Chapter 2

Bioinformatics analysis and annotation of *Lactobacillus acidophilus* surface proteins and secretome

2.1. Introduction

The proteins that are exposed/embedded on the bacterial cells plays a crucial role in interactions with the environment. The Gram-positive bacteria have a single lipid bilayer cell membrane (CM) with usually a thicker outer peptidoglycan layer as compared to the thinner cell wall of the gram-negative bacteria (Beveridge, 2001). Additionally, the gram-negative bacteria also contains an additional lipid bilayer on the cell exterior known as the outer membrane (OM), separated from the inner cell membrane (IM) by an intermediate space called the periplasm (Zuber *et al.*, 2006). In probiotic gram-positive bacteria, the protein embedded in the cell membrane and the extracellular proteins play a significant role in bacterial interaction especially adhesion to the intestinal milieu. Many extracellular proteins are attached to the components of the bacterial cell wall mainly to peptidoglycan or teichoic acids. Few of the secreted proteins remain anchored to the membrane through one or more transmembrane helices or covalent coupling to lipids (Boekhorst, Wels*, et al.*, 2006). The extracellular and surface-associated proteins are those which are either exposed (GO:0046658- membrane-anchored; GO:0031233- intrinsic to external side of the plasma membrane and GO:0005618- the cell wall) or released in medium (GO:0005576) from the bacterial cell surface. This subset of proteome which includes exoproteome and surface proteome but excluding the integral membrane proteins (GO: 0005887) and proteins which are intrinsic to the internal side of plasma membrane (GO:0031235) is known as "secretome" as defined by Desvaux et al. (Desvaux *et al.*, 2009).

These proteins at the cell surface are involved in vivid processes such as signal transduction, recognition, binding and degradation of complex nutrients like polysaccharides & complex carbohydrates, cell communication, microbe-host interaction and adherence to host cells (Buck *et al.*, 2005; Roos & Jonsson, 2002; Zhou *et al.*, 2008). One such classic example is the protein Msa of *L. plantarum* which is a mannose-specific adhesin (Pretzer *et al.*, 2005). The secreted / surface exposed proteins are exported by various mechanisms and are retained by the bacterial cell via different interaction and/or secreted into the external medium (Figure 2.1). On genome scale, experimental studies have carried out earlier (Bumann *et al.*, 2002; Molloy *et al.*, 2000; Molloy *et al.*, 2001) and thus recent high-throughput genome sequencing have created an opportunity for computational prediction of secretome. The secretome of *Lactobacillus* contains two main categories of proteins the secreted proteins which release from the cell and the surface associated proteins (Figure 2.1).

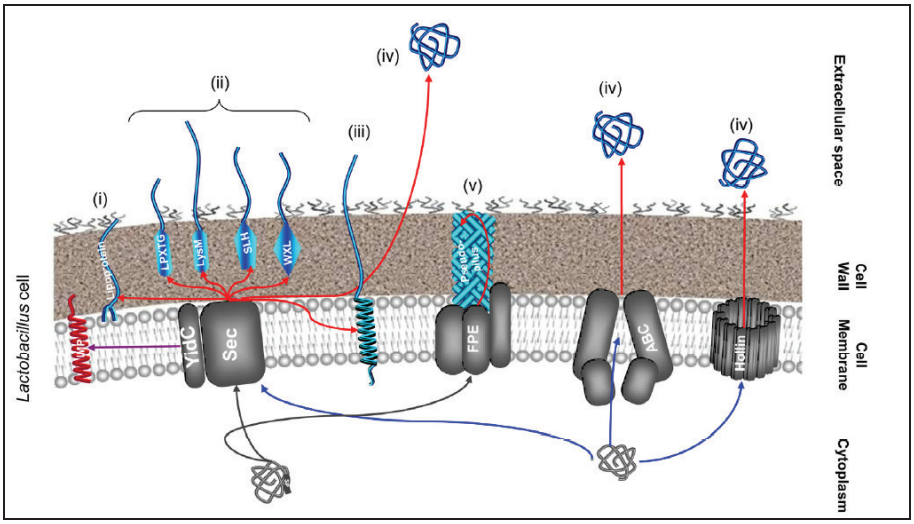


Figure 2.1: **Schematic representation of the secretome in *Lactobacillus*.** The secreted proteins (blue) grouped by various mechanisms: (i) anchored to cytoplasmic membrane (CM) i.e. Lipoproteins; (ii) attached to the cell wall (CW) either covalently (LPxTG proteins) or non-covalently (LysM, SLH or WXL domains/motifs); (iii) anchored to the CM via the N-/C- terminal transmembrane helix; (iv) released into the extracellular medium via Sec, holin or ABC transporters; (v) being part of cell-surface appendages (assembled via FPE). (SP) indicates that the proteins carry an N-terminal signal peptide and their route targeting to the CM is depicted as black arrows, whereas the blue arrow denotes proteins lacking any signal peptide. Source: (Desvaux *et al.*, 2009).

The latter are categorized into several sub-categories: the proteins anchored in the cytoplasmic membrane via N-terminal single hydrophobic domain; the proteins anchored in the cytoplasmic membrane via C-terminal single hydrophobic domain; the proteins anchored in the cytoplasm via lipid-anchored membrane i.e. Lipoproteins; the proteins attached to the cell wall either covalently (LPxTG- motif) or non-covalently (WxL domain, SLH domain, LysM domain, PG-binding domain, SH3 domain and the choline binding domain).

2.1.1. Secreted proteins

Till date seven primary protein secretion mechanisms have been identified and characterized in Gram-positive bacteria, the Sec (secretion), FEA (flagella export apparatus), FPE (fimbrial-protein exporter), Tat (twin-arginine translocation), peptide-efflux ABC transporter, Wss (WXG100 secretion) and pore-forming holing (Desvaux *et al.*, 2009; Driessen & Nouwen, 2008; Lee *et al.*, 2006; van Wely *et al.*, 2001). Earlier studies on *Lactobacillus* genomes revealed that they do not encode the main factors

involved in the Tat, FEA and Wss protein secretion pathways and has only genes encoding for the remaining secretion pathways (Kleerebezem *et al.*, 2010). The proteins targeted to the Sec pathways have an N-terminal signal peptide, which typically contains: a positively charged N-terminus; a stretch of 15-25 hydrophobic residues; and the C-terminal region that may include a signal peptidase cleavage site (Driessen & Nouwen, 2008).

2.1.2. Covalently anchored proteins

2.1.2.1. N- or C- terminally anchored proteins

The N-terminal signal peptides that target proteins for secretion contains the characteristic N, H and C regions as mentioned above. After completion of the secretion process the C region of the signal peptide remains exposed on the extracytoplasmic side of the membrane and as it contains the type-I or type-II SPase, the signal peptide can be cleaved, and the mature protein is then released. However, many C-region does not possess this cleavage motif, and they remain N-terminally anchored in the cell membrane. In case a signal peptide C region contains a typical Type-I SPase cleavage site, after processing for secretion by Sec pathway will remain anchored to the cytoplasmic membrane due to the presence of C-terminal transmembrane domain and thus exposing the protein to the extracellular milieu.

2.1.2.2. Lipoproteins

Few proteins possess a lipoprotein signal peptide which is also similar to N, H and C regions of Sec pathway with the exception that the H-region is shorter, and the C-region contains the lipobox motif [L-(A/S)-(A/G)-C] which directs them to the lipoprotein biogenesis machinery after transport. The conserved Cys residue in the lipobox motif undergoes diacylglycerol modification by the lipoprotein diacylglycerol transferase which then enables the covalent binding of the lipoprotein to the cell membrane. After lipidation, cleavage occurs at N-terminal of the Cys residue by the Type-II SPase, thereby anchoring the mature protein to the membrane via thioester linkage (Hutchings *et al.*, 2009).

2.1.2.3. LPxTG anchored proteins

Some secreted proteins contain an LPxTG motif anchor, located in the C-terminal region. This motif is followed by a C-terminal membrane anchor domain which consists a stretch of hydrophobic residues and a positively charged tail (Marraffini *et al.*, 2006).

The LPxTG motif is recognised by a family of enzymes known as Sortase (SrtA), which cleaves between the Thr and Gly residues and finally links the exported protein covalently to the cell wall peptidoglycan layer and displayed on the microbial surface (Marraffini *et al.*, 2006). The *Lactobacillus* adhesion to mucus involves mucus binding protein (MubP) which has mucus binding domain as well as N-terminal signal peptide and C-terminal LPxTG motif.

2.1.3. Non-covalently anchored proteins

2.1.3.1. LysM domain

The LysM domain module is a highly conserved carbohydrate binding module found in plants, viruses, bacteria, fungi and animals. LysM domain (PF01476) recognises polysaccharides having N-acetylglucosamine (GlcNAc) residues found in peptidoglycan layer of the cell membrane (Mesnage *et al.*, 2014). Surface proteins having LysM domain are also involved in adhesion recognition of host molecules.

2.1.3.2. Choline-binding domain

The choline binding domain (PF01473) is mainly found in extracellular enzymes and protein as conserved tandem repeats of glycine and aromatic acids (Y and G) in a stretch of 20 amino acids. They can bind choline residues of cell wall teichoic and lipoteichoic acids (LTA), thereby anchoring the protein to the cell surface (Wren, 1991).

2.1.3.3. Peptidoglycan binding domain (PG binding domain)

A peptidoglycan-binding domain is composed of three alpha-helices located at the N- or the C-terminus of cell wall degrading enzymes (Pfam PF01471).

2.1.3.4. S-layer protein with SLH domain

Surface-layers (S-layers) proteins form a para-crystalline assembly of proteins/glycoproteins on the cell surface of bacteria and archaea which are primarily involved in mediating adhesion to host surfaces (Sleytr *et al.*, 2014). S-layer proteins are often non-covalently linked with the cell wall polysaccharides, teichoic acids and EPS, thus providing structural stability as well as adhesion factor. The Pfam database contains different S-layer protein domains responsible for non-covalent anchoring to the cell wall: S_layer_C (PF05124), S_layer_N (PF05123), SLAP (PF03217) and SLH (PF00395).

2.1.3.5. WxL domain

The WxL domain is a C-terminal cell wall binding domain which was first identified in proteins from *Lactobacillus* (Boekhorst, Wels*, et al.*, 2006; Siezen *et al.*, 2006). They are found in gene clusters that also encodes additional extracellular proteins with C-terminal membrane anchors and LPxTG type peptidoglycan anchors (Siezen *et al.*, 2006).

2.1.3.6. SH3 domain

The SH3 domain is a eukaryotic domain supposed to be the counterpart of the prokaryotic SH3b domain and have been proposed to be involved in targeting and binding the proteins to the peptidoglycan layer and are thought to recognise specific sequences within the cross-linking peptide bridges (Baba & Schneewind, 1996).

In this chapter, we have predicted the surface exposed extracellular protein and secreted protein which together constitute the secretome of the *Lactobacillus acidophilus*. Further, the proteins were subjected to categorization based on their functions and class of proteins. We have used different prediction algorithms and software/ tools for subcellular localisation prediction. Also, the domains of the proteins have been identified from the Pfam database, and signature patterns were recognised if any.

2.2. Methodology

2.2.1. Tools and software used for prediction

The genome sequence of Gram-positive bacteria *Lactobacillus acidophilus* was downloaded from UniProt (http://uniprot.org). Standalone BLAST 2.21 (Altschul *et al.*, 1990) downloaded from NCBI FTP site was used for sequence homology search against NR database. Multiple sequence alignments were built using MUSCLE package as well as in Clustal W/X (Larkin *et al.*, 2007). For prediction of the signal peptide, HMM method of SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP-3.0/) and TM based network of SignalP 4.0 (http://www.cbs.dtu.dk/services/SignalP-4.0/) were used, while for predicting transmembrane helix TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) and Phobius (Kall *et al.*, 2004) were used. PredSi (http://www.predisi.de) tool was used for prediction of signal peptidase cleavage site. HMM profiles were generated if not available from Pfam using hmmbuild program from HMMER 3.1b2 package (Eddy, 1998). PRED-LIPO and LipoP were used to

identify lipoproteins (Bagos *et al.*, 2008; Juncker *et al.*, 2003). ScanProsite (http://prosite.expasy.org/scanprosite/) tool was used to check for the presence of signature motifs. For accurate prediction of cell anchoring motif LPxTG and Lipoproteins, HMM profiles were constructed based on multiple sequence alignment output of conserved families or protein of known functions following which the hmm models were used to search for similar sequences in the genomic data. PERL, grep, awk and shell commands were use for processing a large number of sequences and further characterization. MS Excel was used for data filtering and data management. Manual inspection of all the sequence was further carried to ensure the prediction accuracy. Pfam (Finn *et al.*, 2014) and CDD (Marchler-Bauer & Bryant, 2004) databases used for prediction of protein domains in the sequences.

2.2.2. Genome annotation

For prediction of surface anchored and secreted proteins from *L. acidophilus* genome, a pipeline was designed using several tools and a modified strategy based on the finding from *L. plantarum* secretome prediction (Boekhorst, Wels, *et al.*, 2006) was developed. Various methods were employed for different category of prediction class and are mentioned further in detail. A schema of the prediction and methodology is mentioned below (Figure 2.2).

2.2.2.1. Secreted/ Surface-associated protein prediction

The source of protein sequence information: The genome sequence of Gram-positive bacteria *Lactobacillus acidophilus* were downloaded from UniProt (release 2016) by downloading proteome set (up000006381) using search term "*Lactobacillus acidophilus* NCFM". The total genome sequences constituted of secreted proteins, single and multi-transmembrane proteins and intracellular proteins. The downloaded sequences in Fasta format were further searched for hypothetical genes, putative function gene and uncharacterized genes. Partial sequences or fragment were removed from the genome set using keyword [Fragment] by the grep command searching from the annotation line.

Intracellular proteins: For prediction of intracellular proteins, the presence of transmembrane region and the signal peptide was evaluated using TMHMM 2.0, Phobius and SignalP 3.0 tools. After prediction of TM and signal peptide, the sequences subjected to manual check and those proteins having no TM, and neither signal peptide was termed as intracellular proteins. These intracellular proteins were excluded from further studies,

and the resulting protein sequences were termed as secreted and membrane proteins. If the transmembrane region was found in first 60 amino acid region of protein, it was further confirmed using NN model of SignalP that whether the predicted transmembrane helix in the N-terminal is a signal peptide or not.

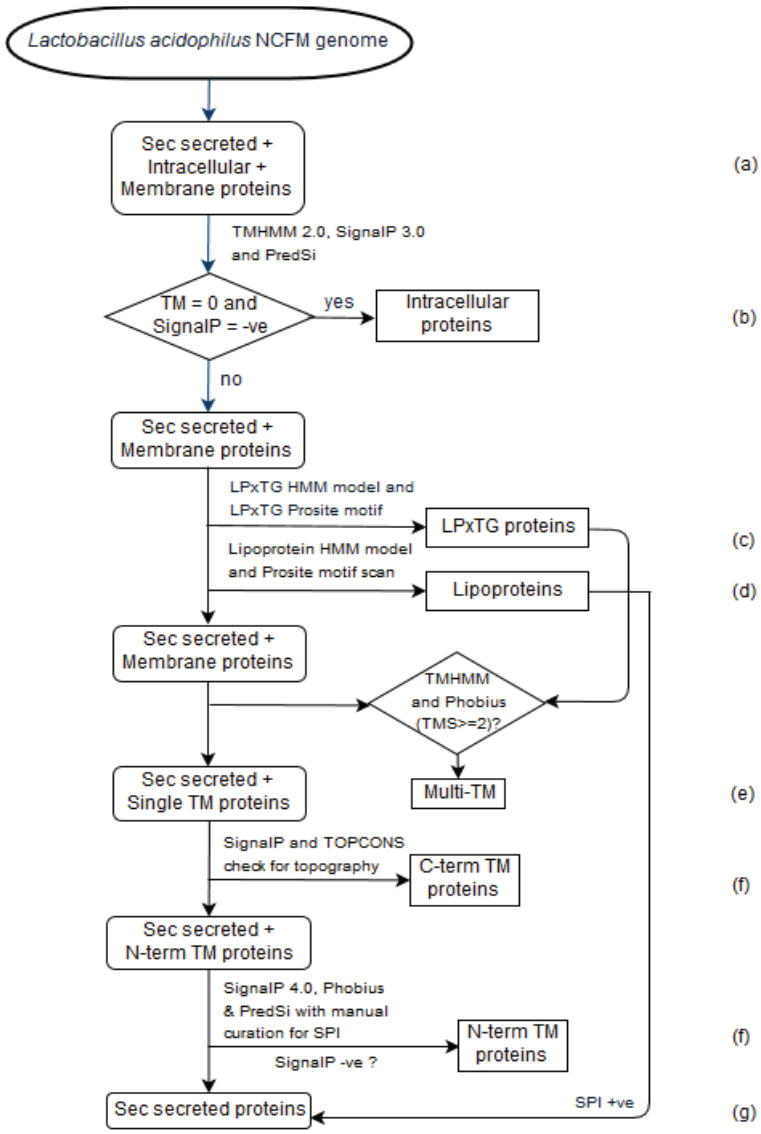


Figure 2.2: Flowchart schema of secretome prediction from *L. acidophilus* genome. Complete protein sequences were collected which comprises of secreted, intracellular and membrane proteins (a). The proteins without any predicted TM segments and signal peptide were considered as intracellular proteins (b). The protein with LPxTG motif (c) and lipoproteins (d) was categorised based on HMM profiles, motifs and manual curation. Proteins with TM segments and signal sequences were divided into multi-TM proteins (e), N-/C- anchored membrane proteins (f) and secreted/released proteins (g). Abbreviation: A-S =Anchored-Secreted; TM = Trans Membrane Segment; SignalP = Signal peptide; N-/C- term = N-/C-terminal transmembrane-anchored; HMM Hidden Markov Model Profiles; SPI /SPII = Type I/II SPases.

LPxTG anchor proteins: For prediction of LPxTG cell-walled anchor proteins, the following criteria was used for detection: (i) the LPxTG proteins contains an N-terminal signal peptide having Type I SPase cleavage site; (ii) the C-terminal of proteins contains the LPxTG motif; (iii) the LPxTG motif is followed by a C-terminal membrane anchor domain which consists of hydrophobic amino acids and a positively charged tail (Arg/Lys) (Marraffini *et al.*, 2006). The sequences were further subjected to predict the Gram-positive anchor motif (PS50847) using ScanProsite. The LPxTG proteins were also pattern searched using a motif used in earlier study: [LY]PX[TSA][GNAST]X(0,10)(DEQNKRP)(DEQNKRP) (DEQNKRP) (DEQNKRP) (DEQNKRP) (DEQNKRP) (DEQNKRP) (DEQNKRP)X(0,15)[DEQNKRH]X(0,5)> which detects LPxTG like motif which is followed by at least 10 non-polar residues capable of spanning the membrane and at least one polar or positive charged residues within the last five residues representing the positive charged tail (Roche *et al.*, 2003). A set of 168 well-known sequences containing LPxTG motif were downloaded from UniProt by using search term: database: [(type: prosite PS50847) AND reviewed: yes]. The sequences were further aligned using Clustal W/X followed by generation of HMM profile for training set using hmmbuild. The sequence logo for an aligned training set of sequences was generated to check for the presence of LPxTG residues (Figure 2.3). The sequences from secreted and membrane fraction of *L. acidophilus* were then further tested for the presence of HMM profile using hmmsearch and all proteins sequences with an E-value cut-off below 1e-05 were considered as putative LPxTG anchoring proteins. The predicted LPxTG sequences were manually checked for the presence of single C-terminal transmembrane domain and only sequences having a C-terminal transmembrane segment and positively charged tail was further considered as LPxTG motif-containing proteins.

Lipoproteins: The proteins containing N-terminal lipobox motif having a common mechanism for protein secretion and membrane anchoring through covalent binding of a conserved cysteine residue to lipid membrane were searched using LipoP and PRED-LIPO tools. The sequences predicted by previous tools were pattern searched for the presence of prokaryotic membrane lipid attachment site profile (PS51257) using ScanProsite. The hmmbuild program was used to generate the hmm profile from the multiple sequence alignments of known sequences and used for searching Lipoprotein sequences using hmmsearch. Putative proteins with more than single transmembrane domain were excluded and further considered as multi-transmembrane proteins. As the

lipoproteins are known to have shorter H region in the N-terminal signal peptide with Type II SPase, the predictions were refined based on the presence of signal peptide cleavage site along with the presence of conserved lipobox motif [L-(A/S)-(A/G)-C] in the C region of signal peptide. The putative sequences lacking Type II SPase (SPII) were removed from the final prediction. The presence of indispensable conserved cysteine throughout the training set along with the presence of Type II SPase cleavage site was final indicator of lipobox motif (Hutchings *et al.*, 2009).

Multi-transmembrane proteins: The secreted and membrane proteins of *L. acidophilus* proteins were evaluated for the presence of transmembrane segments using TMHMM 2.0 and PhobiuS. The proteins thus having at least two transmembrane segment were further considered as multi-transmembrane proteins (multi-TM) which were excluded from further analysis and prediction of secretome. The predicted proteins having cell wall anchoring motif LPxTG and Lipobox motif were also checked for the occurrence of the transmembrane segment and if more than or equal to 2, and if found were further considered as multi-TM proteins.

N-terminal and C-terminal membrane proteins: The N-terminal membrane proteins were detected by evaluating the presence of Type I and Type II SPase cleavage site in the C-region of N-terminal signal peptide sequences. If the C-region does not possess the cleavage motif or contains a Type I SPase motif that is not cleaved, they were considered N-terminal anchored single membrane proteins (Kleerebezem *et al.*, 2010). The C-terminal anchored proteins were detected based on the presence of Type I SPase cleavage site along with a single C-terminal transmembrane segment which anchors the protein in the C-terminal membrane after signal peptide processing. The signal peptide and transmembrane were detected as mentioned above using TMHMM 2.0 and SignalP 4.0. PredSi was used to assess the cleavage site and additionally the sites were also manually evaluated. The topography of transmembrane segment has been assessed using TOPCONS prediction (Tsirigos *et al.*, 2015).

Sec secreted proteins: The proteins which have SPase cleavage site in the N-terminal signal peptide region were classified as proteins secreted by Sec translocation pathway. The presence of Signal peptide with cleavage site and absence of any transmembrane region in the C-terminal of proteins were considered as proteins secreted or released by Sec pathway. The Lipoproteins with Type I SPase cleavage site was also further included

in the Sec secreted protein category.

2.3.2.2. Non-covalent cell-wall binding/anchored proteins

The complete proteome of *L. acidophilus* was searched for the occurrence of the specific cell-wall binding domain which has been reported earlier in cell-wall anchoring or binding (Kleerebezem *et al.*, 2010). Pfam database (Finn *et al.*, 2014) search was used to identify the presence of domains in the proteins and further checked manually to verify their accurate prediction The presence of well-known cell wall binding domain such as LysM (lysin motif) domain (PF01476), choline-binding domains (PF01473), peptidoglycan binding domain (PF01471), SLAP (PF03217), SLH (PF00395), S_layer_C (PF05124), S_layer_N (PF05123), etc. Other domains found in cell-wall anchoring proteins in *L. plantarum* (Boekhorst, Wels*, et al.*, 2006) were also searched for their presence in the *L. acidophilus* genome like CW_binding_1 (PF01473), SH3_3 (PF08239), ABC_sub_bind (PF04392), Big_2 (PF02368), FMN_bind (PF04205), Peripla_BP_2 (PF01497) and Lipoprotein_9 (PF03180).

2.3.2.3. Adherence proteins

As adhesion factors are considered to play a significant role in interactions of host-microbe and as an important virulence factor in pathogens, prediction of adherence proteins or domain involved in binding is very vital (Navarre & Schneewind, 1999). The genome of *L. acidophilus* was searched for the presence of domains involved in binding or adherence identified from earlier evidence. Pfam search was performed to search protein domains for the presence of MucBP (PF06458), Collagen_bind (PF05737), Cna_B (PF05738), Fibronectin type III domain - fn3 (PF00041) and FbpA (PF05833).

2.3. Results

2.3.1. Secreted/ Surface-associated protein prediction

The source of protein sequence information: A total of 1,873 protein sequences were downloaded from UniProt with "*Lactobacillus acidophilus* NCFM" as a search keyword. The sequences were filtered to remove fragment and other non-relevant proteins making the final count of sequences to 1,859 proteins. From the complete protein sets, only 282 protein sequences are reviewed, and rest of 1,577 proteins are un-reviewed sequences. From the un-reviewed protein sequences, 718 protein sequences are annotated as putative or assigned a putative function.

Intracellular protein prediction: Analysis using transmembrane and signal peptide search revealed that a total of 1,274 proteins from complete protein set have no transmembrane helix /segment and neither a signal peptide for secretion. The 1,274 protein sequence set thus was labelled as intracellular proteins, and they were further excluded from the analysis of secretome prediction. The remaining set of 585 protein sequences were further analysed for secreted and surface exposed protein prediction.

LPxTG **motif anchoring proteins:** The resulting set of 585 proteins sequences which are putative members of secretome were further examined for the presence of cell-wall sorting motif LPxTG which enables a protein to be covalently attached to the peptidoglycan by the activity of Sortase enzyme (Marraffini *et al.*, 2006). As the LPxTG motif containing proteins generally consists of the N-terminal signal peptide (Type I) cleavage site and the motif is located in the C-terminal of protein followed with membrane anchor region composed of hydrophobic residues and positively charged tail residues, the putative secretome sequences were searched for the same pattern in the C-terminal of the mature domain. Various other tools and methods were employed to detect the LPxTG motif proteins and the sequence logo for an aligned training set of sequences were generated to evaluate the presence of lipobox motif residues [L-(A/S)-(A/G)-C] (Figure 2.3). The hits were further confirmed through manual verification of the motif residues and the presence of C-terminal single transmembrane domain (Table 2.1). A total of 17 proteins were evaluated further and out of which only 12 proteins were assigned to be LPxTG motif containing protein from the genome (Table 2.1; marked in bold). The rest of 5 proteins had a multi-transmembrane segment with at least 2 TM segments. Either the other criteria for LPxTG proteins like the LPxTG conserved residues or the positively charged tail was found to be incomplete and thus excluded from the annotation of LPxTG proteins.

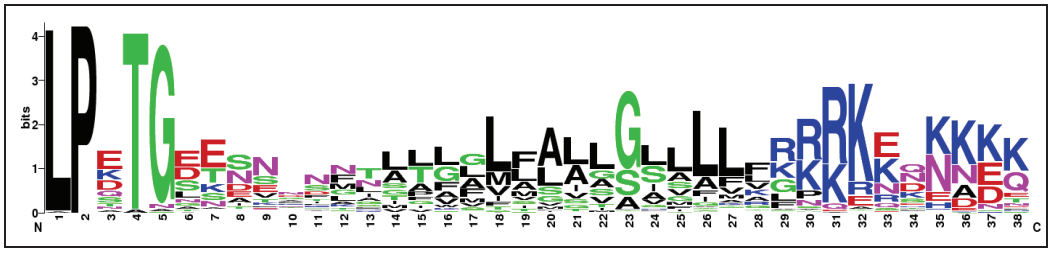


Figure 2.3: Sequence logo representing conserved LPxTG residues and motif in the alignment of well-known 168 LPxTG **protein sequences.**

Table 2.1: List of the 12 putatively identified LPxTG motif-containing proteins.

No.	Protein ID	Protein	Length	A	B	C	D	E	F
1	**Q5FMY3**	hypothetical protein	435	+	+	+	SPI	LPKTG	1
2	**Q5FKA7**	mucus binding protein	346	+	+	+	-	LPETG	1
3	**Q5FKA6**	mucus binding protein	2650	+	+	+	SPI	LPQTG	1
4	**Q5FJA7**	mucus binding protein precursor	4326	+	+	+	SPI	LPQTG	1
5	**Q5FJ09**	fibrinogen-binding protein	991	-	+	+	-	LPQTG	1
6	**Q5FIP8**	surface protein	2539	+	+	+	-	LPQTG	1
7	**Q5FIM8**	surface protein	1659	-	+	+	-	LPQTG	1
8	**Q5FIM7**	surface protein	1924	+	+	+	SPI	LPQTG	1
9	**Q5FIL0**	mucus binding protein precursor	1174	+	-	+	SPI	LPATG	1
10	**Q5FIF3**	mucus binding protein precursor	1208	+	+	+	SPI	LPQAG	1
11	**Q5FIC2**	hypothetical protein	1376	+	+	+	SPI	LPQTG	1
12	**Q5FI76**	hypothetical protein	438	+	+	+	SPI	LPSTG	1
13	Q5FKH1	Putative uncharacterized protein	81	-	-	+	-	LPLIV	2
14	Q5FMG8	Putative uncharacterized protein	205	-	-	+	SPI	-	2
15	Q5FK90	Branched-chain amino acid transport system carrier protein	456	-	+	-	-	-	12
16	Q5FKT0	Cell division protein	394	-	+	-	-	LPITG	10
17	Q5FHX0	Putative uncharacterized protein	288	-	-	-	-	LPLFG	8

A: ScanProsite motif prediction (PS50847); **B:** Motif prediction; **C:** HMM prediction; **D:** Secretory cleavage (SPI- Type I SPases); **E:** LPxTG motif cleavage site (non-conserved residues are marked in bold); **F:** No. of TM segment. (+) and (-) indicates positive and negative hits respectively.

Lipoprotein prediction: The remaining 573 protein sequences from the above prediction step were further evaluated for the presence of lipoproteins motif box. The lipoproteins in *L. acidophilus* were found to be a second largest membrane-anchored group in the predicted secretome with total 42 protein sequences. A known training set of bacterial 65 well-known lipoproteins were used for generation of lipoprotein HMM profile and s sequence logo from the aligned sequences were generated for evaluating the presence of lipoprotein motif (Figure 2.4). Initially, 47 protein sequences were identified as putative lipoproteins using HMM profile model trained with 65 known lipoproteins. Alternatively, 44 proteins were found to be positive for the prosite motif (PS51257). Out of the 47 putative proteins, three protein were found to have Type I SPase cleavage site

(SPI) and two proteins were identified as multi-TM segments which were further excluded from the annotation of lipoproteins. A final set of 42 proteins were identified as putative lipoproteins based on the detection of HMM profile and conserved lipoprotein box motif with SPII cleavage site (Table 2.2; marked in bold).

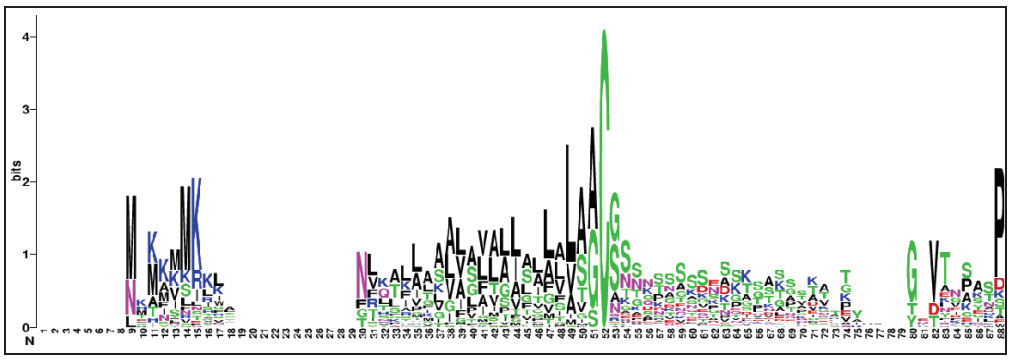


Figure 2.4: Sequence logo representing conserved lipobox motif in the alignment of well-known 65 prokaryotic Lipoprotein sequences.

Table 2.2: List of identified 42 Lipoproteins from *L. acidophilus*.

No.	Protein ID	Protein	Length	A	B	C	D
1	**Q5FN03**	Putative alkylphosphonate ABC transporter	313	+	+	SPII	+
2	**Q5FMM2**	Alkylphosphonate ABC transporter	309	+	+	SPII	+
3	**Q5FMJ5**	DNA polymerase (EC 2.7.7.7)	264	+	+	SPII	+
4	**Q5FMJ4**	Putative uncharacterized protein	253	+	+	SPII	+
5	**Q5FMI2**	Oligopeptide binding protein	541	+	+	SPII	+
6	**Q5FMI1**	Peptide binding protein	542	+	+	SPII	+
7	**Q5FM21**	ABC transporter	288	+	-	SPII	+
8	**G1UB60**	ABC transporter sugar-binding protein	431	+	+	SPII	+
9	**Q5FLL3**	Putative uncharacterized protein	382	+	+	SPII	+
10	**Q5FLH1**	Glutamine ABC transporter	272	+	+	SPII	+
11	**Q5FLG4**	Glutamine-binding protein	272	+	+	SPII	+
12	**Q5FLF9**	Glycerol-3-phosphate ABC transporter	432	+	+	SPII	+
13	**Q5FKM7**	Putative uncharacterized protein	198	+	+	SPII	+
14	**Q5FK78**	Putative uncharacterized protein	200	+	-	SPII	+
15	**Q5FK25**	Putative flavodoxin	203	+	+	SPII	+
16	**Q5FK24**	Putative uncharacterized protein	187	+	+	SPII	+
17	**Q5FJZ3**	Diaphanous-like protein	202	+	+	SPII	+
18	**Q5FJS3**	Pheromone-peptide binding protein	541	+	+	SPII	+
19	**Q5FJJ4**	Oligopeptide ABC trasporter substrate binding protein	585	+	+	SPII	+

No.	Protein ID	Protein	Length	A	B	C	D
20	**Q5FJJ3**	Oligopeptide ABC trasporter substrate binding protein	589	+	+	SPII	+
21	**Q5FJE7**	Putative alpha-beta superfamily hydrolase	299	+	+	SPII	+
22	**Q5FJ99**	Oligopeptide ABC trasporter substrate binding protein	581	+	+	SPII	+
23	**G1UB54**	ABC transporter sugar binding protein	418	+	+	SPII	+
24	**Q5FJ47**	Putative surface layer protein	294	+	+	SPII	+
25	**Q5FJ23**	D-ribose-binding protein	328	+	+	SPII	+
26	**Q5FJ08**	Putative uncharacterized protein	336	+	+	SPII	+
27	**Q5FJ07**	Putative ABC transporter	301	+	+	SPII	+
28	**Q5FIS0**	Foldase protein PrsA (EC 5.2.1.8)	300	+	+	SPII	+
29	**Q5FIM1**	Glycerol-3-phosphate ABC transporter	433	+	+	SPII	+
30	**Q5FIK0**	Putative membrane protein	202	+	+	SPII	+
31	**Q5FIJ7**	Oligopeptide ABC transporter substrate binding protein	546	+	+	SPII	+
32	**Q5FIJ0**	Amino acid ABC transporter	272	+	+	SPII	+
33	**Q5FIG5**	Putative uncharacterized protein	336	+	+	SPII	+
34	Q5FIC3	Putative uncharacterized protein	336	-	+	SPI	+
35	**Q5FI94**	ABC transporter, periplasmatic component	298	+	+	SPII	+
36	Q5FI22	Putative aggregation promoting protein	120	-	+	SPI	+
37	**Q5FI08**	Maltose ABC transporter permease protein	408	+	+	SPII	+
38	**Q5FHT7**	Putative uncharacterized protein	353	+	+	SPII	+
39	**Q5FHT6**	Putative lipoprotein A-antigen	361	+	+	SPII	+
40	**Q5FHS2**	Oligopeptide ABC transporter substrate binding protein	541	+	+	SPII	+
41	**Q5FHR9**	Oligopeptide ABC transporter substrate binding protein	542	+	+	SPII	+
42	Q5FHQ4	Membrane protein insertase YidC (Foldase YidC) (Membrane integrase YidC) (Membrane protein YidC)	291	+	+	Multi TM	-
43	**Q5FHU6**	Putative uncharacterized protein	215	+	+	SPII	-
44	**Q5FIS1**	Putative tropomyosin	111	+	+	SPII	-
45	**Q5FJF0**	Peptide binding protein	184	+	-	SPII	-
46	Q5FLV4	Putative uncharacterized protein	122	-	+	Multi TM	-
47	Q5FIY2	Putative family protein	322	+	+	SPI	-

A: HMM prediction; **B:** ScanProsite motif prediction (PS51257); **C:** Secretory cleavage (SPI- Type I SPase, SPII- Type II SPase); **D:** LipoP prediction. (+) And (-) indicates positive and negative hits respectively. The identified 42 lipoproteins are marked in bold.

Multi-transmembrane proteins: The remaining 531 proteins in the secretome were checked for the occurrence of the multi-transmembrane helix. The proteins having more than and equal to 2 transmembrane segment were classified as multi-transmembrane proteins. A total of 39 proteins were found which did not possess any transmembrane segment, while 128 proteins were identified to contain one transmembrane segment. A set of total 362 proteins were identified to have more than or equal to 2 trans-membrane segments and subsequently was classified as multi-transmembrane proteins and were further excluded from the secretome analysis. The remaining 169 putative proteins in the secretome were subjected for further prediction. The proteins identified as putative lipoproteins and LPxTG motif containing proteins which have more than two transmembrane segments were also included as the multi-transmembrane protein.

N-terminal and C-terminal membrane anchor proteins: The proteins left in the putative secretome prediction group was analysed for the presence of Type I SPase cleavage site. A total of 108 proteins were identified to be N-terminally anchored (Table 2.3), while 2 proteins were identified as putative C-terminal anchor proteins (Table 2.4).

Table 2.3: List of 108 identified proteins as putative N-terminal anchored protein.

No.	Protein ID	Protein name	Length
1	Q5FMX1	Putative ABC transporter substrate binding protein	330
2	Q5FMW1	Penicillin-binding protein	374
3	Q5FMU2	Putative uncharacterized protein yycH	452
4	Q5FMU1	Putative uncharacterized protein yycI	274
5	Q5FMT9	Putative heat shock related serine protease (EC 3.4.21.-)	423
6	Q5FMT6	Putative uncharacterized protein	91
7	Q5FMS6	Putative uncharacterized protein lemA	186
8	Q5FMR9	Putative uncharacterized protein	52
9	Q5FMR6	Putative beta-glucanase (EC 3.2.1.-)	374
10	Q5FMQ1	Putative extracellular protein	313
11	Q5FMK5	Putative uncharacterized protein	149
12	P35829	S-layer protein (Surface layer protein) (SA-protein)	444
13	Q5FMJ0	Transcriptional regulator	367
14	Q5FMA6	Putative uncharacterized protein	125
15	Q5FM25	Protein translocase subunit SecE	55
16	Q5FLX1	Protein translocase	136
17	Q5FLV5	Tropomyosin-related protein	141
18	Q5FLV4	Putative uncharacterized protein	122
19	Q5FLV1	Putative glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46)	455
20	Q5FLQ4	Putative uncharacterized protein	46

No.	Protein ID	Protein name	Length
21	Q5FLN2	Putative uncharacterized protein	48
22	Q5FLL6	N-acetylmuramidase (EC 3.5.1.28)	215
23	Q5FLF6	Ribose-5-phosphate isomerase A (EC 5.3.1.6) (Phosphoriboisomerase A) (PRI)	230
24	Q5FLC9	Putative uncharacterized protein	234
25	Q5FLC0	UTP--glucose-1-phosphate uridylyltransferase (EC 2.7.7.9)	300
26	Q5FLB5	Putative uncharacterized protein	305
27	Q5FL38	Spermidine and putrescine periplasmatic ABC transporter	357
28	Q5FL36	Putative uncharacterized protein	319
29	Q5FL26	Putative uncharacterized protein	315
30	Q5FL24	Putative uncharacterized protein	33
31	Q5FL19	Putative Competence protein	119
32	Q5FL18	Putative uncharacterized protein	142
33	Q5FL17	Putative uncharacterized protein	200
34	Q9RGY5	ATP synthase subunit b (ATP synthase F(0) sector subunit b) (ATPase subunit I)	169
35	Q5FKX1	Cell division regulator	570
36	Q5FKV6	Cell division protein FtsL	120
37	Q5FKV5	Penicillin-binding protein	720
38	Q5FKV1	Cell division protein DivIB	285
39	Q5FKT6	Putative uncharacterized protein	559
40	Q5FKS6	Putative uncharacterized protein	346
41	Q5FKS5	Competence protein	227
42	Q5FKQ5	Penicillin-binding protein	369
43	Q5FKP6	Putative uncharacterized protein	68
44	Q5FKN9	Cellobiose-specific PTS IIC (EC 2.7.1.69)	111
45	Q5FKL3	Lipoprotein	284
46	Q5FKJ0	Putative uncharacterized protein	27
47	Q5FKF7	Putative N-acetylmuramidase (EC 3.5.1.28)	153
48	Q5FKE2	DNA topoisomerase 1 (EC 5.99.1.2)	705
49	Q5FKD4	Putative hydrolase	284
50	Q5FKB5	Secreted protein	401
51	Q5FK82	Glutamine ABC transporter substrate-binding protein	286
52	Q5FK19	Putative flavodoxin	115
53	Q5FJZ4	Lysin	382
54	Q5FJX6	Penicillin-binding protein 1A	776
55	Q5FJX2	Putative uncharacterized protein	165
56	Q5FJV5	Putative signal peptidase	80
57	Q5FJU4	Putative enterolysin A	221
58	Q5FJT1	Putative enterolysin A	213
59	Q5FJR5	Putative uncharacterized protein	501

No.	Protein ID	Protein name	Length
60	Q5FJR3	Putative uncharacterized protein	182
61	Q5FJR1	Putative uncharacterized protein	53
62	Q5FJQ6	Putative lactocepin S-layer protein (EC 3.4.21.96)	180
63	Q5FJP7	Putative sortase	229
64	Q5FJE5	Lysin	155
65	Q5FJC1	Putative biofilm-associated surface protein	66
66	Q5FJA4	Lipoprotein	283
67	Q5FJ84	Putative uncharacterized protein	267
68	Q5FJ76	Putative uncharacterized protein	44
69	Q5FJ73	Putative uncharacterized protein	252
70	Q5FJ52	Putative alpha-galactosidase	62
71	Q5FJ01	Putative rhodanese-related sulfurtransferase	133
72	Q5FIZ6	Penicillin-binding protein	702
73	Q5FIZ3	PrtP	1627
74	Q5FIU3	Putative membrane protein	293
75	Q5FIS9	Putative serine protease	694
76	Q5FIR5	Penicillin-binding protein	685
77	Q5FIQ8	Putative cell surface protein	229
78	Q5FIQ5	Putative uncharacterized protein	109
79	Q5FIQ4	Putative penicillin-binding protein	343
80	Q5FIP6	Putative uncharacterized protein	75
81	Q5FIM3	Putative beta-lactamase (EC 3.5.2.6)	318
82	Q5FIL5	Putative DNA nuclease	285
83	Q5FIK1	Putative membrane protein	180
84	Q5FIJ5	Putative uncharacterized protein	193
85	Q5FII5	Conserved domain	227
86	Q5FII1	Putative uncharacterized protein	423
87	Q5FIH6	Putative uncharacterized protein	45
88	Q5FIH5	L-asparaginase (EC 3.5.1.1)	330
89	Q5FIH3	Maltose-6'-phosphate glucosidase	445
90	Q5FIH2	Putative membrane protein	280
91	Q5FIF2	Thermostable pullulanase (EC 3.2.1.41)	1185
92	Q5FIF0	Putative uncharacterized protein	314
93	Q5FIE3	UTP--glucose-1-phosphate uridylyltransferase (EC 2.7.7.9)	294
94	Q5FIC9	Phospho-glucosyltransferase	217
95	Q5FIC5	Exopolysaccharide biosynthesis protein	351
96	Q5FIB1	Putative uncharacterized protein	240
97	Q5FIA1	Transcriptional regulator	424
98	Q5FI84	Putative uncharacterized protein	33
99	Q5FI82	Putative uncharacterized protein	64
100	Q5FI75	Uncharacterized protein LBA1794	196
101	Q5FI59	Putative uncharacterized protein	63

No.	Protein ID	Protein name	Length
102	Q5FI26	Secreted protein	402
103	Q5FHW7	Signal peptidase I (EC 3.4.21.89)	210
104	Q5FHW5	Putative cell surface hydrolase	303
105	Q5FHV4	Extramembranal transfer protein	428
106	Q5FHV0	Putative uncharacterized protein	378
107	Q5FHU3	Phage capsid protein	204
108	Q5FHR7	Putative uncharacterized protein	125

Table 2.4: List of 2 identified proteins as putative C-terminal anchored protein

No.	Protein ID	Protein name	Length
1	Q5FJ90	Putative uncharacterized protein	262
2	Q5FIQ0	Putative mucus binding protein	643

Sec- dependent secreted proteins: The rest of the proteins in the putative secretome group were further evaluated for the presence of signal peptide cleavage site. A final set of 59 sequences were identified as Sec secreted proteins with a Type I SPase (Table 2.5).

Table 2.5: List of 59 identified putative secreted proteins.

No.	Protein ID	Protein name	Length
1	Q5FLN0	SlpX	499
2	Q5FMX4	Putative uncharacterized protein	118
3	Q5FMU8	Putative uncharacterized protein	154
4	Q5FMT3	Putative uncharacterized protein	63
5	Q5FMR8	Putative cellulose synthase (EC 2.4.1.12)	156
6	Q5FMN7	Putative uncharacterized protein	412
7	Q5FMK0	S-layer	457
8	Q5FMJ9	N-acetylmuramidase (EC 3.5.1.28)	409
9	Q5FMJ8	Autolysin, amidase	364
10	Q5FMI7	Putative fibronectin domain	463
11	Q5FMI5	Putative uncharacterized protein	118
12	Q5FMF9	Putative S-layer	172
13	Q5FMF7	Putative uncharacterized protein	282
14	Q5FLP6	Aggregation promoting protein	231
15	Q5FLP5	Putative surface exclusion protein	355
16	Q5FL81	Ribonuclease Y (RNase Y) (EC 3.1.-.-)	543
17	Q5FL54	Putative uncharacterized protein	550
18	Q5FKW0	Cell shape-determining protein MreC (Cell shape protein MreC)	283
19	Q5FKQ1	Putative uncharacterized protein	497
20	Q5FKP7	Putative uncharacterized protein	156
21	Q5FKB8	Penicillin-binding protein	364
22	Q5FKA5	Putative mucus binding protein	2310
23	Q5FK97	Putative surface layer protein	385
24	Q5FK57	Putative uncharacterized protein	217
25	Q5FK50	Putative cell surface protein	202

No.	Protein ID	Protein name	Length
26	Q5FK41	Cystathionine beta-lyase (EC 4.4.1.8)	85
27	Q5FK22	Putative uncharacterized protein	59
28	Q5FJS0	Putative lipase	425
29	Q5FJL6	UPF0154 protein LBA1278	72
30	Q5FJA8	Putative uncharacterized protein	222
31	Q5FJ43	Putative mucus binding protein	339
32	Q5FJ10	Putative fibrinogen-binding protein	264
33	Q5FIW6	Putative uncharacterized protein	171
34	Q5FIU0	Aminopeptidase	505
35	Q5FIT9	Putative surface protein	353
36	Q5FIQ6	D-alanyl-d-alanine carboxypeptidase	432
37	Q5FIP7	Fibrinogen-binding protein	906
38	Q5FIK8	Putative surface protein	685
39	Q5FIF1	Putative uncharacterized protein	169
40	Q5FIC1	Cell wall-associated hydrolase	299
41	Q5FIC0	Cell wall-associated hydrolase	262
42	Q5FIB9	Putative glycosidase	184
43	Q5FHZ4	Putative uncharacterized protein	130
44	Q5FHZ1	Probable NLP-P60 family secreted protein	160
45	Q5FHV9	Lysin	323
46	Q5FIC3	Putative uncharacterized protein	336
47	Q5FI22	Putative aggregation promoting protein	120
48	Q5FM66	30S ribosomal protein S11	130
49	Q5FM23	50S ribosomal protein L11	141
50	Q5FJH6	Ribosomal RNA small subunit methyltransferase B (EC 2.1.1.-)	455
51	Q5FJF1	Oligopeptide ABC transporter substrate binding protein	123
52	Q5FIW8	50S ribosomal protein L35	66
53	Q5FHV5	Penicillin-binding protein	313
54	Q5FML8	DNA-3-methyladenine glycosidase (EC 3.2.2.20)	190
55	Q5FL84	Putative transcriptional regulator	374
56	Q5FKB4	Putative uncharacterized protein	468
57	Q5FIZ1	Putative low temperature requirement A protein	45
58	Q5FIR8	Putative uncharacterized protein	832
59	Q5FJH8	Serine-threonine protein kinase (EC 2.7.1.-)	674

2.4.2. Non-covalent binding/ anchored proteins

The complete set of proteins were searched for possible protein domains which are known to have interaction with the cell wall of the bacteria and thus enables a protein to be non-covalently anchored to the cell wall. A total of 20 sequences were found with different known cell-wall binding domains (Table 2.6). A majority of the proteins had SLAP (PF03217) domain, which is a bacterial surface layer protein domain. The genome of *L. acidophilus* encodes around 16 proteins with SLAP domain.

Table 2.6: List of identified proteins containing cell-wall anchoring/binding domain

No	Protein ID	Protein name	Length	Cell-wall binding domain	Category of secretome
1	Q5FKF7	Putative N-acetylmuramidase (EC 3.5.1.28)	153	LysM (PF01476)	N-terminal anchor protein
2	Q5FMX1	Putative ABC transporter substrate binding protein	330	ABC_sub_bind (PF04392)	N-terminal anchor protein
3	Q5FJA4	Lipoprotein	283	Lipoprotein_9 (PF03180)	N-terminal anchor protein
4	Q5FKL3	Lipoprotein	284	Lipoprotein_9 (PF03180)	N-terminal anchor protein
5	Q5FIF2	Thermostable pullulanase (EC 3.2.1.41)	1185	Alpha-amylase (PF00128); CBM_48 (PF02922); PUD (PF03714); SLAP (PF03217);	N-terminal anchor protein
6	Q5FK97	Putative surface layer protein	385	SLAP (PF03217);	Secretory protein
7	Q5FJE6	Lysin	249	Glyco_hydro_25 (PF01183); SLAP (PF03217);	Intracellular
8	Q5FIZ3	PrtP	1627	fn3_5 (PF06280); Peptidase_S8 (PF00082); SLAP (PF03217);	N-terminal anchor protein
9	Q5FLN0	SlpX	499	SLAP (PF03217);	Secretory protein
10	Q5FMF8	Putative S-layer	177	SLAP (PF03217);	Intracellular
11	Q5FHV9	Lysin	323	Glyco_hydro_25 (PF01183); SLAP (PF03217);	Secretory protein
12	Q5FJ47	Putative surface layer protein	294	SLAP (PF03217);	Lipoprotein
13	Q5FMK0	S-layer	457	SLAP (PF03217);	Secretory protein
14	P35829	S-layer protein (Surface layer protein) (SA-protein)	444	SLAP (PF03217);	N-terminal anchor protein
15	Q5FMJ8	Autolysin, amidase	364	Amidase_2 (PF01510); SLAP (PF03217);	Secretory protein
16	Q5FJQ6	Putative lactocepin S-layer protein (EC 3.4.21.96)	180	SLAP (PF03217);	N-terminal anchor protein
17	Q5FIT9	Putative surface protein	353	SLAP (PF03217);	Secretory protein
18	Q5FJZ4	Lysin	382	Glyco_hydro_25 (PF01183); SLAP (PF03217);	N-terminal anchor protein

No	Protein ID	Protein name	Length	Cell-wall binding domain	Category of secretome
19	Q5FMJ9	N-acetylmuramidase (EC 3.5.1.28)	409	Glucosaminidase (PF01832); SLAP (PF03217);	Secretory protein
20	Q5FMF6	Cell separation protein	599	SLAP (PF03217);	Intracellular

(The cell-wall binding domains or anchoring domains are marked in bold).

2.4.3. Adherence proteins

The proteins sequences were searched for the presence of domains involved in adhesion or binding. Only two proteins were found to be having adhesion domain i.e. Fibronectin type III domain - fn3 (PF00041) and FbpA (PF05833). None of the protein was found to have Mucus binding domain MubP (PF06458), which is very well known to be found in other *Lactobacillus* species in multiple copies with multiple repeats. Although 13 proteins in the genome were annotated as putative mucus binding protein, none of them was found to have mucus binding domain by Pfam search. A potential set of 14 mucus-binding proteins (Table 2.7) were identified by searching the complete genome with an HMM profile generated on the MUB domains of Mub proteins as mentioned in Boekhorst *et al.,* 2006 (Boekhorst, Helmer*, et al.*, 2006).

Table 2.7: List of identified mucus-binding proteins by HMM using MUB profile search.

No	Protein ID	Protein name	Length	Pfam domain	A	B	C
1	Q5FJA7	Mucus binding protein Mub	4326	YSIRK_signal (PF04650)	Y	2	LPxTG protein
2	Q5FKA5	Putative mucus binding protein	2310	YSIRK_signal (PF04650)	Y	1	Secretory protein
3	Q5FKA6	Putative mucus binding protein	2650	YSIRK_signal (PF04650)	N	2	LPxTG protein
4	Q5FIQ0	Putative mucus binding protein	643	-	N	1	C-terminal anchor protein
5	Q5FKK8	Putative mucus binding	508	-	N	0	Intracellular
66	Q5FIF3	Mucus binding protein Mub	1208	YSIRK_signal (PF04650)	Y	1	LPxTG protein
7	Q5FIL0	Mucus binding protein; Mub	1174	-	Y	1	LPxTG protein
8	Q5FJ43	Putative mucus binding protein	339	-	N	0	Intracellular
9	Q5FKA8	Putative mucus binding protein	294	-	N	0	Intracellular
10	Q5FKA7	Putative mucus binding protein	346	-	N	1	LPxTG protein

No	Protein ID	Protein name	Length	Pfam domain	A	B	C
11	Q5FJC2	Putative mucus binding protein	1017	-	N	0	Intracellular
12	Q5FKA9	Putative mucus binding protein	185	Dipeptidyl-peptidase IV (PF00930)	N	0	Intracellular
13	Q5FJS1	Putative mucus binding protein	697	-	N	0	Intracellular
14	Q5FIC2	Putative membrane protein	1376	DUF285 (PF03382); Gram_pos_anchor (PF00746); YSIRK_signal (PF04650);	Y	2	LPxTG protein

A: Signal peptide prediction; **B:** TM segment; **C:** Category of secretome.

Six putative mucus-binding proteins were found to be LPxTG motif-containing proteins. One protein was of secretory prediction, while six proteins were found intracellular with no detectable signal peptide and transmembrane segment. Other proteins apart from intracellular were found to have 1 or 2 transmembrane segment. Interestingly one of the proteins was found to have a single transmembrane domain at C-terminal without signal peptide sequence and is classified under C-terminal anchor proteins. Few of them contain the YSIRK signal sequence domain which is signal sequence found in other gram-positive bacteria just before the transmembrane. A comprehensive listing of the proteins in the secretome based on above classification is given below (Table 2.8).

Table 2.8: List of predicted secretome proteins with their functional classification based on the protein domain.

Functional class	LPxTG anchor [12]	Lipoproteins [41]	N-terminal anchor [103]	C-terminal anchor [0]	Secreted (release) [54]	TOTAL [210]
Surface/membrane protein	4 (1)	2 (2)	8 (6)	0	8 (1)	**22**
Enzyme	0	5 (3)	43 (18)	0	14 (1)	**62**
Adherence	6 (3)	6	4	1 (1)	5	**22**
Regulator	0	0	8	0	6	**14**
Transporter	0	19 (3)	3 (2)	0	1	**23**
Unknown/Uncharacterized	2 (2)	10 (10)	42 (42)	1 (1)	25 (25)	**80**
TOTAL	**12**	**42**	**108**	**2**	**59**	**223**

Numbers in parenthesis () indicated number of putative protein. Prediction in each category through LocateP database (Zhou et al., 2008) is shown in [].

2.4. Discussion

The secretome of *Lactobacillus* contained two main categories of proteins (Figure 2.1) the secreted proteins which are released from the cell and the surface associated proteins which are further categorised into several sub-categories.

Our analysis of the *L. acidophilus* genome predicts to encode 223 extracellular proteins (~12% of its total proteins) as putative secretome members as compared to 210 proteins identified through LocateP database (Zhou *et al.*, 2008) (Table 2.8). Few of the earlier identified members of secretome from LocateP are not included in our prediction viz. few from our findings are not annotated in the previous studies too (Zhou *et al.*, 2008). Twelve proteins possess a C-terminal LPxTG anchoring motif (Boekhorst *et al.*, 2005) for covalent attachment to cell wall, while 42 of these secretome proteins contain an N-terminal lipobox motif which is a common mechanism for protein secretion and their attachment to membrane through covalent binding of a conserved cysteine residue (Hutchings *et al.*, 2009). A total of 20 proteins have been found to contain a cell-wall binding domain for non-covalent attachment to the cell wall, and the majority of them contain a surface layer protein domain (SLAP). A single protein was also found having LysM domain for non-covalent attachment to peptidoglycan layer (Buist *et al.*, 2008). The remaining 108 of extracellular proteins were found to be N-terminal anchor proteins due to lack of signal cleavage site in the signal peptide. Finally, the residual 59 proteins are predicted to be either released (i.e. secreted) or associated with cell wall with another unknown mechanism. From the total of 223 proteins predicted in secretome, 80 of the proteins are uncharacterized, and their function is yet to be determined. A graphical representation of 223 predicted proteins category wise is mentioned below (Figure 2.5). Though the rest of the proteins were assigned a functional classification based on the Pfam domain, 54 proteins were still annotated as the putative whose function needs to be confirmed (Table 2.8). From the functional annotation, the highest number of proteins was uncharacterized, while the second largest group was of enzymes which may play a role in the secretion of the protein, maintaining bacterial cell wall as well as modification and degradation of extracellular compounds for the source of nutrients. A large number of combined surface/membrane and adhesion proteins were also encoded by genome supporting its ability and the requirement for bacterial adhesion to the gut. Sixteen proteins were found to have adherence binding domain, of which 14 are having mucus binding domain while 2 have fibronectin binding domain.

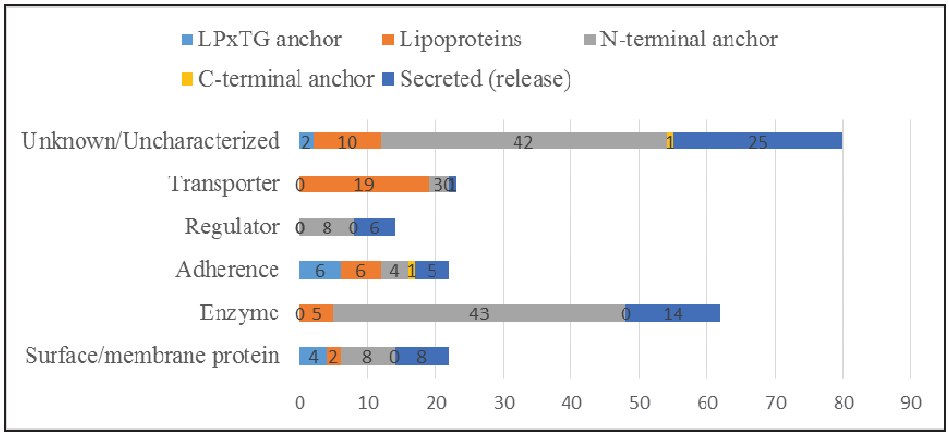


Figure 2.5: Graphical representation of predicted 223 proteins by category-wise. Numbers in the bar indicated the total number of proteins in the respective category.

Of the 223 proteins predicted, 12 proteins were found to have LPxTG motif responsible for cell wall anchoring. The motif is recognised by a transpeptidase sortase (SrtA) enzyme. The *L. acidophilus* genome encodes a single sortase (SrtA) (Q5FJP7) which was identified by HMM profile of known sortases. Out of the 12 LPxTG, six proteins were found to be of adherence class, while four were found to be surface associated proteins with two proteins of unknown function. The majority of predicted LPxTG proteins were mucus-binding proteins with varying length from 4326 to 346 residues supporting the occurrence of multi-repeats of the mucus binding domain (Table 2.1). Though only 8 LPxTG sequences were found with a signal peptide with cleavage site, the remaining sequences were still annotated as LPxTG, due to the presence of LPxTG motif at C-terminal followed by hydrophobic transmembrane residues and positively charged tail (Table 2.1).

In *L. acidophilus* genome, 42 lipoproteins were identified which are a second largest group of membrane-anchored proteins. The majority of the 42 identified lipoproteins by the *L. acidophilus* constitute the substrate binding proteins of ABC transporters, but also some proteins that are involved in adhesion, surface activity, enzymatic and protein secretion (Table 2.8). In the prediction set, three proteins: Q5FIC3, Q5FIY2 and Q5FI22 were also detected by lipoprotein HMM and signature motif, but with SPI site instead of SPII cleavage site. Another protein: Q5FHQ4 was also identified by HMM profile and had multiple TM segments. All these above four proteins were not annotated further as lipoproteins. Among the *Lactobacillus* genus, the *L. acidophilus* genome encodes the highest number of lipoproteins next to *L. plantarum* and

L. casei. One of the probable explanation for the *Lactobacillus* genome to encode a higher number of lipoproteins is to allow functionally diverse class of peripheral membrane proteins. Around ~45% of the total lipoproteins are of solute binding proteins of ABC transporter systems. These are an important class of importers which helps in a wide range of substrate transport and contributes to the ability of bacteria to acquire diverse nutrients and substrates from the environment (Hutchings *et al.*, 2009).

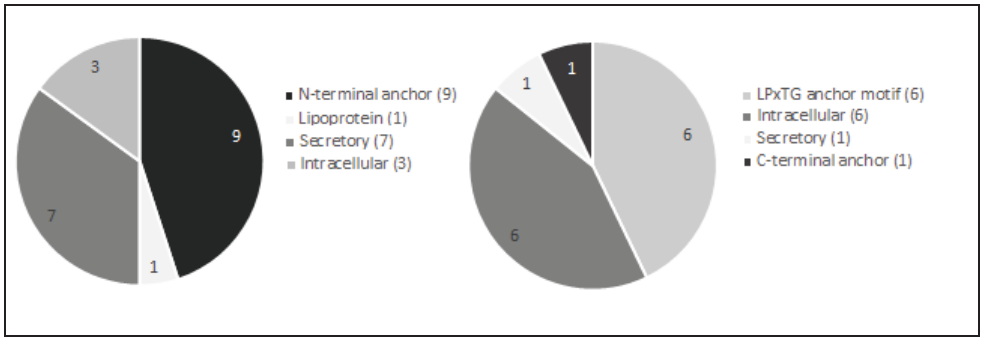
The N-terminal anchored proteins form the highest number of proteins in predicted secretome. A maximum number of the proteins in this class is of enzyme functionality followed by uncharacterized proteins. The majority of the proteins that are predicted to be N-terminally anchored contains typical extracellular domains which may support the wide range of functionalities perform at the cell surface. These proteins are found to be mainly involved in transport, cell-envelope metabolism, signalling, protein sorting, cell division, etc. (Table 2.3). Few surface proteins with S-layer domains are also N-terminally anchored. A Sortase (SrtA) (Q5FJP7) protein was also identified to be N-terminally anchored which recognises the LPxTG proteins, cleaves the site between T and G residues and covalently attaches the threonine carboxyl group to the peptidoglycan layer. A set of 7 penicillin-binding proteins was also found in the N-terminal anchor proteins which are the critical class of proteins involved in the final stage of peptidoglycan synthesis. A fibronectin binding domain 3 containing protein (Q5FIZ3) was also found which may participate in the adherence with the extracellular matrix component fibronectin and thus facilitating adhesion. A putative biofilm-associated protein (Q5FJC1) was also found to be N-terminally anchored proteins. Amongst the *Lactobacillus* genus, the *L. acidophilus* genome encodes the highest number of N-terminally anchored proteins next to *L. plantarum* and *L. casei.* Only two proteins were found to be C-terminally anchored: Q5FJ90-Putative uncharacterized protein and Q5FIQ0- Putative mucus binding protein. The first one is an uncharacterized protein while the latter is a putative mucus binding protein with no detectable LPxTG motif and neither mucus binding domain using HMM profile. The C-terminal anchoring might be a way of expressing the protein for adherence on the cell surface.

The *L. acidophilus* genome encodes 59 proteins which are released in to the environment or secreted through Sec pathway after Type I SPase cleavage. The majority of the secreted proteins are of the unknown function followed by enzymes. Few proteins like mucus binding protein, fibronectin binding proteins and S-layer proteins are a

secretory protein which might play a crucial role in adherence. A wide variety of proteins is found in secretory proteins with vivid protein domains supporting a great diversity in the functionality of the bacteria (Table 2.3).

The genome of *L. acidophilus* also encodes 20 proteins which are found with domains involved in non-covalently binding with the cell wall (Table 2.6). The SLAP domain was found in 16 proteins which among the *Lactobacillus* genus is the highest number of S-Layer Protein domain in *L. acidophilus*. Most of the *Lactobacillus* genome except *L. delbrueckii* and *L. Helvetica* do not encode for SLAP proteins. Most of the proteins also had another domain which might be essential for other functionality of the protein. Apart from SLAP domain, other domains in association with S-layer-like SLH (PF00395), S_layer_N (PF05123) and S_layer_C (PF05124) were not found in the *L. acidophilus* genome. None of the other binding domains as found in the *L. plantarum* were found in the *L. acidophilus* such as PG_binding_1 (PF01471), CW_binding_1 (PF01473), MucBP (PF06458), SH3_3 (PF08239) and collagen binding domain (Boekhorst, Wels*, et al.*, 2006). A few proteins containing bacterial solute binding proteins as well as bacterial periplasmic binding protein were also present. A single protein with LysM domain (Q5FKF7) was present in *L. acidophilus*. The LysM domain containing protein are found to be involved in peptidoglycan binding protein. From the 20 identified proteins, 9 were N- terminally anchored, 7 were secretory proteins, 1 is lipoproteins and the rest 3 are intracellular proteins (Figure 2.6a) (Table 2.6). Though intracellular proteins are not part of secretome, the presence of these cell-wall binding domains indicates an alternate mechanism for secretion of such proteins.

Fourteen mucus binding protein were identified through HMM profile search, although none of them had mucus binding domain annotated by Pfam. All of the proteins are annotated as putative mucus binding protein except one which is annotated as a putative membrane protein. The length of the predicted proteins varies from 4396 to 185 amino acids, which is supporting the fact that mucus binding protein has multiple repeats of the Mub – mucus binding domain. All the predicted proteins had at least 2 Mub domain except for Q5FKA7 which has only one domain.



(a) (b)

Figure 2.6: Distribution of predicted proteins with non-covalent binding domain and MUB domain: (a) Distribution of proteins with non-covalent cell-wall binding domain into different prediction category; (b) Distribution of proteins with mucus binding protein into different prediction category.

Few of the sequence has YSIRK_signal (PF04650) at the start of the transmembrane region and all they had LPxTG anchor suggesting a strong co-relation between YSIRK signal and LPxTG anchor motif. Some of the sequences contain a signal peptide, and few of them has a transmembrane domain. Six out of the 14 proteins had LPxTG motif, six as intracellular proteins, 1 with C-terminal domain and 1 is secretory (Figure 2.6b). The absence of signal sequence or any other secretory site in the intracellular proteins suggests an alternate pathway for protein secretion or anchoring.

Genus *Lactobacillus* is a heterogeneous group of bacteria which display high variation in their habitat as well as molecular factors. This variation reflects the variation of bacterial cell surface properties that might be a key element for the functioning of bacteria on colonisation communication with the host environment. Current high throughput sequencing have greatly enhanced the understanding of the extracellular biology of this genus, though many of the proteins targeted to the extracellular surface lack a functional annotation. The extracellular and surface exposed proteins which referred as secretome of bacterium plays a vital role in its interaction with the host's environment. An extensive bioinformatics analysis of all *L. acidophilus* proteins was done to evaluate the surface-exposed and secreted proteins. All of them were classified into different categories of secretome and further classified according to their functional class based on domains.

A well-known LPxTG protein identified as a mannose-specific adhesin (lp_1229) is a lectin-like protein fulfilling the critical role of adhesion by bacteria to ECM (Pretzer

et al., 2005). Recently, a fibrinogen binding adhesin containing LPxTG motif anchor was identified from *Streptococcus agalactiaie* which has an essential role in adhesion (Buscetta *et al.*, 2014). Lipoproteins are the well-identified class of proteins directly involved in adhesion and colonisation (Kohler *et al.*, 2016). Thus a large number of predicted lipoproteins in *L. acidophilus*, particularly the one classified as adherence proteins, are worthwhile to investigate. Our analysis revealed that the *L. acidophilus* encodes two protein for fibronectin binding and 1 for fibrinogen binding which is a very important target for further study as fibronectin binding protein are earlier known as a protein involved in adhesion (Christie *et al.*, 2002). As mucus binding domain containing proteins are one of the best-characterized adhesin in gut bacteria (Jensen *et al.*, 2014), the 13 proteins with mucus binding domain identified by HMM profile in our secretome prediction are a good candidate to undertake adhesion studies. We believe that functional analysis of few proteins from above-predicted secretome with the extracellular and unusual domain is worthwhile investigating as it will enhance our understanding towards the unusual mechanism of secretion and important unexplored functions of novel domains.