature review. *Cell Stress & Chaperones, 20*, 885–892.

Qu, B. G., Wang, H., Jia, Y. G., Su, J. L., Wang, Z. D., Wang, Y. F., Han, X. H., Liu, Y. X., Pan, J. D., & Ren, G. Y. (2015b). Changes in tumor necrosis factor-alpha, heat shock protein 70, malondialdehyde, and superoxide dismutase in patients with different severities of alcoholic fatty liver disease: A prospective observational study. *Medicine (Baltimore), 94*, e643.

Radons, J. (2016). The human HSP70 family of chaperones: Where do we stand? *Cell Stress & Chaperones, 21*, 379–404.

Ritossa, F. (1962). A new puffing pattern induced by temperature and DNP in Drosophila. *Experientia, 18*, 571–573.

Singh, K., Agrawal, N. K., Gupta, S. K., Mohan, G., Chaturvedi, S., & Singh, K. (2015). Decreased expression of heat shock proteins may lead to compromised wound healing in type 2 diabetes mellitus patients. *Journal of Diabetes and its Complications, 29*, 578–588.

Tissieres, A., Mitchell, H. K., & Tracy, U. M. (1974). Protein synthesis in salivary glands of Drosophila melanogaster: Relation to chromosome puffs. *Journal of Molecular Biology, 84*, 389–398.

Tóth, M. E., Gombos, I., & Sántha, M. (2015). Heat shock proteins and their role in human disease. *Acta Biologica Szegediensis, 59*, 21–141.

Verghese, J., Abrams, J., Wang, Y., & Morano, K. A. (2012). Biology of the heat shock response and protein chaperones: Budding yeast (Saccharomyces cerevisiae) as a model system. *Microbiology and Molecular Biology Reviews, 76*, 115–158.

Vierling, E. (1997). The small heat shock proteins in plants are members of an ancient family of heat induced proteins. *Acta Physiologiae Plantarum, 19*, 539–547.

Whitley, D., Goldberg, S. P., & Jordan, W. D. (1999). Heat shock proteins: A review of the molecular chaperones. *Journal of Vascular Surgery, 29*, 748–751.

Xu, Z. S., Li, Z. Y., Chen, Y., Chen, M., Li, L. C., & Ma, Y. Z. (2012). Heat shock protein 90 in plants: Molecular mechanisms and roles in stress responses. *International Journal of Molecular Sciences, 13*, 15706–15723.

Yaglom, J. A., Gabai, V. L., & Sherman, M. Y. (2007). High levels of heat shock protein Hsp72 in cancer cells suppress default senescence pathways. *Cancer Research, 67*, 2373–2381.

Yang, X., He, H., Yang, W., Song, T., Guo, C., Zheng, X., & Liu, Q. (2010). Effects of HSP70 antisense oligonucleotide on the proliferation and apoptosis of human hepatocellular carcinoma cells. *Journal of Huazhong University of Science and Technology. Medical Sciences, 30*, 337–343.

Yeo, M., Park, H. K., Kim, D. K., Cho, S. W., Kim, Y. S., Cho, S. Y., Paik, Y. K., & Hahm, K. B. (2004). Restoration of heat shock protein 70 suppresses gastric mucosal inducible nitric oxide synthase expression induced by Helicobacter pylori. *Proteomics, 4*, 3335–3342.

Zhao, J. H., Meng, X. L., Zhang, J., Li, Y. L., Li, Y. J., & Fan, Z. M. (2014). Oxygen glucose deprivation post-conditioning protects cortical neurons against oxygen-glucose deprivation injury: Role of HSP70 and inhibition of apoptosis. *Journal of Huazhong University of Science and Technology. Medical Sciences, 34*, 18–22.

Zhu, X., Zhao, X., Burkholder, W. F., Gragerov, A., Ogata, C. M., Gottesman, M. E., & Hendrickson, W. A. (1996). Structural analysis of substrate binding by the molecular chaperone DnaK. *Science, 272*, 1606–1614.

Published online 18 May 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W239–W243***
doi: 10.1093/nar/gkaa385

mRNALoc: a novel machine-learning based *in-silico* tool to predict mRNA subcellular localization

Anjali Garg, Neelja Singhal, Ravindra Kumar and Manish Kumar *

Department of Biophysics, University of Delhi South Campus, New Delhi 110021, India

Received March 03, 2020; Revised April 14, 2020; Editorial Decision April 27, 2020; Accepted April 30, 2020

ABSTRACT

Recent evidences suggest that the localization of mRNAs near the subcellular compartment of the translated proteins is a more robust cellular tool, which optimizes protein expression, post-transcriptionally. Retention of mRNA in the nucleus can regulate the amount of protein translated from each mRNA, thus allowing a tight temporal regulation of translation or buffering of protein levels from bursty transcription. Besides, mRNA localization performs a variety of additional roles like long-distance signaling, facilitating assembly of protein complexes and coordination of developmental processes. Here, we describe a novel machine-learning based tool, mRNALoc, to predict five sub-cellular locations of eukaryotic mRNAs using cDNA/mRNA sequences. During five fold cross-validations, the maximum overall accuracy was 65.19, 75.36, 67.10, 99.70 and 73.59% for the extracellular region, endoplasmic reticulum, cytoplasm, mitochondria, and nucleus, respectively. Assessment on independent datasets revealed the prediction accuracies of 58.10, 69.23, 64.55, 96.88 and 69.35% for extracellular region, endoplasmic reticulum, cytoplasm, mitochondria, and nucleus, respectively. The corresponding values of AUC were 0.76, 0.75, 0.70, 0.98 and 0.74 for the extracellular region, endoplasmic reticulum, cytoplasm, mitochondria, and nucleus, respectively. The mRNALoc standalone software and web-server are freely available for academic use under GNU GPL at http://proteininformatics.org/mkumar/mrnaloc.

INTRODUCTION

Localization of mRNA is an evolutionarily conserved phenomenon that controls many important biological processes like cell-fate determination and polar cell growth (1). After post-transcriptional modifications, such as 5′ capping, splicing and addition of 3′ poly (A) tail, the nascently transcribed mRNA either gets localized within the nucleus or alternatively travels out of the nucleus. It has been suggested that mRNA localization has many advantages over protein localization (2–6). These are: (a) localization of mRNA to a specific location helps the cell to build a local repository of proteins at the site of function instead of transporting individual protein molecules to the site of function. This also compartmentalizes protein synthesis and forms a protein gradient within the cells, which ultimately results in local synthesis of encoded proteins at the target site; (b) mRNA localization works as a translation/co-translational regulator; (c) mRNA localization is a better energy-efficient pathway compared to protein targeting and; (d) mRNA localization aids in formation of only functional and non-harmful multi-protein complexes which aids in avoiding unnecessary protein-protein interactions that might be harmful to the cells (7,8). Not all protein synthesis occurs after mRNA localization. A large number of mRNA sequences are also transported co-translationally (9).

Five different mechanisms namely, diffusion and localized entrapment, localized degradation, localized synthesis, active transport and, polarized nuclear export are considered important for mRNA localization. However, ribonucleoprotein transport complex is the main mode by which majority of RNA is transported. Building the ribonucleoprotein complex is a sequence specific phenomenon, which is guided by a short stretch of 20–200 *cis*-acting nucleotide sequences known as 'zipcode'. It is located at the 3′ untranslated region of the mRNA sequence, although in some cases they can also be present in the 5′UTR or in the coding sequence (10,11). Proteins present in a subcellular compartment are related to the physiological and metabolic function associated with that subcellular compartment. Hence, prediction of subcellular location of mRNA might suggests the biological function of the gene from which the mRNA was transcribed. Thus, a tool that can predict the correct intracellular location of transcripts may also help in understanding how gene expression is regulated and, how cells achieve polarity.

To our knowledge, computational predictors that can predict the subcellular localization of eukaryotic mRNA are unavailable, till date. Hence, we developed a Support Vector Machine (SVM) based *in-silico* tool which can predict the eukaryotic mRNA subcellular locations on the

*To whom correspondence should be addressed. Tel: +91 11 24157263; Email: manish@south.du.ac.in

basis of primary sequence information of mRNA/cDNA. Named as mRNALoc (acronym for 'mRNA Localization), this tool is based on the experimentally validated localization data of mRNA retrieved from 'RNALocate' (12).

ANALYSIS WORKFLOW

Data sources

In the present work, we collected the mRNA sequences and their subcellular location information from RNALocate database (version 2.0) (12). RNALocate is a manually curated database that provides complete subcellular location annotation of RNA with experimental support. Initially, a total of 28829 mRNA sequences with annotated subcellular localization were obtained. The downloaded mRNA sequences revealed their localization to both single and multiple subcellular locations. In the present study, we considered only those mRNA sequences which showed single locations. The mRNA dataset was classified in five subgroups on the basis of subcellular locations namely, cytoplasm, endoplasmic reticulum, extracellular, mitochondria and nucleus. The number of mRNA sequences in the five locations were as follows: 6964 in cytoplasm, 1998 in endoplasmic reticulum, 1131 in extracellular region, 442 in mitochondria and 6346 in nucleus.

Since, redundant mRNA sequences results in overestimation of prediction capability, hence to reduce the redundancy and to avoid homology bias in prediction, we used NCBI BLASTCLUST program to retain only sequences showing alignment identity ≤40% over 70% or more of their full length (BLASTCLUST with '-S 40 and -L 0.7' option) (13). The final non-redundant mRNA dataset contained 6376 sequences of cytoplasm, 1426 sequences of endoplasmic reticulum, 855 sequences of extracellular region, 421 sequences of mitochondria and 5831 mRNA sequences of nucleus. 5/6 part of total 40% non-redundant data was used for training the model. Remaining 1/6 data was used for the independent evaluation of the trained model. For detail about collection of dataset, redundancy removal, constructions of training and independent datasets please see supplementary material. The NCBI gene accession numbers, mRNA sequences and subcellular locations are available in the download section of mRNALoc webserver (http://proteininformatics.org/mkumar/mrnaloc/download.html).

Overview of mRNALoc

mRNALoc is a web resource to predict the subcellular localization of eukaryotic mRNA. The overall workflow of mRNALoc is shown in Figure 1. Users have to provide the mRNA sequences in a FASTA format. The submitted mRNA sequence will be converted into numerical encoding using pseudo oligonucleotide composition or pseudo K-tuple nucleotide composition (PseKNC) (14–17). On the basis of SVM prediction score, the mRNA will be predicted to localize at one of the five subcellular locations, namely cytoplasm, endoplasmic reticulum, extracellular location, mitochondria and nucleus. mRNALoc prediction is based on the five trained SVM models, each specific for one location. During prediction each model provides the prediction score for its corresponding location. The subcellular location, whose SVM model gets the maximum score, will be the predicted location. The final outcome of mRNALoc depends on the user-selected threshold. Higher thresholds would result in more specific predictions, while lower threshold would result in low specificity predictions.

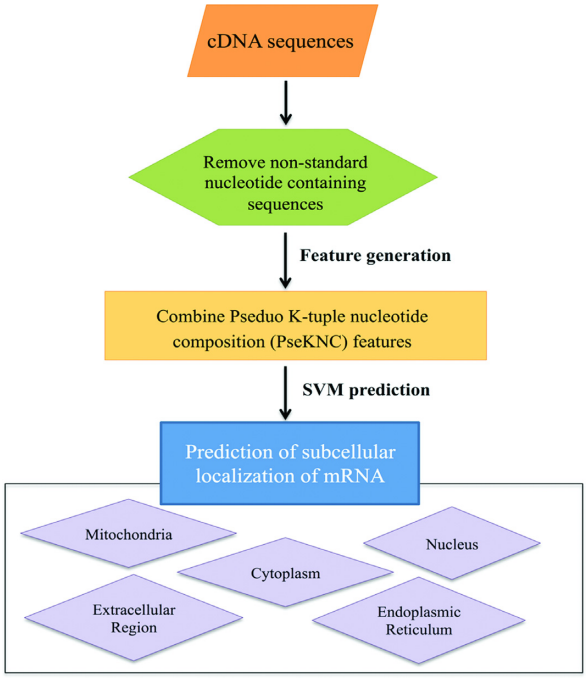

Figure 1. Overall schema of mRNALoc. mRNALoc predicts five subcellular locations *viz.,* mitochondria, cytoplasm, nucleus, endoplasmic reticulum and extracellular. Firstly, it removes the sequences from the query that has non-standard nucleotides then generates combined features from pseudo K-tuple nucleotide composition, which is further used as input for Support Vector Machine (SVM) prediction.

TRAINING OF PREDICTIVE SVM MODELS

The performance mRNALoc for all subcellular locations during five-fold cross-validation mode of training is shown in Table 1. Using the combined input of PseKNC ($K = 2$, 3, 4 and 5) we found 65.19, 75.36, 67.10, 99.70 and 73.59% accuracy of prediction for mRNA whose subcellular locations were extracellular region, endoplasmic reticulum, cytoplasm, mitochondria and nucleus, respectively. When evaluated on an independent dataset, mRNALoc did prediction with sensitivity, specificity, accuracy and MCC values of 81.38, 56.67, 58.10 and 0.18 for extracellular region, 75.10, 68.60, 69.23 and 0.27 for endoplasmic reticulum, 73.26, 58.06, 64.55 and 0.31 for cytoplasm, 87.32, 97.16, 96.88 and 0.63 for mitochondria and 50.20, 81.62, 69.35 and 0.34 for nucleus, respectively (Supplementary Figures S1 and S2, Supplementary Table S1).

COMPARISON WITH EXISTING MRNA SUBCELLULAR LOCALIZATION PREDICTION METHODS

Though, the role of mRNA localization is unambiguously established in cellular physiology, attempts to build *in-silico*

Table 1. The performance metrics for mRNA subcellular localization under hybrid *K*-mer feature (2+3+4+5), and performance of the SVM based classifiers (mRNALoc) on independent data

Location	Sen (%)	Spe (%)	ACC (%)	MCC	THR	AUC
Training dataset						
Extracellular region	62.67	65.34	65.19	0.14	−0.20	0.69
Endoplasmic reticulum	74.09	75.49	75.36	0.32	0.40	0.81
Cytoplasm	66.69	67.41	67.10	0.34	0.40	0.69
Mitochondria	96.28	99.79	99.70	0.95	0.10	0.98
Nucleus	74.17	73.22	73.59	0.47	0.40	0.76
Independent dataset						
Extracellular region	81.38	56.67	58.10	0.18	−0.20	0.76
Endoplasmic reticulum	75.10	68.60	69.23	0.27	0.40	0.75
Cytoplasm	73.26	58.06	64.55	0.31	0.40	0.70
Mitochondria	87.32	97.16	96.88	0.63	0.10	0.98
Nucleus	50.20	81.62	69.35	0.34	0.40	0.74

Sen: sensitivity, Spe: specificity, ACC: accuracy, MCC: Mathews correlation coefficient, THR: threshold, and AUC: area under ROC curve.

Table 2. Comparative evaluation of mRNALoc and iLoc-mRNA. In extracellular region and mitochondria no human mRNA was present, hence these two locations were not included in the evaluation

		mRNALoc		iLoc-mRNA	
Location	Number of human mRNA sequences	True positive	False negative	True positive	False negative
Cytoplasm	50	35	15	18	32
Endoplasmic reticulum	50	34	16	37	13
Extracellular region	0	0	0	0	0
Mitochondria	0	0	0	0	0
Nucleus	50	33	17	13	37

tools to predict the subcellular localizations of mRNA are negligible in comparison to protein subcellular localization prediction tools. Recently, Yan *et al.* proposed a deep-learning based method, named as RNATracker (18), to predict the subcellular localization of mRNA using data from CeFra-Seq (19) and APEX-RIP (3). Using the data from RNALocate, a human mRNA subcellular localization method iLoc-mRNA was also developed (20).

Though, both RNATracker and iLoc-mRNA are based on two different mRNA subcellular localization datasets and, were developed using two different approaches, mRNALoc has several advantages over both RNATracker and iLoc-mRNA. For example, (a) localization data produced by CeFra-Seq/APEX-RIP are inherently noisy and sometimes inaccurate also (18). The mRNALoc was developed from datasets retrieved from RNALocate (12), which contains manually curated mRNA subcellular localization information with experimental evidences. (b) The RNATracker among all the isoforms, considered only the longest isoform while, mRNALoc did not made any such distinction. (c) Redundant mRNA sequences were not removed from RNATracker and in iLoc-mRNA the redundancy threshold was 80%. While in mRNALoc, we used 40% non-redundant mRNA sequences to train the predictor. This may be the reason underlying high MCC and AUC for RNATracker and iLoc-mRNA. (d) Both, RNATracker and iLoc-mRNA were developed using only localization data of human mRNA. On the contrary, mRNALoc is a general-purpose eukaryotic mRNA subcellular localization prediction tool, which is applicable to all eukaryotes. (e) RNATracker also excluded low expressed genes, but mRNALoc made no such distinction (Supplementary Table S2).

We also conducted one-to-one comparison of performance of iLoc-mRNA and mRNALoc. As RNATracker required gene expression and coordination files for prediction, it was not possible to include it in the evaluation. For comparison we used the independent dataset of mRNALoc. Since, iLoc-mRNA is specifically designed for human mRNA subcellular localization prediction, we used 50 human mRNA sequences of independent dataset of mRNALoc. The number of human mRNA in different locations and prediction result of mRNALoc and iLoc-mRNA is shown in Table 2. In extracellular region and mitochondria, we didn't find human mRNA sequences in mRNALoc independent dataset hence, these locations were not included in the evaluation.

As shown in Table 2, for cytoplasm and nucleus the performance of mRNALoc was better than iLoc-mRNA but, in endoplasmic reticulum the performance of iLoc-mRNA was better than mRNALoc. It is also pertinent to mention that in iLoc-mRNA prediction were made for one of the following locations namely, cytosol/cytoplasm, ribosome, endoplasmic reticulum, and nucleus/exosome/dendrite/mitochondrion. We feel that combining nucleus, exosome, dendrite, and mitochondria as a single location is not appropriate as these are diverse subcellular locations which should not be merged in a single category.

DESCRIPTION OF THE WEBSERVER

Implementation of mRNALoc

The web server is hosted on a Linux system. The back-end pipeline is implemented in the Perl language. The webserver has an intuitive interface and 'how-to' guide to help the user.

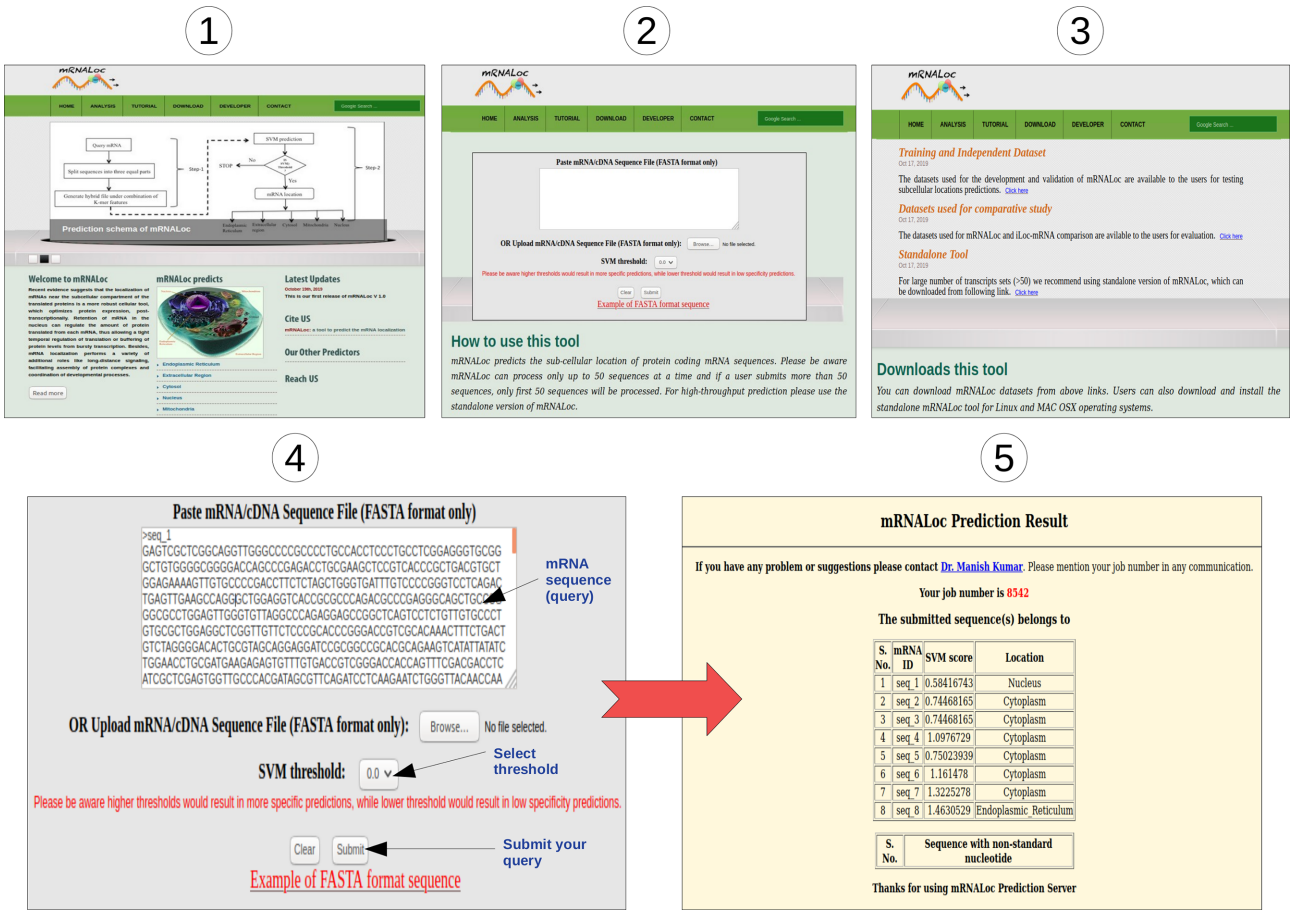

Figure 2. Screenshots of mRNALoc webserver.

Each mRNA query sequence must be at least 100 bp long and contains only valid characters, namely 'A', 'C', 'G' and 'T/U'. Sequences having non-standard nucleotides will be omitted from the prediction pipeline (Figure 2).

The output of mRNALoc

The output of mRNALoc is presented in a tabular format. It contains the highest scores obtained from the five SVM models and the location to which the mRNA is assigned. A maximum of fifty sequences can be processed by mRNALoc webserver in one go. Hence, for genome scale prediction a standalone version will be required (Figure 2 and Supplementary Figure S3).

CONCLUSIONS AND FUTURE PROSPECTS

The annotation of subcellular localization has been addressed mainly at the protein level. Many *in silico* tools were developed to predict protein subcellular location using machine-learning techniques. It has been unequivocally established that both mRNA and protein localization play an equal role in protein translocation. In future versions of mRNALoc we would like to overcome some of the limitations of the present tool. The first and foremost is that our tool is currently limited by the accuracy of the RNALocate datasets. Though, RNALocate contain data from 65 organisms, most of the data is enriched with the common model organisms like, *Homo sapiens*, *Mus musculus*, and *Saccharomyces cerevisiae* etc. Moreover, considering at the biological level, instead of cytosol, mitochondria or extracellular locations, axons, dendrites, dendritic spines, or anterior/posterior vs dorsal/ventral locations are more relevant. Another, limitation is that due to lesser availability of plant mRNA localization data compared to other domains of life, mRNALoc performance might be compromised (21). The performance of a machine-learning method depends on the data on which it is trained. We believe that with development of new and better RNA localization finding techniques, information about RNA localization in plants would also be available in the near future and future versions of mRNALoc would then support prediction of plant mRNA sequences, also. We admit that mRNALoc is in an early stage of development and training on additional datasets is needed to further improve our tool. Further prediction of mRNA localization will also help in predicting the novel zipcodes which may guide researchers to cast new hypothesis for unraveling the finer details of mechanism of mRNA-protein complex formation which is actually responsible for mRNA location. Though, the current version of mRNALoc supports prediction of only eukaryotic mRNA, the future versions of mRNALoc would definitely include data from other organisms and locations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Indian Council of Medical Research (ICMR)-JRF scheme [3/1/3 J.R.F.-2016/LS/HRD-(32262) to A.G.]; CSIR Senior Research Associateship (Scientists' Pool Scheme) [9089A/2019-Pool to N.S.]. Funding for open access charge: Indian Council of Medical Research.
Conflict of interest statement. None declared.

REFERENCE

1. Kloc,M., Zearfoss,N.R. and Etkin,L.D. (2002) Mechanisms of subcellular mRNA localization. *Cell*, **108**, 533–544.
2. Lecuyer,E., Yoshida,H., Parthasarathy,N., Alm,C., Babak,T., Cerovina,T., Hughes,T.R., Tomancak,P. and Krause,H.M. (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, **131**, 174–187.
3. Kaewsapsak,P., Shechner,D.M., Mallard,W., Rinn,J.L. and Ting,A.Y. (2017) Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. *eLife*, **6**, e29224.
4. Medioni,C., Mowry,K. and Besse,F. (2012) Principles and roles of mRNA localization in animal development. *Development*, **139**, 3263–3276.
5. Wang,E.T., Taliaferro,J.M., Lee,J.A., Sudhakaran,I.P., Rossoll,W., Gross,C., Moss,K.R. and Bassell,G.J. (2016) Dysregulation of mRNA localization and translation in genetic disease. *J. Neurosci.*, **36**, 11418–11426.
6. Hughes,S.C. and Simmonds,A.J. (2019) Drosophila mRNA localization during later Development: Past, Present, and Future. *Front. Genet.*, **10**, 135.
7. Di Liegro,C.M., Schiera,G. and Di Liegro,I. (2014) Regulation of mRNA transport, localization and translation in the nervous system of mammals (Review). *Int. J. Mol. Med.*, **33**, 747–762.
8. Vandepoele,K., Simillion,C. and Van de Peer,Y. (2002) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. *Trends Genet.*, **18**, 606–608.
9. Weis,B.L., Schleiff,E. and Zerges,W. (2013) Protein targeting to subcellular organelles via mRNA localization. *Biochim. Biophys. Acta*, **1833**, 260–273.
10. Heasman,J., Wessely,O., Langland,R., Craig,E.J. and Kessler,D.S. (2001) Vegetal localization of maternal mRNAs is disrupted by VegT depletion. *Dev. Biol.*, **240**, 377–386.
11. Kloc,M. and Etkin,L.D. (1994) Delocalization of Vg1 mRNA from the vegetal cortex in Xenopus oocytes after destruction of Xlsirt RNA. *Science (New York, N.Y.)*, **265**, 1101–1103.
12. Zhang,T., Tan,P., Wang,L., Jin,N., Li,Y., Zhang,L., Yang,H., Hu,Z., Zhang,L., Hu,C. *et al.* (2017) RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.*, **45**, D135–D138.
13. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
14. Liu,B., Wang,S., Long,R. and Chou,K.C. (2017) iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*, **33**, 35–41.
15. Liu,B., Liu,F., Fang,L., Wang,X. and Chou,K.C. (2015) repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics*, **31**, 1307–1309.
16. Chen,W., Feng,P., Ding,H., Lin,H. and Chou,K.C. (2015) iRNA-Methyl: identifying N(6)-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.*, **490**, 26–33.
17. Liu,B., Fang,L., Long,R., Lan,X. and Chou,K.C. (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, **32**, 362–369.
18. Yan,Z., Lecuyer,E. and Blanchette,M. (2019) Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*, **35**, i333–i342.
19. Benoit Bouvrette,L.P., Cody,N.A.L., Bergalet,J., Lefebvre,F.A., Diot,C., Wang,X., Blanchette,M. and Lecuyer,E. (2018) CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells. *RNA*, **24**, 98–113.
20. Zhang,Z.Y., Yang,Y.H., Ding,H., Wang,D., Chen,W. and Lin,H. (2020) Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinformatics*, doi:10.1093/bib/bbz177.
21. Tian,L., Chou,H.L., Fukuda,M., Kumamaru,T. and Okita,T.W. (2020) mRNA localization in plant cells. *Plant Physiol.*, **182**, 97–109.

SCIENTIFIC REPORTS
nature research

OPEN

Comparative *in-silico* proteomic analysis discerns potential granuloma proteins of *Yersinia pseudotuberculosis*

Manisha Aswal[1,2], Anjali Garg[1,2], Neelja Singhal[1] & Manish Kumar[1*]

Yersinia pseudotuberculosis **is one of the three pathogenic species of the genus *Yersinia*. Most studies regarding pathogenesis of *Y. pseudotuberculosis* are based on the proteins related to Type III secretion system, which is a well-known primary virulence factor in pathogenic Gram-negative bacteria, including *Y. pseudotuberculosis*. Information related to the factors involved in *Y. pseudotuberculosis* granuloma formation is scarce. In the present study we have used a computational approach to identify proteins that might be potentially involved in formation of *Y. pseudotuberculosis* granuloma. A comparative proteome analysis and conserved orthologous protein identification was performed between two different genera of bacteria - *Mycobacterium* and *Yersinia*, their only common pathogenic trait being ability to form necrotizing granuloma. Comprehensive analysis of orthologous proteins was performed in proteomes of seven bacterial species. This included *M. tuberculosis*, *M. bovis* and *M. avium paratuberculosis* - the known granuloma forming *Mycobacterium* species, *Y. pestis* and *Y. frederiksenii* - the non-granuloma forming *Yersinia* species and, *Y. enterocolitica* - that forms micro-granuloma and, *Y. pseudotuberculosis* - a prominent granuloma forming *Yersinia* species. *In silico* proteome analysis indicated that seven proteins (UniProt id A0A0U1QT64, A0A0U1QTE0, A0A0U1QWK3, A0A0U1R1R0, A0A0U1R1Z2, A0A0U1R2S7, A7FMD4) might play some role in *Y. pseudotuberculosis* granuloma. Validation of the probable involvement of the seven proposed *Y. pseudotuberculosis* granuloma proteins was done using transcriptome data analysis and, by mapping on a composite protein-protein interaction map of experimentally proved *M. tuberculosis* granuloma proteins (RD1 locus proteins, ESAT-6 secretion system proteins and intra-macrophage secreted proteins). Though, additional experiments involving knocking out of each of these seven proteins are required to confirm their role in *Y. pseudotuberculosis* granuloma our study can serve as a basis for further studies on *Y. pseudotuberculosis* granuloma.**

The genus *Yersinia* is comprised of Gram-negative, catalase-positive, facultative anaerobic enteric-bacteria. Though, the optimal temperature for growth is 28 °C, some members of the genus can survive at low temperatures *ca*. 4 °C[1]. Most species of the genus *Yersinia* grow extracellularly, except *Y. pseudotuberculosis* and *Y. pestis* which are capable of intracellular growth, *i.e.* inside the host macrophages[2]. Of the sixteen known species of *Yersinia*, only three are pathogenic *Y. pestis*, *Y. pseudotuberculosis* and *Y. enterocolitica*[3,4]. In humans, infection with *Y. pestis* results in plague, while infection with *Y. pseudotuberculosis* and *Y. enterocolitica* results in gastroenteritis, which is usually self-limiting[5]. Besides man, *Y. pseudotuberculosis* can infect a wide range of animals including swines, dogs, rodents, birds *etc*.[6–9]. *Y. pseudotuberculosis* infection in animals can lead to tuberculosis-like symptoms, including localized tissue necrosis and granulomas in the liver, spleen, and lymph nodes.

Plasmid-encoded *Yersinia* outer proteins (Yops) have been regarded as an essential virulence factor of the pathogenic *Yersinia* spp., which restrain the host immune mechanisms to the local lymph nodes. When *Yersinia* and the target host cell come in mutual contact, Yops are delivered in the host cells with the help of type III secretion system (T3SS)[10]. Since polymorphonuclear neutrophils are the first cells to reach the infection site, these are regarded as the main targets of Yps T3SS-mediated Yops translocation[11]. Though, multiple factors underlie

[1]Department of Biophysics, University of Delhi South Campus, New Delhi, 110021, India. [2]These authors contributed equally: Manisha Aswal and Anjali Garg. *email: manish@south.du.ac.in

the virulence mechanism, the current knowledge about virulence of *Y. pseudotuberculosis* is based mainly on the secretion systems. To date, information about *Y. pseudotuberculosis* proteins, which help in granuloma formation, sustenance, expansion and dissemination in the host is fragmentary. Thus, the present study was conducted to identify *Y. pseudotuberculosis* proteins involved in granuloma formation using a system biology approach.

The present study is based on three *Mycobacterium* spp. that included *M. tuberculosis* (Mtb), *M. bovis* (Mbov) and *M. avium paratuberculosis* (Map) - the known granuloma forming *Mycobacterium* species, *Y. pestis* (Ype) and *Y. frederiksenii* (Yfr) - the non-granuloma forming *Yersinia* species and, *Y. enterocolitica* (Yen) - that forms micro-granuloma and, *Y. pseudotuberculosis* (Yps) - a prominent granuloma forming *Yersinia* species. We have considered Yen as non-granuloma forming since it forms non-prominent micro-granuloma, unlike Yps, which forms a prominent granuloma. Despite, the fact that Mtb and Yps belong to two different microbial genera, the granuloma formed by both share several common features: (i) both form a similar type of granuloma in the host which is both necrotizing and infectious, different from other forms of granuloma, (ii) the pathological symptoms of *Yersinia* infection *i.e.* granulomatous ileitis, colitis and appendicitis (causative agent - Yen and Yps) are similar with the symptoms of tuberculosis (causative agent – Mtb) (iii) cellular infiltration of neutrophils and histiocytes is found in the lesions produced by both Yps and mycobacteria[12–15]. Despite many similarities, there are also a few differences between tubercular and pesudotubercular granuloma *viz.* (i) mycobacterial lesion (non-suppurative) involves activation of T-cell mediated immune response[16,17] while in a *Yersinia* lesion (suppurative) despite the presence of T-cells and macrophages, there is suppression of T- and B-cells (adaptive immunity) due to the release of virulence factors[18,19] (ii) both form a central necrotic granuloma, the structure underlying necrotic tissue is maintained in *Yersinia* granuloma but is inconspicuous in a tuberculous granuloma[17]. Along with Yps, Yen also causes granulomatous ileitis, colitis and appendicitis. Researchers have reported that evident granuloma are formed in case of *Mycobacterium* species, Yps, *Chlamydia* species and *Treponema* species, whereas micro-granulomas are formed by Yen, *Salmonella* spp. and *Campylobacter* spp. Yps infection is characterized by a granulomatous process with central micro abscess, while Yen granulomas are accompanied by an acute inflammation and hemorrhagic necrosis[20]. Also, it has been suggested that Yen infection in gastrointestinal tract is usually not associated with granuloma. Infection with Yen has been characterized by mucosal ulceration, often initially overlying Peyer's patches, with accompanying hemorrhagic necrosis, palisading histiocytes and lymphoid hyperplasia[21–23]. Gastrointestinal infection with Yps is usually described as a granulomatous process with central micro abscesses, and almost always accompanied by mesenteric adenopathy[23–25].

In the present study, similarities and differences were discerned in the proteomes of seven bacteria using a bottom-up approach. The proteomes that were compared, included proteomes of three mycobacterial spp. (*M. tuberculosis* H37Rv, *M. bovis* ATCC BAA-935/AF2122/97 and *M. avium paratuberculosis* ATCC BAA-968/K-10 and four *Yersinia* spp. like *Y. pestis* CO-92/Biovar *orientalis, Y. enterocolitica* NCTC 13174/8081, *Y. pseudotuberculosis* IP 31758 and *Y. frederiksenii* ATCC 33641. The primary step was to define the core proteome, which was done using sequence homology. Gradually, we narrowed down our study towards species-specific proteomes. The rationale of our study was based on following basic premises: (a) proteins, which are present in all the seven proteomes, should be the part of conserved core gene set and should be involved in house-keeping functions, (b) proteins which are present in all the species of either *Yersinia* or mycobacteria should also be involved in genus-specific house-keeping functions and, (c) if we remove the proteins of category (a) and (b) the only common proteins between Yps and *Mycobacterium* spp. might be the proteins which help in granuloma formation, as these proteins are not shared with other species of *Yersinia*. In the present work, we have proposed the proteins of category (c) as putative granuloma forming proteins. Functional annotation of these proteins suggested involvement of these proteins in granuloma formation in Yps.

Results

Comparative analysis of genome and proteome relatedness. On the basis of average nucleotide indentity (ANI) of genomes and percentage of conserved proteins (PCOP) content of proteomes, relatedness among all the seven bacterial species was analyzed. Figure 1(a,b) shows relationships among the seven genomes and proteomes, respectively. Both the genomic and proteomic trees divided all the species into two branches; one specific to *Yersinia* spp. and the other to *Mycobacterium* spp. Among the four *Yersinia* spp. Yen and Yfr formed a common cluster while Ype and Yps formed a separate cluster. Among *Mycobacterium* spp., Mbov and Mtb were present in one cluster, and Map was present in a separate cluster.

Comparative analysis of proteomes of mycobacteria and *Yersinia*. A pairwise comparative analysis between different proteomes revealed that Mbov (causative agent of bovine TB) and Mtb (causative agent of human TB) shared 3845 proteins. Map shared 2602 and 2632 proteins with Mbov and Mtb, respectively (Table 1). Interestingly, Map shared more protein with Yen (926) and Yfr (959), than with Yps (896) and Ype (882). Among the different species of *Yersinia*, the two top most closely related proteomes on the basis of number of shared proteins, were Yps and Ype (number of shared proteins was 3387), and, Yen and Yfr (number of shared proteins was 3304).

Clustering of orthologous proteins. All the possible combinations of the proteomes yielded 127 different protein clusters, of which only 85 clusters contained proteins (Fig. 2). An inter-genus comparison of proteins conserved across the seven proteomes resulted in a conserved set of 693 proteins. We also performed an intra-genus conservation analysis and found 1684 proteins for *Mycobacterium* spp. and 1727 proteins for *Yersinia* spp. The number of proteins which were present only in Yps, Ype, Yfr, Yen, Mtb, Map and Mbov were 642, 369, 738, 428, 106, 1483 and 119, respectively.

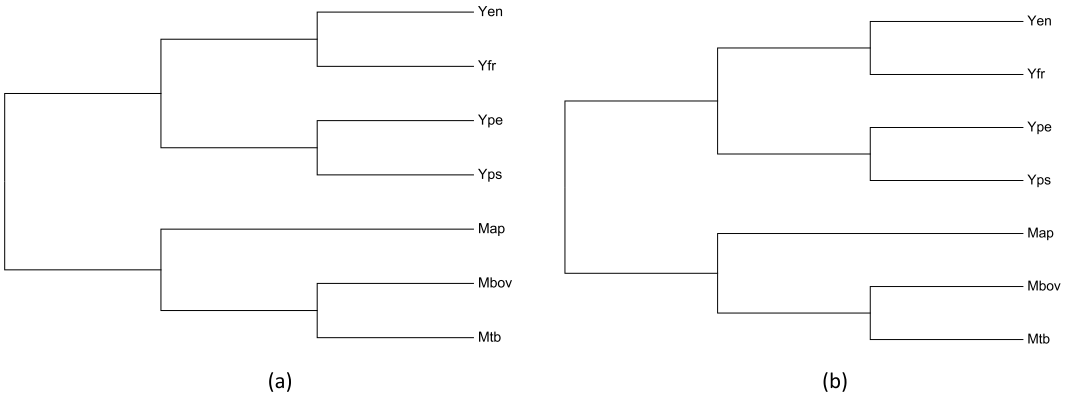

Figure 1. Relatedness among the seven bacterial species on the basis of their genome and proteomes. Cladograms were generated using Neighbor-Joining method[78] using (**a**) average nucleotide identity (ANI) and (**b**) Percentage of conserved proteins (PCOP).

S. No.	Microbe	Number of proteins in complete proteome	Number of shared orthologous proteins						
			Mbov	Mtb	Map	Yen	Yfr	Ype	Yps
1	Mbov	3976	—	3845	2602	892	917	868	868
2	Mtb	3993		—	2632	893	919	872	874
3	Map	4316			—	926	959	882	896
4	Yen	4026				—	3303	2809	2885
5	Yfr	3909					—	2789	2933
6	Ype	4354						—	3387
7	Yps	4305							—

Table 1. Total proteome size of the seven microbes and the number of shared orthologous proteins. Mbov- *M. bovis*, Mtb- *M. tuberculosis*, Map- *M. avium paratuberculosis*, Yen- *Y. enterocolitica*, Yfr- *Y. frederiksenii*, Ype- *Y. pestis*, Yps- *Y. pseudotuberculosis*.

Functional analysis of orthologous protein cluster. Inter genus analysis of proteomes revealed presence of a conserved set of 693 proteins in the seven proteomes. Presence of these 693 proteins across the seven proteomes indicated their involvement in vital functions of the seven microbes. This also implied that these 693 proteins might be considered as the core proteome. To validate this assumption, we analyzed the top 10 enriched GO terms of biological processes associated with these proteins. Enrichment analysis revealed involvement of these 693 proteins in carbohydrate biosynthesis, amino acid biosynthesis and transport, sulfur metabolism, cell wall biosynthesis and overall metabolic processes (Fig. 3). Thus the GO term enrichment analysis also validated our assumption that these proteins were involved in housekeeping functions. Among the 1684 proteins that were conserved only in *Mycobacterium* spp., the enriched GO terms of biological process was biosynthesis of sulfur containing amino acids (cysteine and methionine) and metabolic proteins, in addition to the functions that were enriched in inter-genus core protein set. The 1727 proteins conserved only in *Yersinia* spp., were mainly metabolic and reproduction/mitotic cell cycle proteins. Yen shared 18 proteins with *Mycobacterium* spp. (Mbov, Map and Mtb). These proteins were mostly related to vitamin metabolism, DNA regulation and transport. While, Yfr and *Mycobacterium* spp. (Mbov, Map and Mtb) shared 40 proteins, which were mainly involved in DNA replication, repair, regulation, and metabolism. The five proteins, which were common between Ype and *Mycobacterium* spp. (Mbov, Map and Mtb) were involved in metabolism, DNA replication and protein modification. Interestingly, it was observed that seven proteins were common between Yps and *Mycobacterium* spp. (Mbov, Map and Mtb). These proteins were involved in lipid metabolism, cell wall synthesis and, pyruvate and aldehyde metabolism. It is pertinent to mention here that proteins of each ortholog cluster were mutually exclusive. For example, the seven proteins common among Yps and *Mycobacterium* spp. were not a part of any other ortholog clusters.

Functional characterization of common orthologs of Yps and *Mycobacterium spp*. and their probable involvement in granuloma formation. A comparative analysis of protein conservation in the three *Mycobacterium* spp. and four *Yersinia* spp. revealed that seven Yps proteins were present in the three *Mycobacterium* spp. but absent in other species of *Yersinia*. Since, the only common feature in the three *Mycobacterium* spp. and Yps is their capability to form macro-granuloma, it might be anticipated that these seven proteins might play a potential role in granuloma formation (Table 2). To validate the role of these proteins in Yps granuloma, a detailed functional characterization of all the seven proteins was performed using UniProtKB, STRING and KEGG databases. It was observed that, of the seven proteins, two proteins were functionally uncharacterized, while functions of five proteins were known. The details of the seven proteins with their UniProt ids, interaction partners and pathway information analysis are presented in Table 3.

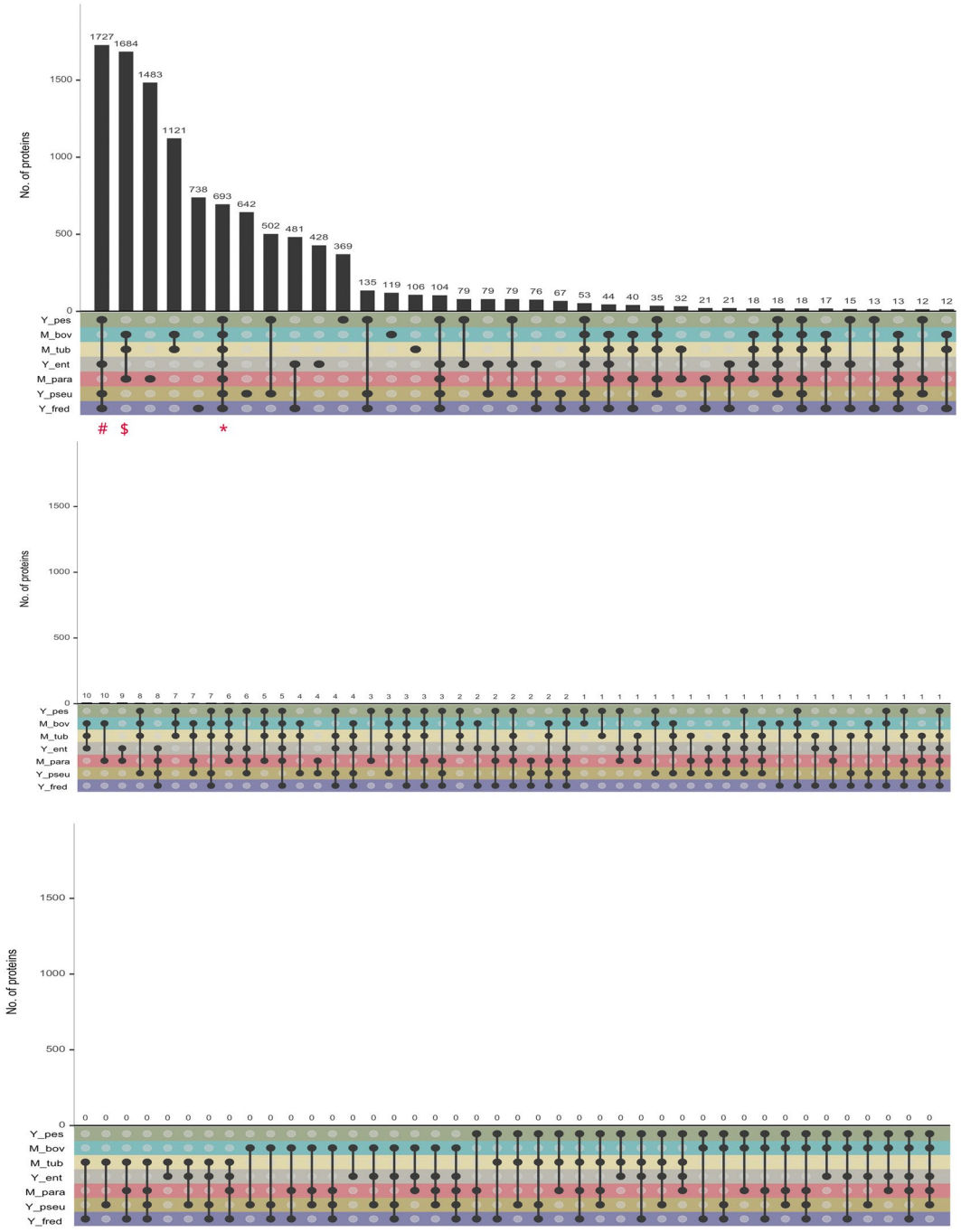

Figure 2. Number of proteins in mutually exclusive protein clusters formed due to different combinations of seven proteomes (*M. tuberculosis, M. bovis, M. paratuberculosis* and non-granuloma forming *Yersinia* species like *Y. pseudotuberculosis, Y. enterocolitica and Y. frederiksenii*). Solid dots show presence of proteins in the corresponding proteome and line between two solid dots shows presence of orthologous proteins in the two proteomes. The vertical bars show the number of shared homologous proteins (orthologous proteins) among the proteomes. Single solid dot represents species unique proteins. The plot was drawn from UpSet plot tool (https://gehlenborglab.shinyapps.io/upsetr/). Panel (a–c) Shows combination of proteomes which shared >10, ≤ 10 and 0 orthologous proteins respectively[85]. '*' represent proteins shared among all seven proteomes, '$' and '#' shows intra-genus conserved proteins of *Mycobacterium* spp. and *Yersinia* spp respectively.

Validation of the identified putative granuloma proteins with the gene expression data. To validate the expression status of the seven Yps granuloma proteins, RNAseq gene expression dataset (ID GSE55292) of *Yersinia pseudotuberculosis* YPIII strain NC_010465.1[26] was obtained from the GEO database[27]. We observed that, of the seven proteins, five proteins were expressed *in-vitro* (above 90% sequence identity).

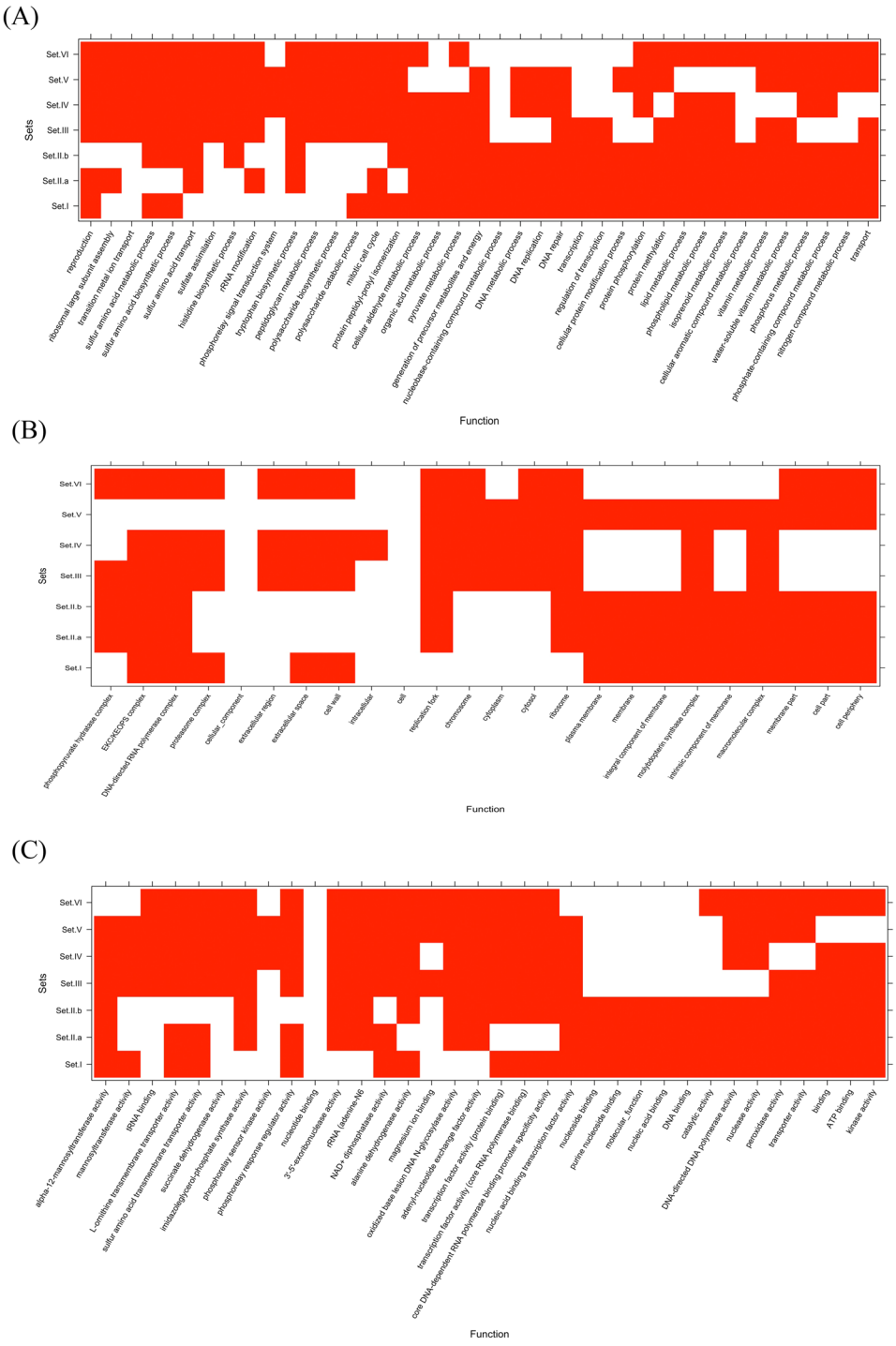

Figure 3. Functional enrichment analysis of different mutually exclusive protein clusters. The enrichment analysis was done for six sets, Set I - inter-genus, Set II - intra-genus: (**A**) *Mycobacterium* spp. and (**B**) *Yersinia* spp., and *Mycobacterium* spp. with Ype Set III; Yen Set IV; Yps Set V and Yfr Set VI. Top 10 enriched GO terms of GO Biological Processes (BP), GO Cellular Components (CC), and GO Molecular Functions (MF) are shown in panels a–c, respectively. Red color shows absence of function where as white shows presence.

However, the expression of remaining two proteins (UniProt id: A0A0U1QT64 and A0A0U1QTE0) could not be ascertained.

Validation of the role of identified Yps granuloma proteins using experimentally identified Mtb granuloma proteins. Earlier studies have proved that Mtb RD1 locus proteins[28], ESAT-6 secretion system proteins[29] and intra-macrophage secreted proteins[30] play an important role in the formation and regulation of granuloma. To ascertain the probable involvement of the seven proposed Yps proteins in Yps granuloma

S. No.	UniProt ID	Protein Name	Domain		STRING Annotation			KEGG pathway
			Position(s)	Description	Definition	Protein	E-value	
1.	A0A0U1QT64	Uncharacterized protein	9–134	AAA - ATPases associated with a variety of cellular activities	—	—	—	—
2.	A0A0U1QTE0	Uncharacterized protein	64–307	Formylglycine-generating sulfatase enzyme	—	—	—	—
3.	A0A0U1QWK3	ABC transporter, ATP-binding protein	8–239, 332–561	ABC transporter	ABC transporter family protein	yadG	2.8e-33	—
4.	A0A0U1R1R0	5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (EC 1.2.1.60)	16–474	Aldedh- Aldehyde dehydrogenase family	5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase	hpaE	5.7e-288	Tyrosine metabolism Microbial metabolism in diverse environments Degradation of aromatic compounds
5.	A0A0U1R1Z2	Uncharacterized protein	30–295	Cellulase - glycoside hydrolase family 5	glycosyl hydrolase 10 family protein	DJ40_3168	6.6e-225	—
6.	A0A0U1R2S7	Transcriptional regulator, TetR family	14–74	tetR- DNA-binding, helix-turn-helix (HTH) domain	bacterial regulatory s, tetR family protein	yxaF	6.6e-115	—
7.	A7FMD4	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	—	—	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	ispH	5.2e-179	Terpenoid backbone biosynthesis Metabolic pathways Biosynthesis of secondary metabolites Biosynthesis of antibiotics

Table 2. Characterization of protein domains present in Yps granuloma proteins, interacting proteins and metabolic pathways as discerned using UniProtKB, STRING and KEGG, respectively.

formation, we constructed a composite protein-protein interaction (PPI) network map of the Mtb RD1 locus proteins, ESAT-6 secretion system proteins and intra-macrophage secreted proteins. The orthologs of the proposed Yps granuloma proteins in the Mtb proteome were mapped on the composite PPI network. Interestingly, all the mapped proteins showed moderate to strong connections with other proteins of the composite PPI network (Fig. 4).

Discussion

Identification of the orthologous protein(s)/gene(s) is a useful method for determining relatedness among different taxonomic groups *viz.* genera, species and strains. In the present study this approach was used to identify Yps proteins, which might be involved in granuloma formation. A comparative *in-silico* analysis of the conserved orthologs of Yen, Ype, Yps, Yfr, Map, Mbov and Mtb proteomes was performed to predict proteins that might be involved in Yps granuloma formation. Initially, we analyzed the genomic and proteomic relatedness among the seven species. At genomic level, analysis was done using pair-wise comparison of ANI values, which is routinely used as a measure of overall similarity between two genome sequences[31]. Results of the present study reiterated the results of previous phylogenetic studies, that the number of shared homologs between different organisms is directly proportional to their evolutionary relatedness[32]. At the proteome level, comparison was done on the basis of pair-wise conservation of orthologs. Our results indicated that the evolutionary relatedness at both genomic (Fig. 1a) and proteomic levels (Fig. 1b) remained the same. A higher number of conserved proteins in Mbov and Mtb reiterated the close relationship between the two species of Mtb-complex. Our results were also in-line with the 16S rRNA gene sequences based phylogenetic studies on Mbov and Mtb[33]. Also, the number of proteins shared by Mbov and Mtb, with Map (atypical mycobacteria) was less than the proteins shared between Mbov and Mtb. This was similar to the earlier reports based on 16S rRNA based phylogenetic study[34]. During pair-wise comparison of conserved proteins among the four species of *Yersinia*, Ype and Yps shared maximum number of proteins. Similar to Mtb and Mbov, a large number of conserved proteins in Ype and Yps can be attributed to their phylogenetic proximity[3]. An earlier study has also reported that Ype has evolved from Yps[35], which might be a probable reason behind their closeness. On the other hand Yfr, which is an opportunistic pathogen[36], shared more proteins with Yen than with Ype or Yps. Earlier phylogenetic studies using multi locus sequence typing have also shown that Ype and Yps belonged to the same cluster and, Yen and Yfr clusters were close to each other[32]. This further confirms their relatedness at the proteome level. We also observed an interesting pattern of shared orthologs of Map with Yen and Yfr. The number of shared orthologs of Map with Yen and Yfr was more, than with Ype and Yps. This might have happened because Map, Yen and Yfr cause gastrointestinal infections and hence occupy the same niche which might have resulted in horizontal transfer of genes among them[37].

The enrichment of sulfur-containing amino acids and metabolic proteins in intra-genus protein cluster of *Mycobacterium* spp. indicates their importance in survival of *Mycobacterium* spp. Earlier reports also suggested that sulfur containing amino acids help Mtb in sustaining the oxidative stress, nutrient starvation and, in dormancy adaptation[38,39]. Due to presence of sulfur-containing amino acid synthesis pathway proteins exclusively in *Mycobacterium* spp., this pathway was also proposed as a potential target candidate for anti-TB therapy[40]. Analysis of intra-genus protein cluster containing proteins of *Yersinia* spp. revealed conservation of proteins involved in reproduction and mitotic cell cycle. This showed that except for a few, functions of most of the proteins were

S. No.	UniProt ID	Interacting proteins	KEGG pathway analysis of the interacting proteins	Interacting proteins/ pathways proposed as potential drug-targets
1.	A0A0U1QT64	Uncharacterized protein	—	—
2.	A0A0U1QTE0	Uncharacterized protein	—	—
3.	A0A0U1QWK3	yadH queF can icaB lepB gstB ybhS	Folate biosynthesis, Metabolic pathways (queF), Nitrogen metabolism (can), Protein export (lepB), Glutathione metabolism (gstB)	Glutathione metabolism[40,86], ABC transporters[87]
4.	A0A0U1R1R0	hpcD hpaD hpcE_1 hpcE_2 hpaH hpaI hpaR hpaX hpaB nifJ	Tyrosine metabolism, Microbial metabolism in diverse environments, Degradation of aromatic compounds (hpcD, hpaD, hpcE_1, hpcE_2, hpaH, hpaI, hpaB) Glycolysis/Gluconeogenesis, Citrate cycle (TCA cycle), Pyruvate metabolism, Butanoate metabolism, Metabolic pathways, Biosynthesis of secondary metabolites, Microbial metabolism in diverse environments, Biosynthesis of antibiotics, Carbon metabolism (nifJ)	D-Alanine metabolism[88], Metabolic pathways[89], tyrosine metabolism, pyruvate metabolism, Butanoate metabolism, Glycolysis/Gluconeogenesis, Citrate cycle (TCA cycle)[87]
5.	A0A0U1R1Z2	nhaR melB2	–	–
6.	A0A0U1R2S7	DJ40_975 rpsC tatD tesB purC tmk icIR	Ribosome (rpsC), Biosynthesis of unsaturated fatty acids (tesB), Purine metabolism, Metabolic pathways, Biosynthesis of secondary metabolites, Biosynthesis of antibiotics (purC) Pyrimidine metabolism, Metabolic pathways (tmk)	Fatty acid biosynthesis, Purine metabolism, pyrimidine metabolism[87]
7.	A7FMD4	ispG ispA lspA rpsA cmk ispF ispD dxs fkpB dxr	Terpenoid backbone biosynthesis, Metabolic pathways, Biosynthesis of secondary metabolites, Biosynthesis of antibiotics (ispG, ispF, ispD, dxs, dxr), Protein export (lspA), Ribosome (rpsA), Pyrimidine metabolism, Metabolic pathways (cmk), Thiamine metabolism (dxs),	Thiamine metabolism[90] Terpenoid backbone biosynthesis[63,64]

Table 3. Information about the interacting proteins of the potential Yps granuloma proteins, the various metabolic pathways in which they are involved and, the proposed drug targets.

conserved in both inter- and intra-genus orthologous protein clusters. This also indicates that though sequences of core proteins might have diverged, functionality is retained during evolution. Also, proteins which were unique to a single species and whose orthologs were absent in other species were considered as unique proteins. No significant functional enrichment in species-specific unique proteins was observed because they might belong to different biological pathways, involved in diverse molecular function and reside in different cellular component.

The main objective of this study was to identify proteins that might help in Yps granuloma formation. Therefore, the function of the seven Yps proteins, which were common in Yps and the three *Mycobacterium* spp. were critically analyzed to investigate their probable role in Yps granuloma formation. The protein with the UniProt id A0A0U1QT64 was an uncharacterized protein with an ATPase domain. Since ATPase domains are capable of unfolding the protein substrates, hence proteins harboring ATPase domains are known to be involved in protein degradation. ATPase domains are also essential for intracellular protein degradation because macromolecular assemblies, for *e.g.* proteasome machinery, confine their proteolytic and protease activity in an inner nano-compartment which is accessible only to the unfolded protein substrates[41]. This suggests that proteolytic machinery might be functionally linked to unfolding machinery (AAA – ATPases domain proteins) and are preserved throughout evolution[42,43]. In Mtb, proteostasis network provides protection from different stresses and host immunity. The machinery used for this comprises a complex network of chaperones, proteases, and a eukaryotic-like proteasome (functionally linked AAA – ATPases domain protein family) which helps in evading the host immunity by maintaining the integrity of the mycobacterial proteome[44]. Besides this, AAA – ATPases domain protein family also play a significant role in recognition of ESAT-6 secretion system (ESX-1) secreted virulence factors[45], which is a type VII secretion system of Mtb and is capable to form pores and rupture phagosomes[46]. This leads to cell toxicity, necrosis and ultimately cell death[47]. On the basis of the functional role of constituent domains in different organism, it can be inferred that this protein might play a probable role in formation of Yps granuloma.

The protein with the UniProt id A0A0U1QTE0 is a functionally uncharacterized protein with a formylglycine-generating sulfatase enzyme domain. Such proteins are reported to be involved in ergotheonine (EGT) synthesis which is a histidine-derived thiol[48]. It reportedly enabled the pathogens in withstanding the host hostile environment during initial phase of infection[49]. EGT-containing proteins are present only in prokaryotes, while plants and animals (including humans) do not produce EGT[50]. Also, macrophages with EGT show an increased cytokine production that enhances Th17 polarization of $CD4^+$ T cells. Therefore, it acts as TLR agonist

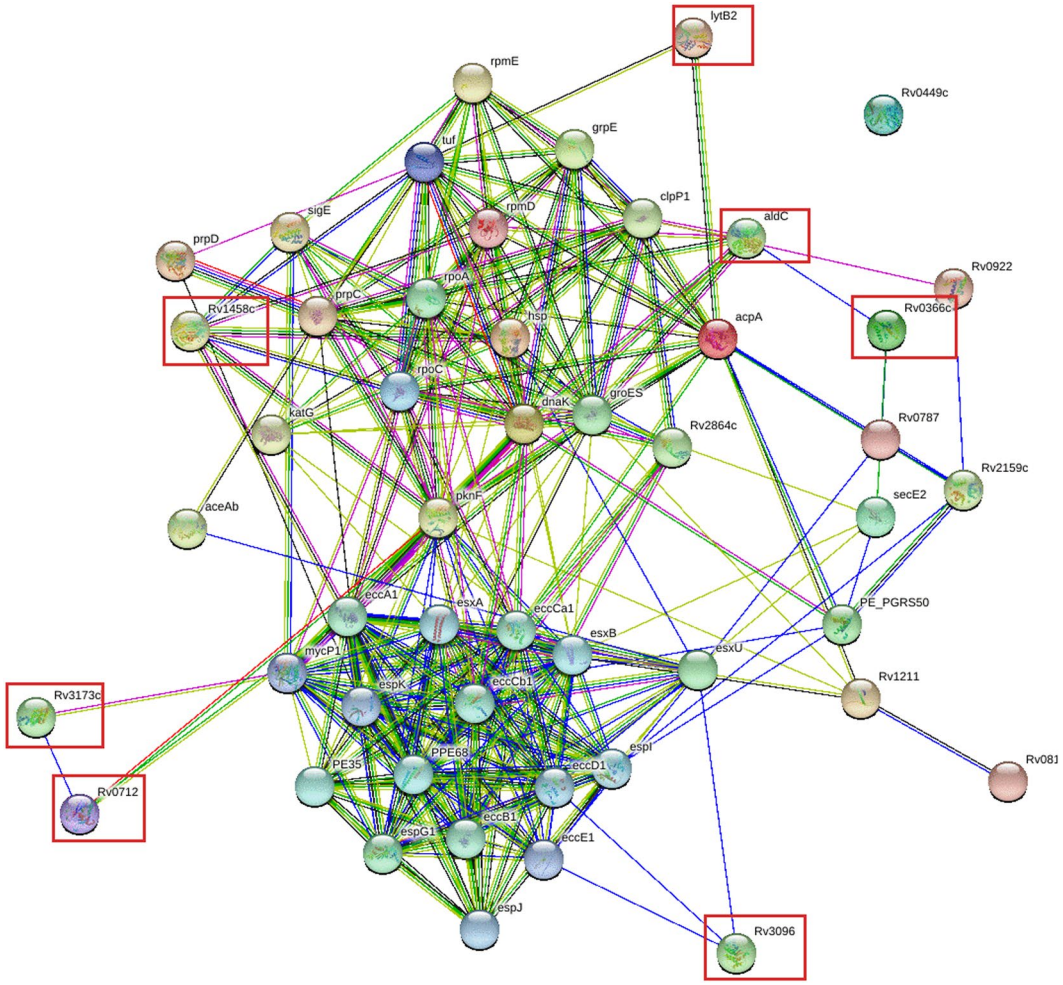

Figure 4. A composite protein-protein interaction network map of Mtb RD1 locus proteins, ESAT-6 secretion system proteins and intra-macrophage secreted proteins, constructed using STRING database.The Mtb orthologs of the seven proposed Yps proteins (Rv0366c, Rv0712, Rv1458c, Rv2858c (aldc), Rv3096, Rv3173c, and Rv3382c (lytB2)) are marked in red colored rectangular boxes. The colour and thickness of edges (lines connecting two proteins or nodes) indicates type and confidence of interaction, respectively. The color coding of edges are as follows: Red line - indicates the presence of fusion evidence; Green line - neighborhood evidence; Blue line - cooccurrence evidence; Purple line - experimental evidence; Yellow line - textmining evidence; Light blue line - database evidence; Black line - coexpression evidence.

(ligand)[51] and show immune enhancing property, which causes more cells to come into contact. This indicates that EgtB might be involved in attracting more cells to the site of granuloma formation and thereby help in the process of granuloma formation.

The protein with UniProt id A0A0U1QWK3 was a protein of ABC transporter family (yadG), an integral membrane protein responsible for active transport of ligands across biological membranes[52]. This ABC transporter, ATP–binding protein-encoding gene is present as a pseudogene in Ype (a closely related species of Yps) but is active in Yps[53]. These transporters couple ATP hydrolysis for the uptake and efflux of solutes across the membrane in both bacterial and eukaryotic cells. These are considered as important bacterial virulence factors due to their role in nutrient uptake, secretion of toxins and antimicrobial agents in the host[54]. In *Yersinia* and *Mycobacterium* iron uptake is important for infection and survival in host macrophages[55,56]. Also, ABC transporter system of *Mycobacterium* is similar to the *Yersinia* YbtPQ system[56]. Since, ATP binding cassette transporter proteins are also enriched in tubercular granuloma[30], it indicates that these proteins might also play an important role in Yps granuloma

The protein with the UniProt id A0A0U1R1R0 is an enzyme, 5-carboxymethyl-2-hydroxymuconate semi aldehyde dehydrogenase (hpaE). The gene encoding this protein is also annotated as a pseudogene in Ype but is actively expressed in Yps[53]. Aldehydes are highly reactive chemical moiety that triggers oxidative stress in both prokaryotes and eukaryotes, that makes them toxic for cells. Enzymes with aldehyde dehydrogenase domain (ALDHs) play an important role in metabolism of both endogenous and exogenous aldehydes. Earlier studies have shown an increased production of ALDHs to cope up with environmental and chemical stress in bacteria[57]. Both *Yersinia* and *Mycobacterium* are intracellular pathogens; hence these bacteria have to combat oxidative stress

inside the cell. Previous reports also indicated oxido-reductase enzymes were present in tubercular granuloma[30]. This suggests that this protein might also play an important role in Yps granuloma.

The protein with the UniProt id A0A0U1R1Z2 belongs to the glycosyl hydrolase 10 family of proteins (DJ40_3168). Proteins containing domains of glycoside hydrolase family are present in cellulases (glycoside hydrolases). These proteins play a crucial role in degrading plant cellulose and bacterial cell walls[58]. It has been reported that for transforming from disease causing active state to persistent stage, Mtb dissolves the polysaccharide biofilm in the mammalian host[59]. This also indicates the role of glycoside hydrolases in Mtb virulence. Linkage of Mtb persistence to biofilm also indicates that this protein might also have an important role in Yps granuloma.

The protein with the UniProt id A0A0U1R2S7 is a bacterial regulatory, tetR family protein (yxaF). This gene is also present as pseudogene in Ype but is active in Yps[53]. These proteins contain a TetR DNA-binding, helix-turn-helix (HTH) domain. Proteins containing HTH domains function as DNA-binding transcriptional regulators. These proteins regulate gene expression by binding to the major grove of DNA. These proteins regulate the expression of mycobacterial membrane protein family transporters which are critical for exporting fatty acids and lipidic elements important for mycobacterial virulence[60]. Also, a high rate of lipid transport and metabolism helps in better survival in diverse environments. In Mtb, TetR proteins are found to induce necrosis in lungs[61]. The above-mentioned function of orthologous proteins in Mtb indicates that A0A0U1R2S7 might also play a significant role in necrosis of Yps granuloma.

The protein with the UniProt id A7FMD4 was an enzyme 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (ispH), which is required in DOXP/MEK pathway. The DOXP pathway plays an important role in the pathogenic potential of mycobacterial species. Disruption of DOXP pathway in Mtb hinders its ability to prevent acidification of the phagosome. This results in a decreased potential in intracellular survival[62]. These proteins are present only in pathogenic bacteria, but not in human[63,64]. In *M. avium* subsp. *paratuberculosis, gcpE* mutants were reported to be less efficient in tissue colonization in mice or calves[65,66], which further confirms the importance of this pathway in virulence. Thus, enzymes of the DOXP pathway have also been proposed as a potential drug target against Mtb[67].

Yersinia and mycobacteria formed two distinct branches on the cladogram drawn on the basis of ANI and PCOP. Despite the differences at genomic and proteomic levels, there are similarities between the granuloma formed by Yps and *Mycobacterium* spp. For example, Yps granuloma is characterized by a central necrosis (caseation) and micro abscess which is also common in tubercular granuloma[24]. Interestingly, functional enrichment revealed that Yps proteins were involved in lipid, phospholipid, isoprenoid, aldehyde and pyruvate metabolism which is similar to the mechanism of granuloma formation in Mtb[30]. Also, lipid metabolism is associated with caseation of granuloma and dissemination of bacteria in the neighboring tissues and organs and, increases the infectivity of bacteria[68]. Despite some similarities, there are also certain subtle differences between the granuloma formed by Yps and *Mycobacterium* spp. For example, the chaperones are well known virulence factors in the formation of tubercular granuloma and are required for bacterial virulence, detoxification and adaptation in Mtb[30,69]. But chaperone proteins were absent in Yps granuloma. Interestingly, protein-protein interaction and metabolic pathway mapping of interacting proteins of the seven common proteins of Yps and *Mycobacterium* spp. revealed that most of these interacting proteins have been proposed as potential drug targets in Mtb (Table 3). Hence, the seven Yps proteins proposed in the present study might be explored as useful drug targets against Yps. Also, an attempt was made to discern if the seven Yps proteins interacted with each other. However, no interaction was observed among these proteins. This might have happened because each of these proteins was involved in a different biological pathway. This also indicates that like Mtb, diverse mechanisms might underlie granuloma formation in Yps[70].

A comparison of the expression patterns of *in vivo* and *in vitro* derived transcriptome analysis revealed that the Yps early phase infection expression pattern was similar to the *in vitro* expression pattern at 37 °C. Also, the expression pattern of persistent Yps bacteria was approximately similar to that of bacteria grown *in vitro* at 26 °C[26]. Hence, to validate our findings regarding expression of seven proteins we used the RNAseq expression data derived from Yps during *in vitro* growth at 26 °C and 37 °C (GSE55292)[26]. We found that out of the seven proteins, five proteins were also expressed during *in vitro* growth. However, the expression of the remaining two proteins (UniProt ids A0A0U1QT64 and A0A0U1QTE0) could not be ascertained. This might have probably happened because we used Yps strain IP 31758 for our genomic and proteomic analysis, while the transcriptomic data used in our study was based on Yps strain YPIII, which is a plasmid curated strain and, the two proteins, *viz*. UniProt id: A0A0U1QT64 and A0A0U1QTE0 were present on the plasmid. Also, we failed to find the expression in RNAseq data generated by *in vivo* derived total RNA samples. This might be due to a very low abundance of Yps transcripts, which ultimately leads to low coverage of Yps ORFs[26].

Granuloma formation is primarily a host-defence mechanism which restricts the spread of bacteria. However, some pathogens use it as a protective shell to survive till the advent of favourable conditions. The pathogen resumes its activity and starts multiplication when the conditions become favourable. The best known example of a well-studied pathogen and granuloma is Mtb. Experimental studies in Mtb suggested that RD1 locus proteins[28], ESAT-6 secretion system proteins[29] and proteins of intra-macrophage secretome[30] were mainly involved in Mtb granuloma. Several studies indicated the importance of RD1 locus and ESAT-6 secretion system in Mtb granuloma[71–73]. Mtb strains devoid of RD1 proteins failed to induce Mtb granuloma[74]. Thus, it can be inferred that proteins of the RD1 region, ESAT secretion system and intra-macrophage secretome are important for Mtb granuloma formation. Even after an extensive literature survey we could not find study regarding the mechanistic details of granuloma formation in Yps. Hence, we constructed a composite PPI network of proteins encompassing the proteins of the Mtb RD1 locus, ESAT-6 secretion system and intra-macrophage secretome. The orthologs of the proposed Yps granuloma proteins present in the Mtb proteome were mapped on the composite PPI network map. It was interesting to note that all the Mtb proteins mapped on the composite PPI network showed moderate

to strong connections with other proteins of the network. This, suggested that the seven proteins identified in this study might be important for Yps granuloma formation.

To summarize, using a comparative *in silico* proteome analysis of Yps with Map, Mbov and Mtb we identified seven proteins that were absent in Yen, Yfr and Ype. The *in-silico* functional characterization and validation with experimental Mtb granuloma proteins further strengthen our findings that the proposed seven proteins might play some role in Yps granuloma. However, additional experiments involving knocking out of each of these seven proteins are required to confirm their role in Yps granuloma. Additionally, the seven proteins proposed in the present study might not only be the proteins responsible for Yps granuloma and, despite adoption of stringent parameters, many potential granuloma proteins might have been missed. We understand that a detailed functional characterization of Yps proteins is required to unravel the complex mechanisms underlying Yps granuloma formation. Nevertheless, our study provides some useful insights and can serve as a basis for further studies on Yps granuloma.

Materials and Methods

Genomes and proteomes used for analysis. The proteome and genome data sets used in the present study, were obtained from UniProtKB (release 2017_09)[75] (Table 1) and NCBI (http://www.ncbi.nlm.nih.gov) respectively. The accession numbers of proteomes and genomes of the seven microbes used in the present work are as follows: Mtb (UniProt ID: UP000001584; NCBI ID: NC_000962.3), Mbov (UniProt ID: UP000001419; NCBI ID: AP010918.1), Map (UniProt ID: UP000000580; NCBI ID: NC_002944.2), Ype (UniProt ID: UP000000815; NCBI ID: NC_003143.1), Yen (UniProt ID: UP000000642; NCBI ID: AM286415.1), Yps (UniProt ID:UP000002412; NCBI ID: NZ_CP008943.1), and Yfr (UniProt ID: UP000005500; NCBI ID: NZ_CP009364.1).

Determination of relatedness and distinctiveness among different species. To determine the genomic relatedness among the seven bacterial species, Average Nucleotide Identity (ANI) was calculated using OrthoANI[76]. To estimate the evolutionary distance among proteomes of different species, the percentage of conserved proteins (PCOP) was calculated[77]. The values of ANI and PCOP were used to construct the Neighbor-Joining (NJ) tree[78] using MEGA[79].

Identification of orthologous proteins. All the possible combinations of the seven proteomes were made and, orthologous proteins in each group were identified. To find orthologous proteins we used InParanoid (version 4.1) at default parameters. InParanoid performs reciprocal BLAST and labels protein sequences based on sequence similarity as orthologs[80]. For each ortholog, InParanoid provides bit score in the range of 0.5–1. In this study we considered two proteins as orthologs, if the InParanoid score was $\geq$0.8.

Clustering of orthologous sequences. On the basis of the number of proteomes in which a set of orthologous proteins was present, orthologous sequences were categorized into mutually exclusive clusters. Proteins of each cluster represented a specific chunk of proteins that was not shared by other clusters. For example proteins in inter-genus ortholog cluster contained proteins that were present in all the seven proteomes. Similarly, intra-genus ortholog cluster contained proteins, which were present only in *Yersinia* or *Mycobacterium*.

Functional enrichment of proteins. Functional annotation of each protein cluster was done by assigning them gene ontology (GO) terms. The GO terms were retrieved from the Gene Ontology Consortium[81]. The functional enrichment of orthologous protein clusters was done using topGO tools (v2.24.0) of Bioconductor package[82]. In the present work, high-level view of GO terms, namely GO-slim terms, was used to determine the enriched functions. These terms were extracted from the GO annotation dataset by GO Slim Mapper OWL Tool (https://github.com/owlcollab/owltools.git). During enrichment, all the proteins present in the seven proteomes were divided into seven broad categories (or "test datasets") and a unique background was used during enrichment of each test-dataset. The enrichment was done on cluster of orthologous sequences (or "test datasets") and a unique background was used during enrichment of each test-dataset.

Set I (inter-genus conserved proteins). Inter genus conserved set included proteins that were conserved across all the seven proteomes. During functional enrichment of this category of proteins, combined GO-slim terms of all seven complete proteomes was used as background.

Set II (intra-genus conserved proteins). Intra genus conserved set included the proteins present in all species of genus *Mycobacterium* or *Yersinia*. During functional enrichment of this group of proteins, collective GO-slim terms of proteome of respective genus was used as background.

Set III (Conserved in Ype and Mycobacterium spp.). This contains the proteins which were common in Ype and the three *Mycobacterium spp*., The functional enrichment of proteins of Set III proteins were determined against using all four *Yersinia* spp. as background.

Set IV (Conserved in Yen and Mycobacterium spp.). The functional enrichment of proteins of Yen whose orthologs were present in all *Mycobacterium* spp. (as test-dataset) were determined against all the four *Yersinia* spp. (as background).

Set V (Yfr with Mycobacterium spp.). enriched function of Yfr proteins, whose orthologs were present in *Mycobacterium* spp. (as test-dataset) were determined against all the four *Yersinia* spp. (as background).

Set VI (Yps with Mycobacterium spp.). enriched function of Yps proteins which showed orthology with the proteins of *Mycobacterium* spp. (as test-dataset) were determined against all the four *Yersinia* spp. (as background).

Set VII (with-in the species). This set includes proteins unique to a particular species. To find functional enrichment in these proteins, GO-slim terms retrieved from the complete proteome of the same species were used as the background.

Characterization of Yps proteins involved in granuloma formation. Since, the aim of the current study was identification of Yps proteins involved in granuloma formation, hence only those proteins of Yps whose orthologs were present in the MTB- complex members, were functionally characterized. For functional annotation a three-step process was followed: (a) domain information of each protein was collected from UniProtKB; (b) the protein and their interaction partners were identified using STRING database (https://string-db.org/)[83] and characterized; (c) the metabolic pathways in which the interacting protein partners were involved were identified using the KEGG database[84] and, (d) information on whether the interacting proteins and/or pathways have been used as drug targets was retrieved from the published literature.

Mapping of Yps predicted proteins on composite PPI interaction network of experimentally identified Mtb granuloma proteins. A composite interaction network of RD1 locus proteins[28], ESAT-6 secretion system proteins[29] and intra-macrophagic secretome of Mtb[30] was created using STRING database, at confidence score 0.150. The Mtb orthologs of the proposed seven Yps granuloma proteins were mapped on this interaction network.

Received: 10 September 2019; Accepted: 3 February 2020;
Published online: 20 February 2020

References

1. Long, C. *et al*. Yersinia pseudotuberculosis and Y. enterocolitica infections, FoodNet, 1996–2007. *Emerg. Infect. Dis.* **16**, 566–567 (2010).
2. Pujol, C. & Bliska, J. B. The ability to replicate in macrophages is conserved between Yersinia pestis and Yersinia pseudotuberculosis. *Infect. Immun.* **71**, 5892–5899 (2003).
3. McNally, A., Thomson, N. R., Reuter, S. & Wren, B. W. Add, stir and reduce: Yersinia spp. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.* **14**, 177–190 (2016).
4. Reuter, S. *et al*. Parallel independent evolution of pathogenicity within the genus Yersinia. *Proc. Natl Acad. Sci. USA* **111**, 6768–6773 (2014).
5. Westermark, L., Fahlgren, A. & Fallman, M. Yersinia pseudotuberculosis efficiently escapes polymorphonuclear neutrophils during early infection. *Infect. Immun.* **82**, 1181–1191 (2014).
6. Thoerner, P. *et al*. PCR detection of virulence genes in Yersinia enterocolitica and Yersinia pseudotuberculosis and investigation of virulence gene distribution. *Appl. Environ. microbiology* **69**, 1810–1816 (2003).
7. Wang, X. *et al*. Distribution of pathogenic Yersinia enterocolitica in China. *Eur. J. Clin. microbiology Infect. diseases: Off. Publ. Eur. Soc. Clin. Microbiology* **28**, 1237–1244 (2009).
8. Wang, X. *et al*. Pathogenic strains of Yersinia enterocolitica isolated from domestic dogs (Canis familiaris) belonging to farmers are of the same subtype as pathogenic Y. enterocolitica strains isolated from humans and may be a source of human infection in Jiangsu Province, China. *J. Clin. Microbiol.* **48**, 1604–1610 (2010).
9. Liang, J. *et al*. Prevalence of Yersinia enterocolitica in pigs slaughtered in Chinese abattoirs. *Appl. Environ. microbiology* **78**, 2949–2956 (2012).
10. Galan, J. E. & Wolf-Watz, H. Protein delivery into eukaryotic cells by type III secretion machines. *Nat.* **444**, 567–573 (2006).
11. Durand, E. A., Maldonado-Arocho, F. J., Castillo, C., Walsh, R. L. & Mecsas, J. The presence of professional phagocytes dictates the number of host cells targeted for Yop translocation during infection. *Cell Microbiol.* **12**, 1064–1082 (2010).
12. Asano, S. Granulomatous lymphadenitis. *J. Clin. Exp. hematopathology: JCEH* **52**, 1–16 (2012).
13. Riedel, D. D. & Kaufmann, S. H. Chemokine secretion by human polymorphonuclear granulocytes after stimulation with Mycobacterium tuberculosis and lipoarabinomannan. *Infect. Immun.* **65**, 4620–4623 (1997).
14. Silva Miranda, M., Breiman, A., Allain, S., Deknuydt, F. & Altare, F. The tuberculous granuloma: an unsuccessful host defence mechanism providing a safety shelter for the bacteria? *Clin. developmental immunology* **2012**, 139127 (2012).
15. Zhang, L., English, D. & Andersen, B. R. Activation of human neutrophils by Mycobacterium tuberculosis-derived sulfolipid-1. *J. Immunol.* **146**, 2730–2736 (1991).
16. Almadi, M. A. *et al*. New insights into gastrointestinal and hepatic granulomatous disorders. *Nat. Rev. Gastroenterol. Hepatol.* **8**, 455–466 (2011).
17. Brown, I. & Kumarasinghe, M. P. Granulomas in the gastrointestinal tract: deciphering the Pandora's box. *Virchows Arch.* **472**, 3–14 (2018).
18. Autenrieth, I. B., Hantschmann, P., Heymer, B. & Heesemann, J. Immunohistological characterization of the cellular immune response against Yersinia enterocolitica in mice: evidence for the involvement of T lymphocytes. *Immunobiology* **187**, 1–16 (1993).
19. Yao, T., Mecsas, J., Healy, J. I., Falkow, S. & Chien, Y. Suppression of T and B lymphocyte activation by a Yersinia pseudotuberculosis virulence factor, yopH. *J. Exp. Med.* **190**, 1343–1350 (1999).
20. Ye, Z., Lin, Y., Cao, Q., He, Y. & Xue, L. Granulomas as the Most Useful Histopathological Feature in Distinguishing between Crohn's Disease and Intestinal Tuberculosis in Endoscopic Biopsy Specimens. *Med.* **94**, e2157 (2015).
21. Bradford, W. D., Noce, P. S. & Gutman, L. T. Pathologic features of enteric infection with Yersinia enterocolitica. *Arch. Pathol.* **98**, 17–22 (1974).
22. Gleason, T. H. & Patterson, S. D. The pathology of Yersinia enterocolitica ileocolitis. *Am. J. Surg. Pathol.* **6**, 347–355 (1982).
23. Lamps, L. W. *et al*. The role of Yersinia enterocolitica and Yersinia pseudotuberculosis in granulomatous appendicitis: a histologic and molecular study. *Am. J. Surg. Pathol.* **25**, 508–515 (2001).
24. El-Maraghi, N. R. & Mair, N. S. The histopathology of enteric infection with Yersinia pseudotuberculosis. *Am. J. Clin. Pathol.* **71**, 631–639 (1979).
25. Huang, J. C. & Appelman, H. D. Another look at chronic appendicitis resembling Crohn's disease. *Mod. Pathol.* **9**, 975–981 (1996).
26. Avican, K. *et al*. Reprogramming of Yersinia from virulent to persistent mode revealed by complex *in vivo* RNA-seq analysis. *PLoS Pathog.* **11**, e1004600 (2015).
27. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. *Methods Mol. Biol.* **1418**, 93–110 (2016).

28. Soman, S. *et al*. Presence of region of difference 1 among clinical isolates of Mycobacterium tuberculosis from India. *J. Clin. Microbiol.* **45**, 3480–3481 (2007).
29. Gey Van Pittius, N. C. *et al*. The ESAT-6 gene cluster of Mycobacterium tuberculosis and other high G + C Gram-positive bacteria. *Genome Biol* **2**, RESEARCH0044 (2001).
30. Chande, A. G. *et al*. Selective enrichment of mycobacterial proteins from infected host macrophages. *Sci. Rep.* **5**, 13430 (2015).
31. Yoon, S. H., Ha, S. M., Lim, J., Kwon, S. & Chun, J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van. Leeuwenhoek* **110**, 1281–1286 (2017).
32. Kotetishvili, M. *et al*. Multilocus sequence typing for studying genetic relationships among Yersinia species. *J. Clin. Microbiol.* **43**, 2674–2684 (2005).
33. Brosch, R., Pym, A. S., Gordon, S. V. & Cole, S. T. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol.* **9**, 452–458 (2001).
34. Zakham, F., Aouane, O., Ussery, D., Benjouad, A. & Ennaji, M. M. Computational genomics-proteomics and Phylogeny analysis of twenty one mycobacterial genomes (Tuberculosis & non Tuberculosis strains). *Microb. Inf. Exp.* **2**, 7 (2012).
35. Achtman, M. *et al*. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proc. Natl Acad. Sci. USA* **96**, 14043–14048 (1999).
36. Ursing, J. & Aleksic, S. Yersinia frederiksenii, a genotypically heterogeneous species with few differential characteristics. *Contributions microbiology immunology* **13**, 112–116 (1995).
37. Fuchsman, C. A., Collins, R. E., Rocap, G. & Brazelton, W. J. Effect of the environment on horizontal gene transfer between bacteria and archaea. *PeerJ* **5**, e3865 (2017).
38. Paritala, H. & Carroll, K. S. New targets and inhibitors of mycobacterial sulfur metabolism. *Infect. Disord. Drug. Targets* **13**, 85–115 (2013).
39. Gengenbacher, M. & Kaufmann, S. H. Mycobacterium tuberculosis: success through dormancy. *FEMS Microbiol. Rev.* **36**, 514–532 (2012).
40. Bhave, D. P., Muse, W. B. III & Carroll, K. S. Drug targets in mycobacterial sulfur metabolism. *Infect. Disord. drug. targets* **7**, 140–158 (2007).
41. Bar-Nun, S. & Glickman, M. H. Proteasomal AAA-ATPases: structure and function. *Biochim. Biophys. Acta* **1823**, 67–82 (2012).
42. Pickart, C. M. & Cohen, R. E. Proteasomes and their kin: proteases in the machine age. *Nat. Rev. Mol. Cell Biol.* **5**, 177–187 (2004).
43. Sauer, R. T. & Baker, T. A. AAA+ proteases: ATP-fueled machines of protein destruction. *Annu. Rev. Biochem.* **80**, 587–612 (2011).
44. Lupoli, T. J., Vaubourgeix, J., Burns-Huang, K. & Gold, B. Targeting the Proteostasis Network for Mycobacterial Drug Discovery. *ACS Infect. Dis.* **4**, 478–498 (2018).
45. Simeone, R., Bottai, D. & Brosch, R. ESX/type VII secretion systems and their role in host-pathogen interaction. *Curr. Opin. Microbiol.* **12**, 4–10 (2009).
46. Champion, P. A., Champion, M. M., Manzanillo, P. & Cox, J. S. ESX-1 secreted virulence factors are recognized by multiple cytosolic AAA ATPases in pathogenic mycobacteria. *Mol. Microbiol.* **73**, 950–962 (2009).
47. Simeone, R. *et al*. Phagosomal rupture by Mycobacterium tuberculosis results in toxicity and host cell death. *PLoS Pathog.* **8**, e1002507 (2012).
48. Seebeck, F. P. *In vitro* reconstitution of Mycobacterial ergothioneine biosynthesis. *J. Am. Chem. Soc.* **132**, 6632–6633 (2010).
49. Cumming, B. M., Chinta, K. C., Reddy, V. P. & Steyn, A. J. C. Role of Ergothioneine in Microbial Physiology and Pathogenesis. *Antioxid. Redox Signal.* **28**, 431–444 (2018).
50. Ey, J., Schomig, E. & Taubert, D. Dietary sources and antioxidant effects of ergothioneine. *J. Agric. Food Chem.* **55**, 6466–6474 (2007).
51. Yoshida, S. *et al*. The Anti-Oxidant Ergothioneine Augments the Immunomodulatory Function of TLR Agonists by Direct Action on Macrophages. *PLoS One* **12**, e0169360 (2017).
52. Linton, K. J. Structure and function of ABC transporters. *Physiol.* **22**, 122–130 (2007).
53. Chain, P. S. *et al*. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. *Proc. Natl Acad. Sci. USA* **101**, 13826–13831 (2004).
54. Davidson, A. L. & Chen, J. ATP-binding cassette transporters in bacteria. *Annu. Rev. Biochem.* **73**, 241–268 (2004).
55. Fetherston, J. D., Bertolino, V. J. & Perry, R. D. YbtP and YbtQ: two ABC transporters required for iron uptake in Yersinia pestis. *Mol. Microbiol.* **32**, 289–299 (1999).
56. Rodriguez, G. M. & Smith, I. Identification of an ABC transporter required for iron acquisition and virulence in Mycobacterium tuberculosis. *J. Bacteriol.* **188**, 424–430 (2006).
57. Singh, S. *et al*. Aldehyde dehydrogenases in cellular responses to oxidative/electrophilic stress. *Free. Radic. Biol. Med.* **56**, 89–101 (2013).
58. Berlemont, R. & Martiny, A. C. Phylogenetic distribution of potential cellulases in bacteria. *Appl. Env. Microbiol.* **79**, 1545–1554 (2013).
59. Varrot, A. *et al*. Mycobacterium tuberculosis strains possess functional cellulases. *J. Biol. Chem.* **280**, 20181–20184 (2005).
60. Chou, T. H. *et al*. Crystal structure of the Mycobacterium tuberculosis transcriptional regulator Rv0302. *Protein Sci.* **24**, 1942–1955 (2015).
61. Repasy, T. *et al*. Bacillary replication and macrophage necrosis are determinants of neutrophil recruitment in tuberculosis. *Microbes Infect.* **17**, 564–574 (2015).
62. Pethe, K. *et al*. Isolation of Mycobacterium tuberculosis mutants defective in the arrest of phagosome maturation. *Proc. Natl Acad. Sci. USA* **101**, 13642–13647 (2004).
63. Obiol-Pardo, C., Rubio-Martinez, J. & Imperial, S. The methylerythritol phosphate (MEP) pathway for isoprenoid biosynthesis as a target for the development of new drugs against tuberculosis. *Curr. medicinal Chem.* **18**, 1325–1338 (2011).
64. Testa, C. A. & Brown, M. J. The methylerythritol phosphate pathway and its significance as a novel drug target. *Curr. Pharm. Biotechnol.* **4**, 248–259 (2003).
65. Shin, S. J., Wu, C. W., Steinberg, H. & Talaat, A. M. Identification of novel virulence determinants in Mycobacterium paratuberculosis by screening a library of insertional mutants. *Infect. Immun.* **74**, 3825–3833 (2006).
66. Wu, C. W. *et al*. Invasion and persistence of Mycobacterium avium subsp. paratuberculosis during early stages of Johne's disease in calves. *Infect. Immun.* **75**, 2110–2119 (2007).
67. Hunter, W. N. The non-mevalonate pathway of isoprenoid precursor biosynthesis. *J. Biol. Chem.* **282**, 21573–21577 (2007).
68. Kim, M. J. *et al*. Caseation of human tuberculosis granulomas correlates with elevated host lipid metabolism. *EMBO Mol. Med.* **2**, 258–274 (2010).
69. Cehovin, A. *et al*. Comparison of the moonlighting actions of the two highly homologous chaperonin 60 proteins of Mycobacterium tuberculosis. *Infect. Immun.* **78**, 3196–3206 (2010).
70. Davis, J. M. & Ramakrishnan, L. The role of the granuloma in expansion and dissemination of early tuberculous infection. *Cell* **136**, 37–49 (2009).
71. Martinot, A. J. Microbial Offense vs Host Defense: Who Controls the TB Granuloma? *Vet. Pathol.* **55**, 14–26 (2018).
72. Mishra, B. B. *et al*. Mycobacterium tuberculosis protein ESAT-6 is a potent activator of the NLRP3/ASC inflammasome. *Cell Microbiol.* **12**, 1046–1063 (2010).
73. Mishra, B. B. *et al*. Nitric oxide controls the immunopathology of tuberculosis by inhibiting NLRP3 inflammasome-dependent processing of IL-1beta. *Nat. Immunol.* **14**, 52–60 (2013).

74. Volkman, H. E. *et al*. Tuberculous granuloma formation is enhanced by a mycobacterium virulence determinant. *PLoS Biol.* **2**, e367 (2004).
75. Apweiler, R. *et al*. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115–119 (2004).
76. Lee, I., Ouk Kim, Y., Park, S. C. & Chun, J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* **66**, 1100–1103 (2016).
77. Qin, Q. L. *et al*. A proposed genus boundary for the prokaryotes based on genomic insights. *J. Bacteriol.* **196**, 2210–2215 (2014).
78. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
79. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
80. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
81. The Gene Ontology, C. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
82. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. (2015).
83. Szklarczyk, D. *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
84. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
85. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinforma.* **33**, 2938–2940 (2017).
86. Takayama, K., Wang, C. & Besra, G. S. Pathway to synthesis and processing of mycolic acids in Mycobacterium tuberculosis. *Clin. Microbiol. Rev.* **18**, 81–101 (2005).
87. Sharma, A. & Pan, A. Identification of potential drug targets in Yersinia pestis using metabolic pathway analysis: MurE ligase as a case study. *Eur. J. Med. Chem.* **57**, 185–195 (2012).
88. LeMagueres, P. *et al*. The 1.9 A crystal structure of alanine racemase from Mycobacterium tuberculosis contains a conserved entryway into the active site. *Biochem.* **44**, 1471–1481 (2005).
89. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
90. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC Bioinforma.* **4**, 41 (2003).

Acknowledgements

The authors acknowlege ICMR as, A.G. is supported by ICMR-JRF (3/1/3J.R.F.-2016/LS/HRD-(32262)) and CSIR as, M.A. is supported by CSIR-JRF (09/045/(1637)/2019-EMR-1). Authors would also like to acknowlege Bandana Kumari and Deeksha pandey for their meaningful help and suggestions.

Author contributions

A.G. and M.A. collected and organized the data. A.G., N.S. and M.A. analyzed the data. A.G., N.S. and M.A. prepared the manuscript. M.K. conceived the idea. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.