International Journal of Research and Review
www.ijrrjournal.com E-ISSN: 2349-9788; P-ISSN: 2454-2237

Original Research Article

Functional Annotation of Uncharacterized Proteins of *Listeria Monocytogenes*

Parvinder Kaur[1*], Rhythm Gandhi[1*], Harbinder Kaur[1], Sana Gupta[1*], Ruchi Sachdeva[2*]**

[1]Graduate, [2]Assistant Professor,
[*]Dept of Bioinformatics, Goswami Ganesh Dutta Sanatan Dharma College, Sector 32-C, Chandigarh, India
[**]School of Computational and Integrated Sciences, Jawaharlal Nehru University, New Delhi, India.

Corresponding Author: Ruchi Sachdeva

ABSTRACT

Listeria monocytogenes is a gram-positive, non-spore forming, facultatively anaerobic rod which is found in soil, water samples, silage, sewage, slaughterhouse waste, milk of normal and mastitic cows, human and animal faeces. *L. monocytogenes* has been implicated as the causative agent in several outbreaks of food-borne listeriosis. Recently sequenced genome of *L. monocytogenes* J1-220 has large number of protein encoding genes annotated as hypothetical proteins. Thus, keeping this in mind, we attempted to predict the functions of 30 randomly selected hypothetical proteins of *L. monocytogenes* J1-220 genome with the help of various analysis including domains prediction, remote homology search, fold recognition, investigation of transmembrane helices and functional partners. Functions of five proteins were successfully assigned based on the consistent predictions obtained from different analysis. A strong agreement was obtained regarding different structural and functional aspects of these proteins, thereby validating the annotation. These proteins were found to play critical roles in cellular processes such as membrane transport, signal transduction, host interaction and catalysis. Gene products unique to *L. monocytogenes* for which no function could be identified were eleven. For five proteins, partial information was obtained based on the results obtained from single analysis. Thus, functional annotation of hypothetical proteins has enhanced the *L. monocytogenes* genomic information and provided a useful basis for experimental design aiming to understand the mechanism of pathogenesis and drug target identification.

Keywords: *hypothetical proteins, prediction, listeriosis, domains, fold recognition.*

INTRODUCTION

Listeria monocytogenes is a gram positive, non-spore forming, facultative anaerobe, rod-shaped bacteria. The bacterium is 0.5 µm in width and 1-1.5 µm in length. It has been found in silage, sewage, slaughterhouse waste, milk of normal and mastitic cows, sheeps, goats and poultry, but infrequently from wild animals. [1] *L. monocytogenes* has been recognized as an opportunistic intracellular pathogen that causes food borne infections in humans worldwide. [2,3] Out of the three serotypes of *L. monocytogenes* (1/2a, 1/2b/ and 4b), the 4b is responsible for the vast majority of listeriosis outbreaks. *L. monocytogenes* causes severe problems especially in pregnant women, neonates, elderly and immunocompromised individuals. Listeriosis patients often develop non-specific flu-like symptoms (fatigue, chills, headache, and pain) and gastroenteritis. Without proper treatment, it may lead to septicaemia, meningitis, abortion and, in some cases, death. [4]

Some virulence associated proteins are involved in *L. monocytogenes* virulence and pathogenicity. The genome sequence of *L. monocytogenes* J1-220 has recently been determined by next generation sequencing. [5] Its genome is circular in shape with the size of 3032271 bp. The nucleotide sequences were annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline. It contains approximately 3046 genes, 67 tRNAs. [5] There are around 680 hypothetical proteins encoded by *L. monocytogenes* genome. No function has been assigned to these proteins. However, complete understanding of structure and functions of proteins encoded by *L. monocytogenes* genome is essential to obtain insights into its pathogenicity.

Recently, functional annotation of pathogen *Mycobacterium tuberculosis* has been carried out. [6] It has resulted in successful assignment of functions to 95 % of *M. tuberculosis* proteins. Thus, in order to improve the quality of *L. monocytogenes* proteome information, present work was carried out to predict the functions of a few hypothetical proteins of *L. monocytogenes.*

METHODOLOGY

Acquisition of Sequences for Analysis

The complete genome of *Listeria monocytogenes* J1-220 is contained in NCBI Nucleotide sequence database having accession number CP006046. Sequences of thirty randomly selected hypothetical proteins of *L. monocytogenes* were retrieved in FASTA format from the corresponding NCBI protein sequence database. Following is the list of accession umbers of the sequences retrieved.

EGF38885.1, AGR29492.1, EGF38896.2, EGF38908.2, ALD11032.1, ALD11035.1, EGF38918.2, EGF38919.1, EGF38940.1, EGF38946.1, ALD11042.1, ALD11043.1, ALD11044.1, EGF38952.1, EGF38822.1, EGF38823.2., AGR29694.1, EGF38872.2, EGF38830.2, ALD11052.1, ALD11068.1, EGF38733.1, AGR29685.2, EGF38749.2, EGF38751.2, EGF38752.1, EGF38753.1, AGR29482.1, EGF38756.1, EGF38762.2.

Analysis

All the retrieved hypothetical proteins were subjected to several different types of analysis. FASTA sequences were provided as input to all the tools used. Each of the retrieved protein sequences were searched against protein families databases to predict the presence of any functional domain. Three tools were used: Pfam (version 29.0), [7] InterProScan [8] and NCBI CD Search. [9] The hypothetical proteins were analyzed in terms of the presence of transmembrane helices using TMHMM. The structural class of these proteins was predicted using 1D server available at http://biomine.ece.ualberta.ca/1D/1D.html. The Protein sequences were subjected to fold recognition using PHYRE 2. [10] Only high confidence matches i.e. > 90% were considered to ensure correctness of the recognized fold. Functional partners of hypothetical proteins were predicted using STRING version 10. [11] *L. monocytogenes* Scott was selected in the Organism box.

RESULTS AND DISCUSSION

Domains Assignment

Domains predicted in the hypothetical proteins are mentioned in table 1. The proteins for which no domains were predicted are not included in the table. Hypothetical protein EGF38885.1 of *L. monocytogenes* was predicted to contain PIN domain in the region 165-291 amino acids and TRAM domain in the region 292 to 353 (Table 1). The PIN domains are small protein domains of ~130 amino acids. PIN domain functions to cleave single stranded RNA in sequence specific manner. In prokaryotes, PIN domain proteins are the toxic components of toxin-antitoxin systems. EGF38885.1 also contains TRAM domain (Table 1). The TRAM domain adopts a beta-barrel fold. It is found in two distinct classes of tRNA-modifying enzymes of the TRAM2 family and enzymes of the miaB family and in a family of small uncharacterised archaeal proteins that may have a role in the regulation of tRNA modification and translation. [12] The

TRAM domain can be found alone or in association with other domains. The TRAM domain is predicted to bind tRNA and deliver the RNA-modifying enzymatic domain to their targets. This data suggests that EGF38885.1 protein may function as a ribonuclease enzyme.

Table 1: Assignment of domains

S.No.	Accession Number	Pfam	InterPro	InterPro Domain Region	CD Search	CD Domain Region
1.	EGF38885.1	no domain	PIN domain TRAM	163-292 292-353	PIN YacL	165-291
2.	AGR29492.1	YacP	no domain	no domain	YacP	1-170
3.	EGF38896.2	Imm63	Imm63	42-130	Imm38	2-133
4.	EGF38918.2	LRR_4 LRR_4	LRR Copper Resistance protein LRR LRR4	54-273 272-336 101-122 123-146 167-192 103-142	LRR_4 LRR_4 LRR_4 LRR_4 LRR_4 LRR_4 LRR	102-144 168-210 146-187 190-232 213-253 234-276 32-350
5.	EGF38919.1	SMI1_KNR4	SMI1/KNR4	32-198	SMI1_KNR4	41-81
6.	EGF38940.1	YycI	YycH	39-251	YycI	1-274
7.	EGF38946.1	DUF998	no domain	no domain	DUF998	13-144
8.	EGF38822.1	SfLAP	no domain	no domain	DUF2910	8-219
9.	EGF38823.2	Bph1	Bacterial Pleckstrin homology domain	3-129	Bph_1	6-126
10.	AGR29694.2	MucBP	LRR LPXTG MucBP	39-177 1278-1313	C-term_anchor LRR_4 LRR_8	290-346 72-112 49-104
11.	EGF38830.2	DUF3130	no domain	no domain	DUF3130	1-89
12.	ALD11052.1	LXG	LXG	1-231	LXG	2-201
13.	EGF38733.1	AzlC	no domain	no domain	AzlC	16-153
14.	AGR29685.2	no domain	Glycoside hydrolase	110-177	HlyD_2	3-198
15.	EGF38751.2	Glyoxalase	Glyoxalase	1-126	Glo_EDI_BRP_LIKE_3	2-125
16.	EGF38752.1	DUF1149			DUF1149	1-124
17.	EGF38753.1	NusG II	NusG II	15-125	Lin431	43-126
18.	AGR29482.1	no domain	no domain	no domain	SPFH_like_u4 TolA_full	170-209 169-213
19.	EGF38762.2	DUF1211	no domain	no domain	DUF1211	1-176

YacP domain was predicted in hypothetical protein AGR29492.1 (Table 1). Proteins containing YacP are suggested to be nucleases because of the presence of a NYN domain (i.e.YacP-like Nuclease). [13] Bacterial YacP proteins interact with the Ribonuclease III and TrmH methylase in a processome complex that catalyzes the maturation of rRNA and tRNA. Thus, the protein AGR29492.1 is predicted to be a nuclease enzyme. The hypothetical protein EGF38896.2 was predicted to contain immunity (IMM) proteins domain in the region 2-133 amino acids. All the three tools predicted the presence of IMM domain in this protein. IMM domain exhibits alpha+beta fold and a conserved E+G and ExxY motifs. [14] A key feature that distinguishes the polymorphic toxins from conventional toxins is the presence of immunity proteins. Based on this data, EGF38896.2 is predicted to be immunity protein that is a part of polymorphic toxin system of *L. monocytogenes*.

The hypothetical protein EGF38918.2 of *L. monocytogenes* was predicted to contain leucine rich repeats (LRR) and copper resistance protein. Leucine rich repeats are short sequence motifs present in a number of proteins with diverse functions and cellular locations. [15] These repeats are usually involved in protein-protein interactions. Each Leucine Rich Repeat is composed of a beta-alpha unit. Leucine rich repeats are abundantly found in internalin proteins which can bind to mammalian receptors, such as E-cadherin. As can be seen from table 1, EGF38918.2 is predicted to constitute another domain: Copper resistance protein

CopC/ internalinimmunoglobulin-likedomain. This domain is found in two proteins: Copper- resistance proteins CopC and PcoC. This domain is also found in Internalin proteins InlA, InlB and InlH which are members of a family of *Listeria* cell surface proteins from the opportunistic pathogen *L. monocytogenes*. [16] The N-terminal regions of internalins consist of a central LRR region flanked by an EF-hand-like cap on one end, and an immunoglobulin-like fold on the other. Together these regions form a domain that directs host cell-specific invasion. Consistent with this observation this protein comprises of leucine rich repeats as predicted by all the three tools. Thus, EGF38918.2 is predicted to be internalin protein. *L. monocytogenes* gains entry into the host cells through interaction between internalin and E-cadherins present on host cell surface. [17,18] Such protein-protein interaction may be mediated through leucine rich repeats. Internalins enable *L. monocytogenes* to evade host immune functions and hence serve as virulence associated protein.

SMI1_KNR4 domain was predicted in EGF38919.1 by all the three tools (Table 1). The SMI1_KNR4 domain is found in the yeast cell wall assembly regulator SMI1 and the cell proliferation protein KNR4. [19] *Saccharomyces cerevisiae* protein SMI1 has a regulatory role in chitin deposition and in cell wall assembly. This data indicates that EGF38919.1 protein may be involved in cell wall assembly and cell proliferation.

In case of protein EGF38940.1, Pfam and CD predicted the presence of YycI domain (Table 1). While according to InterPro, it contains YycH domain. Proteins harboring the YycI and YycH domains are a part of YycFG two-component system which is the only signal transduction system in *Bacillus subtilis*. [20] This system is highly conserved in low GC Gram-positive bacteria and regulates important processes such as cell wall homeostasis, cell membrane integrity, cell division and hence is essential for cell viability. [20] Both YycH and YycI are always found in pair on the chromosome, downstream of the essential histidine kinase YycG. [21] Additionally, both proteins share a function in regulating the YycG kinase with which they appear to form a ternary complex. Lastly, the two proteins always contain an N-terminal transmembrane helix and are localized to the periplasmic space.

YycI and YycH proteins interact to control the activity of the YycG kinase. [22] Both YycI and YycH proteins are localized outside the cytoplasm and attached to the membrane by an N-terminal transmembrane sequence. YycH and YycI control the activity of YycG in the periplasm and that this control is crucial in regulating important cellular processes. Thus, the hypothetical protein EGF38940.1 of *L. monocytogenes* is predicted to be an essential component of YycFG two-component system and may be involved in regulating cellular processes such as cell wall homeostasis, cell membrane integrity, and cell division. The YycFG two-component system is highly conserved in gram positive bacteria. This observation strongly validates the presence of such system in *L. monocytogenes* which happens to be gram positive.

Protein EGF38822.1 contains SsfLAP domain and DUF2910 domain (Table 1). SsfLAP (SAP Sulphoid 1 addressing protein) is a transmembrane transport protein with six predicted transmembrane helices and a hydrophilic domain between helices 3 and 4. [23] SAP also belongs to the LysE protein super family, whose members have been implicated in small molecule transport in bacteria. SsfLAP domain is specifically involved in the transport of sulfolipid-1 across the membrane. This prediction suggests that EGF38822.1 is a transmembrane protein that is involved in the transport of small molecules such as sulfolipid across the cell membrane.

Bph (Bacterial Pleckstrin homology) domain was predicted in hypothetical protein EGF38823.2 in the region 6-126 amino acids. Bph domain is a Pleckstrin homology domain (PH domain) of approximately 120 amino acids that occurs in a wide range of proteins involved in intracellular signaling or as constituents of the cytoskeleton. [24] This domain can bind phosphatidylinositol lipids of plasma membranes and proteins such as the βγ-subunits of heterotrimeric G proteins and protein kinase C. With the help of these interactions, PH domain plays a role in recruiting proteins to different membranes, thus targeting them to appropriate cellular compartments. This data indicates the possible functional role of EGF38823.2 protein in intracellular signaling. Another hypothetical protein ALD11052.1 contains LXG domain which is present in the N-terminal region of a group of polymorphic toxin proteins in bacteria. It is predicted to use Type VII secretion pathway to mediate export of bacterial toxins. [25] This prediction suggests that the protein ALD11052.1 seems to be implicated in the extrusion of bacterial toxins.

Protein EGF38733.1 was predicted to contain AzlC (4-azaleucine) domain in the region of 16-153 amino acids. AzlC domain has 5 potential transmembrane motifs and is predicted to be part of a branched-chain amino acid transport system. Protein AGR29685.2 was predicted to contain hlyD_2 domain and O-Glycosidase hydrolase domain (Table 1). HlyD (hemolysin D family secretion protein) is a member of a large family of polypeptides, the MFP (membrane fusion protein) family, proposed to span the periplasm linking the inner and outer membranes. MFPs, although involved in the export of a variety of compounds, from drug molecules to large polypeptides. Proteins belonging to the MFP family, such as HlyD, are characterized by a single transmembrane domain (TMD), followed by a large helical domain and a C-terminal domain, predicted to be composed largely of β strands. Another domain predicted in this protein is O-Glycosyl hydrolase. Proteins containing this domain are a widespread group of enzymes that hydrolyse the glycosidic bond. This entry represents the catalytic TIM beta/alpha barrel common to many different families of glycosyl hydrolases.

Protein EGF38751.2 was predicted to contain Glyoxalase domain in the region of 1-126 amino acids by Pfam and InterPro (Table 1). Glyoxalase domain is found in Glyoxalase I (lactoylglutathionelyase) that catalyzes the first step of the glyoxal pathway. S-lactoylglutathione is then converted by glyoxalase II to lactic acid. According to CD search, EGF38751.2 protein contains GLO_EDI_BRP domain that is found in a variety of structurally related metalloproteins, including the glyoxalase I. Thus, based on domain assignment, the EGF38751.2 protein can be annotated to be a glyoxalase enzyme.

EGF38753.1 was predicted to contain NusG II domain and Lin0431 domain (Table 1). NusG II domain consists of two-sandwiched four-stranded antiparallel sheets. It is found in transcription factor NusG. [26] The function of this domain is unknown. Lin0431_like (Listerrriainnocua Lin0431) is similar to the N-Utilization Substance G (NusG) N terminal (NGN) insert. Based on its homologous sequences, Lin0431_like domain is predicted to bind negatively charged nucleic acids and/or another anionic binding partner, suggesting a possible role in transcription/translation regulating functions.

Hypothetical Protein AGR29482.1 was predicted to contain SPFH_like_u4 domain and TolA domain (Table 1). SPFH_like_u4 (stomatin, prohibitin, flotillin, and HflK/C) superfamily is summarized as uncharacterized. [27] Individual proteins of the SPFH superfamily may cluster to form membrane microdomains which may in turn recruit multiprotein complexes. TolA domain was also predicted in AGR29482.1. TolA is an inner membrane protein that is involved in the transport of colicins and filamentous DNA, and is implicated in pathogenesis. [28] Based on this prediction, the protein AGR29482.1 is proposed to play a role in transport mechanism across the cell membrane.

In five hypothetical proteins of *L. monocytogenes*, DUF (Domain of unknown function) domain was predicted (Table 1). DUFs, are a large set of families within the Pfam database that do not include any protein of known function. [29] However, there were eleven hypothetical proteins in which no domain could be predicted by any of the three tools used. These proteins may represent a family of uncharacterized proteins that are unique to *L. monocytogenes*.

Table 2: Prediction of Tran membrane Helix (TM) and Structural Class

S. No.	Accession No.	TM Helix	Region	Structural Class

1.	EGF38885.1	4	TMhelix(5-27)
			TMhelix(47-69)
			TMhelix(76-98)
			TMhelix(113-132)
				α/β

2.	AGR29492.1	0	N.A	α/β
3.	EGF38896.2	0	N.A	α+β
4.	EGF38908.2	0	N.A	β
5.	ALD11032.1	1	TMhelix(6-28)	β

6.	ALD11035.1	3	TMhelix(5-27)
			TMhelix(34-54)
			TMhelix(74-96)
				β

7.	EGF38918.2	0	N.A	α
8.	EGF38919.1	0	N.A	α+β
9.	EGF38940.1	1	TMhelix(7-26)	α/β

10.	EGF38946.1	6	TMhelix(7-29)
			TMhelix(54-76)
			TMhelix(83-105)
			TMhelix(120-142)
			TMhelix(155-177)
			TMhelix(187-206)
				α/β

11.	ALD11042.1	4	TMhelix(7-28)
			TMhelix(43-65)
			TMhelix(86-108)
			TMhelix(112-131)
				α/β

12.	ALD11043.1	0	N.A	α/β
13.	ALD11044.1	0	N.A	β
14.	EGF38952.1	0	N.A	α

15.	EGF38822.1	6	TMhelix(10-32)
			TMhelix(39-61)
			TMhelix(71-90)
			TMhelix(122-144)
			TMhelix(159-181)
			TMhelix(194-216)
				β

16.	EGF38823.2	0	N.A	β
17.	AGR29694.2	1	TMhelix(1286-1308)	α
18.	EGF38827.2	0	N.A	α/β
19.	EGF38830.2	0	N.A	α+β
20.	ALD11052.1	0	N.A	α+β
21.	ALD11068.1	0	N.A	β

22.	EGF38733.1	5	TMhelix(12-34)
			TMhelix(59-81)
			TMhelix(127-149)
			TMhelix(164-186)
			TMhelix(193-215)
				α/β

23.	AGR29685.2	0	N.A	β

24.	EGF38749.2	3	TMhelix(7-29)
			TMhelix(39-61)
			TMhelix(73-102)
				α+β

25.	EGF38751.2	0	N.A	α+β
26.	EGF38752.1	0	N.A	β
27.	EGF38753.1	1	TMhelix(13-35)	β

28.	AGR29482.1	2	TMhelix(5-24)
			TMhelix(37-56)
				α/β

29.	EGF38756.1	0	N.A	α

30.	EGF38762.2	5	TMhelix(12-34)
			TMhelix(39-61)
			TMhelix(74-96)
			TMhelix(101-123)
			TMhelix(143-174)
				α

Prediction of Transmembrane (TM) Helices and Structural Class

Hypothetical proteins of *L. monocytogenes* were analyzed for the presence of TM helix. Structural fold in these proteins was also predicted. Table 2 summarizes the results of these predictions. Interestingly, EGF38940.1 protein was predicted to contain one TM helix (Table 2). As mentioned previously, this protein was predicted to contain YycI domain. Usually protein harboring this domain attach to the membrane by an N-terminal TM sequence, consistent with the prediction of one TM helix near N-terminus by TMHMM.

According to our analysis, protein EGF38896.2 was predicted to be immunity protein and contain α+β structural fold (Tables 1 & 2). It is well known that all immunity proteins have an alpha+beta fold and a conserved E+G and ExxY motifs. Thus, there is a strong correlation between the predictions on EGF38896.2 made by two different tools, thereby validating the annotation.

Based on domains analysis, EGF38822.1 may function as transmembrane transport protein (SsfLAP, SAP Sulphoid 1 addressing protein) and involved in the transport of small molecules across the cell (Table 1). Concurrent with this prediction, TMHMM indicated the presence of six TM helices in EGF38822.1 protein of *L. monocytogenes* (Table 2). EGF38751.2 was predicted to contain glyoxalase domain which belongs to alpha and beta class of SCOP. Domain prediction is consistent with the presence of α+β structural fold as analyzed by 1D server (Table 2). According to domain analysis, AGR29482.1 contains TolA domain which is found in inner membrane proteins (Table 1). Consistent with this prediction, the protein AGR29482.1 possesses two TM helices and is proposed to be an integral membrane protein (Table 2). EGF38733.1 protein was predicted to contain five TM helices (Table 2). Thus, this prediction is compatible with the TM helix containing AzlC domain earlier predicted (Table 1).

However, in case of some proteins predictions were not compatible with each other. For example, based on domain assignment EGF38885.1 was predicted to be a ribonuclease enzyme but four TM helices were also predicted in this protein (Table 2). Another example is EGF38918.2 which contains leucine rich repeats which comprise of α/β horseshoe fold (Table 2). But according to 1D server, this protein contains α fold. There is discrepancy in the predictions on AGR29685.2 which contains hlyD_2 and O-Glycosidase hydrolase domains (Table 1). Both of these domains comprise alpha helices and beta sheets. However, AGR29685.2 was predicted to contain beta fold (Table 2). Moreover, hlyD_2 domain contains a transmembrane region. But our analysis indicated the absence of any TM helix (Table 2). As mentioned earlier, EGF38753.1 contains NUS G II domain composed of two-sandwiched four-stranded antiparallel sheets (Table 1). Consistent with this prediction, 1D server has predicted β structural class for this protein (Table 2). However, presence of TM helix in EGF38753.1 cannot be related with the domains predicted (Tables 1 & 2). Thus, predictions in these proteins could not be validated and hence further analysis is required.

Searching of similar sequences

Hypothetical proteins of *L. monocytogenes* were searched for similar protein sequences of known function. Results are summarized in table 3 only for those proteins for which significant hits were obtained. Protein EGF38885.1 was predicted to have the best homolog Putative PIN and TRAM-domain containing protein from *Virgibacillus sp.* sharing 57% sequence identity (Table 3). As previously mentioned, PIN and TRAM domains were predicted in this protein (Table 1). Based on the two predictions, the presence of PIN domain and TRAM domain is strongly validated in EGF38885.1 protein of *L. monocytogenes*. The best homolog for protein AGR29492.1 was YacP from *Bacillus sp.* having 57% sequence identity

(Table 3). Interestingly, this protein was found to contain YacP domain (Table 1). This is another example of strong agreement between domains predicted in AGR29492.1 and its homologous sequence.

Table 3: Results of PSI-Blast

S.No.	Accession Number	Top Hits	E-value	Sequence Identity
1.	EGF38885.1	Putative PIN and TRAM-domain containing protein [*Virgibacillus*]	0.0	59%
2.	AGR29492.1	YacP [*Bacillus sp.*]	5e-107	57%
3.	EGF38918.2	Leucine-rich repeat-containing protein [*Listeria seeligeri*]	6e-126	68%
4.	EGF38940.1	YycH protein [*Listeria seeligeri*]	7e-118	84%
5.	EGF38822.1	Membrane protein [*Rhodococcus erythropolis*]	3e-70	35%
7.	EGF38751.2	Glyoxalase [*Listeria innocua*]	8e-72	84%

Protein EGF38918.2 was predicted to have the homolog Leucine-rich repeat-containing protein from *Listeria seeligeri.* Leucine rich repeats are also predicted from the previous analysis on domains (Table 1), thereby strongly validating the presence of LRR in EGF38918.2 of *L. monocytogenes*. Sequence homolog of protein EGF38940.1 was found to be YycH protein from *Listeria seeligeri* having 84% sequence identity (Table 3). Based on our analysis, protein EGF38940.1 was predicted to contain YcyI domain which along with YycH domain functions as YycFG two-component system. This is another example of strong correlation between domain assignment and sequence homolog.

Protein EGF38822.1 was found to share 35% sequence identity with a membrane protein from *Rhodococcuserythropolis*. Interestingly, protein EGF38822.1 was found to contain SsfLAP domain which is a transmembrane protein. Furthermore, six TM helices were predicted in this protein. Collectively, these predictions strongly validate the functional role of EGF38822.1 in membrane transport. Another hypothetical protein EGF38751.2 was found to share 84% sequence identity with Glyoxalase from *Listeria innocua* and this domain was predicted to function as Glyoxalase enzyme that is involved in the formation of lactic acid.

Analysis of functional partners

Prediction of functional partners was done only on the proteins of *L. monocytogenes* for which domains were assigned successfully. Predictions with score > 0.8 were considered. EGF38885.1 was predicted to lie adjacent to MEP cytidylyl transferase on the genome (Table 4). This prediction indicates the possibility of ribonuclease function of EGF38885.1 as predicted earlier. AGR29492.1, a YacP domain containing protein was predicted to function as a nuclease enzyme (Table 4). Consistent with this observation, this protein was predicted to interact with RNA methyl transferase (Table 4). This interaction is believed to make a process some complex that catalyzes the maturation of rRNA and tRNA.

Table 4: Prediction of Functional Partners

S. No.	Accession Number	Predicted Functional Protein	Score
1.	EGF38885.1	MEP cytidylyl transferase Glutamate-tRNA ligase	0.93 0.92
2.	AGR29492.1	RNA methyl transferase Serine acetyltransferase	0.97 0.96
3.	EGF38919.1	6-phospho-beta-glucosidase	0.85
4.	EGF38940.1	YycH protein Sensor histidine kinase YycG Putative metallo-hydrolase YycJ	0.99 0.96 0.85
6.	EGF38733.1	Transmembrane protein	0.99
7.	AGR29685.2	ABC transporter, permease Macrolide export ATP-binding ABC transporter Permease protein	0.99 0.99 0.98 0.97
8.	EGF38751.2	Methylglyoxal synthase Phosphate transporter family Tryptophan synthase	0.90 0.81 0.80

As discussed previously, protein EGF38940.1 contains YycI domain which is a part of YycFG two-component system (Table 1). YycI domain is always found in pair on the chromosome along with YcyH, downstream of the essential histidine kinase YycG. Both YcyH and YcyI proteins share a function in regulating the YycG kinase with which they appear to form a ternary complex. Consistent with this observation, EGF38940.1 was predicted to associate with both YycH protein and histidine kinase YycG (Table 4). These are the high scoring interactions established by Neighborhood, Cooccurence and text mining methods (Table 4). This data represents a strong correlation between domain assignment and prediction of functional partners and hence validating these annotations.

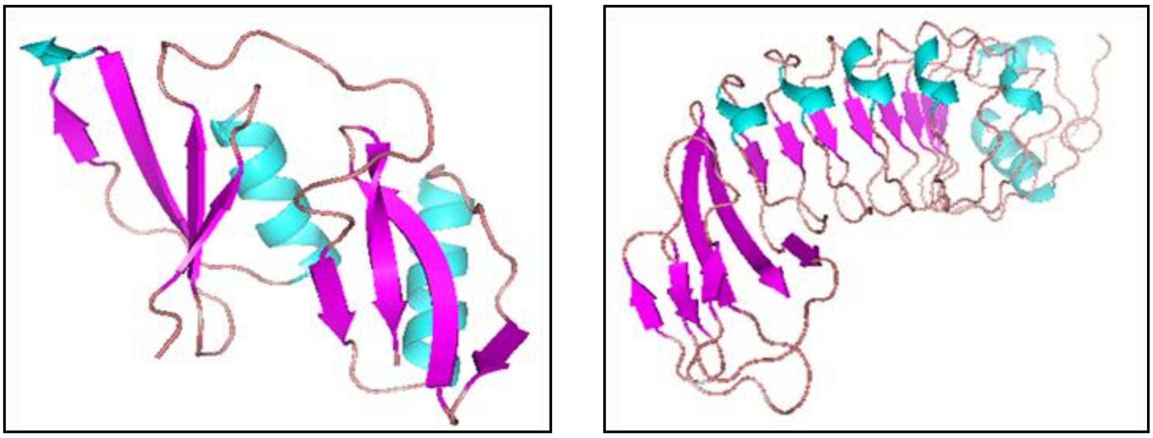
EGF38733.1 was predicted to contain TM helices and AzlC domain. Consistent with this prediction, this protein was strongly predicted to occur along with a membrane protein (Table 4). As mentioned earlier, domain and structural class predicted for AGR29685.2 could not be related. However, this protein is strongly predicted to interact with ABC transporter (Table 4). Thus, this protein may play a role in transport mechanism.

Fold Recognition

Only significant results obtained for fold recognition are enlisted in table 5. Based on our analysis, protein EGF38751.2 was annotated as glyoxalase enzyme. Interestingly, fold recognition results yielded a high confidence match with a putative glyoxalase enzyme (Table 5). This reflects a strong agreement between domain assignment and fold recognition. Its predicted structure is shown in figure 1a. EGF38918.2 was predicted to be LRR containing internal in protein. Consistent with this prediction, the template used for fold recognition was also an internalin protein (Table 5). This protein displays α/β horseshoe fold typically found in LRR domain (Figure 1b). These observations indicate that prediction of α structural class by 1D server for this protein was wrong.

Table 5: Results of Fold Recognition

S. No.	Accession Number	Template Information	Confidence	Coverage	Sequence Identity
1.	EGF38885.1	Uncharacterized protein of *Thermus thermophilus*	100%	37%	41%
3.	EGF38918.2	Internalin-A	100%	76%	37%
4.	EGF38919.1	SMI1/KNR4	99.6%	62%	25%
5.	EGF38940.1	Signaling protein YycI from *Bacillus subtilis*	100%	80%	26%
6.	EGF38823.2	PH domain-like barrel	100%	85%	32%
8.	EGF38751.2	Putative glyoxalase	99.9%	100%	97%
9.	EGF38752.1	SecB-like	100%	92%	31%



(a) **(b)**

Figure 1: Predicted structures of (a) EGF38751.2 and (b) EGF38918.2

As discussed earlier, EGF38940.1 was predicted to be a part of YycFG two-component system. Consistent with this observation, a high confidence match i.e. signaling protein YycI from *Bacillus subtilis* was been used as the template. The structure is shown in figure 2a. Based on our analysis EGF38919.1 is predicted to constitute SMI1_KNR4 domain (Table 1). Consistent with the observation, template used for structure prediction of EGF38919.1 was also a protein containing SMI1/KNR4 (Table 5). This reflects a strong agreement between domain assignment and fold recognition. Figure 2b shows the structure of EGF38919.1. Fold recognition of EGF38752.1 yielded high confidence match with the template SecB-like. SecB is a chaperone associated with the presence of an outer membrane and outer membrane proteins, mainly present in gram negative bacteria. However, secB-like genes are also found in Gram-positive bacteria. Thus, EGF38752.1 protein is homologous to SecB-like protein. According to our prediction, EGF38823.2 contains bacterial pleckstrin homology domain (Table 1). Consistent with this prediction, its structure was predicted using PH domain-like barrel as a template sharing 32 % sequence identity (Table 5).

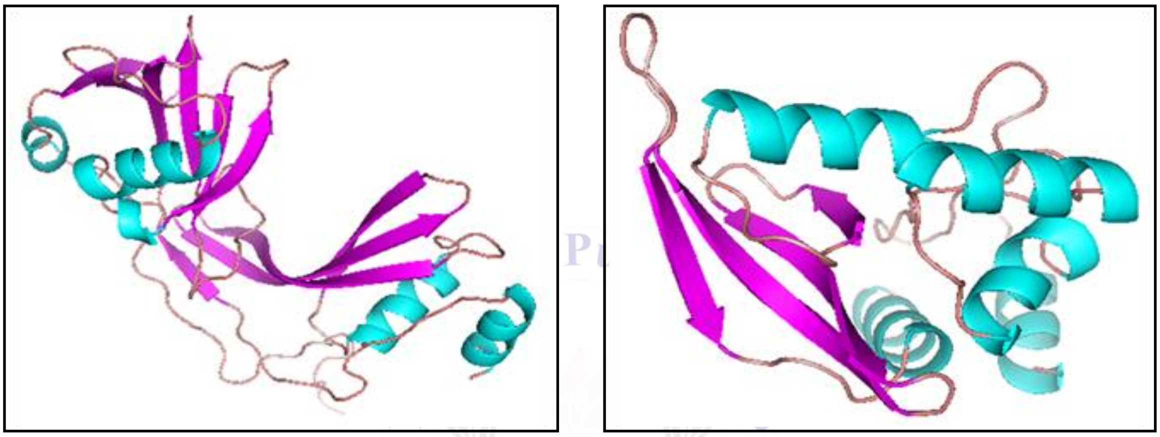

(a) (b)

Figure 2: Predicted structures of (a) EGF38940.1 and (b) EGF38919.1

CONCLUSIONS

Functional annotation of 30 hypothetical proteins of *Listeria monocytogenes* was carried out. Domains were successfully assigned for 19 proteins. It was observed that for some of the proteins, predictions from various tools correlated well with each other and hence pointed to the same conclusion. Functions of such proteins could be directly predicted. Best results were obtained for proteins, EGF38940.1, EGF38751.2, AGR29492.1, EGF38822.1 and EGF38896.2 whose results from different analysis were in strong agreement with each other. EGF38940.1 protein was predicted to be a part of YycFG two-component system. EGF38751.2 was predicted as glyoxalase enzyme. AGR29492.1 was predicted to contain YacP domain and may function as a nuclease enzyme. EGF38822.1 was predicted to be a SsfLAP domain containing transmembrane protein with a possible role in membrane transport. EGF38896.2 was predicted to be immunity protein. EGF38733.1 was predicted to be a membrane protein. AGR29482.1 was implicated to participate in membrane transport because of the presence of TolA domain. EGF38918.2 was predicted to be internalin protein containing leucine rich repeats. EGF38823.2 is predicted to play a role in intracellular signalling. EGF38918.2 is predicted to interact with E-cadherins on host cells via leucine rich repeats and hence enable *L. monocytogenes* gain entry into host cells. However, there was discrepancy in predictions regarding four proteins and hence require further investigation. Discrepancy was found mainly in the prediction of TM helices and structural class. Moreover, no information could be obtained for eleven proteins. These proteins may be novel proteins unique to *L.*

monocytogenes. Use of several approaches has aided in putative structure/function recognition for ten proteins of *L. monocytogenes*. These predictions can be further validated experimentally. The functional inferences drawn based on our predictions can provide valuable basis for designing experiments in order to understand the intricate mechanisms of pathogenesis and identify the potential drug targets.

REFERENCES

1. Jemmi T, Pak SI, Salman MD. Prevalence and risk factors for contamination with *Listeria monocytogenes* of imported and exported meat and fish products in Switzerland, 1992-2000. Prev Vet Med. 2002; 54(1):25-36.
2. Rebagliati V, Philippi R, Rossi M, et al. Prevention of food borne listeriosis. Indian J Pathol Microbiol. 2009; 52(2):145-149.
3. Reda WW, Abdel-Moein K, Hegazi A, et al. *Listeria monocytogenes*: An emerging food-borne pathogen and its public health implications. J Infect Dev Ctries. 2016; 10(2):149-154.
4. Vazquez-Boland JA, Kuhn M, Berche P, et al. *Listeria* pathogenesis and molecular virulence determinants. Clin Microbiol Rev. 2001; 14: 584-640.
5. Chen Y, Strain EA, Allard M, et al. Genome sequences of *Listeria monocytogenes* strains J1816 and J1-220, associated with human outbreaks. J Bacteriol. 2011; 193(13):3424-3425.
6. Ramakrishnan G, Ochoa-Montaño B, Raghavender US, et al. enriching the annotation of Mycobacterium tuberculosis H37Rv proteome using remote homology detection approaches: insights into structure and function. Tuberculosis (Edinb). 2015; 95(1):14-25.
7. Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database. Nucleic Acids Res. 2012 40(Database issue):D290-D301.
8. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30(9):1236-1240.
9. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 2004; 32(Web Server issue):W327-W331.
10. Kelley LA, Mezulis S, Yates CM, et al. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015; 10(6):845-858.
11. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015; 43(Database issue):D447-D452.
12. Anantharaman V, Koonin EV, Aravind L. TRAM, a predicted RNA-binding domain, common to tRNA uracil methylation and adenine thiolation enzymes. FEMS Microbiol Lett. 2001; 197(2):215-221.
13. Anantharaman V, Aravind L. The NYN domains: novel predicted RNAses with a PIN domain-like fold. RNA Biol. 2006; 3(1):18-27.
14. Zhang D, Iyer LM, Aravind L. A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems. Nucleic Acids Res. 2011; 39(11):4532-4552.
15. Kobe B, Deisenhofer J. The leucine-rich repeat: a versatile binding motif. Trends Biochem Sci. 1994; 19(10):415-421.
16. Yu WL, Dan H, Lin M. InlA and InlC2 of Listeria monocytogenes serotype 4b are two internalin proteins eliciting humoral immune responses common to listerial infection of various host species. Curr Microbiol. 2008; 56(5):505-509.
17. Schubert WD, Urbanke C, Ziehm T, et al. Structure of internalin, a major invasion protein of Listeria monocytogenes, in complex with its human receptor E-cadherin. Cell. 2002; 111(6):825-836.
18. Liu D. Identification, subtyping and virulence determination of *Listeria monocytogenes*, an important foodborne pathogen. J Med Microbiol. 2006; 55(Pt 6):645-659.
19. Enderlin CS, Selitrennikoff CP. Cloning and characterization of a Neurosporacrassa gene required for (1, 3) beta-glucan synthase activity and cell

wall formation. Proc Natl Acad Sci USA. 1994; 91(20):9500-9504.

20. Bisicchia P, Noone D, Lioliou E, et al. The essential YycFG two-component system controls cell wall metabolism in *Bacillus subtilis*. Mol Microbiol. 2007; 65(1):180-200.
21. Szurmant, H, Nelson K, Kim EJ, et al. YycH Regulates the Activity of the Essential YycFG Two-Component System in *Bacillus subtilis*. Journal of Bacteriology. 2005; 187(15):5419–5426.
22. Szurmant H, Mohan MA, Imus PM, et al. YycH and YycI interact to regulate the essential YycFG two-component system in Bacillus subtilis. J Bacteriol. 2007; 189(8):3280-3289.
23. Seeliger JC, Holsclaw CM, Schelle MW, et al. Elucidation and chemical modulation of sulfolipid-1 biosynthesis in Mycobacterium tuberculosis. J Biol Chem. 2012; 287(11):7990-8000.
24. Musacchio A, Gibson T, Rice P, et al. The PH domain: a common piece in the structural patchwork of signalling proteins. Trends Biochem Sci. 1993; 18(9):343-348.
25. Zhang D, de Souza RF, Anantharaman V, et al. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. BiolDirect. 2012; 7:18.
26. Steiner T, Kaiser JT, Marinkoviç S, et al. Crystal structures of transcription factor NusG in light of its nucleic acid- and protein-binding activities. EMBO J. 2002; 21(17):4641-4653.
27. Tavernarakis N, Driscoll M, Kyrpides NC. The SPFH domain: implicated in regulating targeted protein turnover in stomatins and other membrane-associated proteins. Trends Biochem Sci. 1999; 24(11):425-427.
28. Levengood SK, Beyer WF Jr, Webster RE. TolA: a membrane protein involved in colicin uptake contains an extended helical region. Proc Natl Acad Sci USA. 1991; 88(14):5939-5943.
29. Bateman A, Coggill P, Finn RD. DUFs: families in search of function. Acta Crystallographica Section F: Structural Biology and Crystallization Communications. 2010; 66(Pt 10):1148-1152.

How to cite this article: Kaur P, Gandhi R, Kaur H et al. Functional annotation of uncharacterized proteins of *Listeria monocytogenes*. International Journal of Research and Review. 2016; 3(12):4-15.

******