International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

Genome Annotation of Uncharacterized Protein in *Corynebacterium diphtheriae*

Diksha Verma

Department of Biotechnology and Bioinformatics, Guru Nanak Girls College, Model Town, Ludhiana (Punjab), India

Abstract: ***Assigning functional information to hypothetical proteins in bacterial genomes is crucial for gaining insight into their proteomes. Corynebacterium diphtheriae is one of the most prominent human pathogens and the causative agent of the communicable disease diphtheria. The genome of the NCTC13129 strain (biotype gravis) consists of a single circular chromosome of 2 488 635 bp, with no plasmids, in which the presence of 13 putative pathogenicity islands (PAIs) was demonstrated. Many genes that could contribute to the pathogenicity of C. diphtheriae are found within these putative islands. Comparative genomics revealed that the core genome consists of 2320 protein-coding sequences of which 45 are pseudogenes. Diphtheria toxin is one of the most widely studied bacterial toxins. The current work suggests a computational approach to annotate the putative function of the Corynebacterium diphtheria. Over all 80 sequences were collected from the NCBI database. The genome annotation were performed by using NCBI, MOTIF SEARCH, SMART, BLAST, TM PRED, CRNPRED, STRING, SOSUI, PROTPARAM, PSORTB and PHYRE2. The result suggests that most of the uncharacterized proteins resemble more to the transport regulators and signal transducers. Functions are assigned to 5 proteins out of the 80 uncharacterized proteins in C. diphtheriae with this approach. 3 out of 5 characterized proteins are predicted with confidence score above 98% by using PHYRE2. Therefore, the prediction of the function of these uncharacterized proteins might be helpful to find out the potential therapeutic drug targets against this deadliest pathogen.***

Keywords: *Corynebacterium diphtheriae, Genome Annotation, Confidence Score.*

1. Introduction

Corynebacterium diphtheriae belongs to family: Corynebacteriaceae , Genus: Corynebacterium , Species: C. diphtheria is a nonmotile, noncapsulated, non-sporulating, club-shaped, aerobic, gram-positive bacterium with a high GC-content and occurs in four biovars: gravis, mitis, intermedius and belfanti based on colonial morphology and biochemical profiles [1]. Toxigenic strains are lysogenic for one of a family of corynebacteriophages that carry the structural gene for diphtheria exotoxin, *tox*. Toxin production (toxigenicity) occurs only when the bacillus is itself infected (lysogenized) by a specific virus (bacterio-phage) carrying the gene for the production of exotoxin (tox gene). Only toxigenic strains can cause severe disease [2]. There are two types of clinical diphtheria: nasopharyngeal and cutaneous. Symptoms include airway obstruction by the pseudomembrane. The involvement of cervical lymph nodes may cause profound swelling of the neck (bull neck diphtheria), and the patient may have a fever (≥ 103 °F). The skin lesions in cutaneous diphtheria are usually covered by a gray-brown pseudomembrane. Life-threatening systemic complications, principally loss of motor function (e.g., difficulty in swallowing) and congestive heart failure, may develop as a result of the action of diphtheria toxin on peripheral motor neurons and the myocardium. The incubation period is short, usually of 1 week [3]. Recently in India, a total of 533 cases were seen in 11 districts of Kerala in 2016, of which 92% occurred in 3 districts of north Kerala; Malappuram, Kozhikode and Kannur. Almost 79% cases reported in >10 years age group. In <18 years age group, 62% were male while in ≥18 years, 69% were females [4].

2. Review of Literature

It is believed that the diphtheria was first described and reported by Hippocrates in the 5th century B.C. Prior to the development of the vaccination; the diphtheria was a common cause of childhood mortality and a major concern to families [5]. Diphtheria is a paradigm of the toxigenic infectious diseases. In 1884 Edwin Klebs and Friedrich Loffler, demonstrated that *Corynebacterium diphtheriae* was the agent of diphtheria, giving it the name Klebs-Loffler bacillus by which it used to be known [6]. In 1888, Pierre-Paul-Émile Roux discovered that a toxin produced by *C. diphtheriae* was the causative agent in the disease Diphtheria [7]. A severe outbreak occurred in 1735 in New England, with many families losing all of their children to the disease. That year, approximately 5,000 people, or 2.5% of the population, died of diphtheria, and the mortality rate in children was nearly 80% [8]. Another epidemic occurred in the late 1800s in New York with nearly 15,000 deaths in a period of 10 years [9]. The global incidence of diphtheria remained steady during 2007–2011 after a steady decline of over 95 % from 1980 to 2006. This was largely due to a resurgence of the disease in India, which alone accounted for 71–83 % of the total cases reported worldwide particularly in remote villages of northern Karnataka in South India in May 2011 [10]. Three recent outbreaks of diphtheria in Dibrugarh district, Assam in two consecutive years have been reported. The study was undertaken in Assam Medical College & Hospital, Dibrugarh after the diagnosis of two Diphtheria cases in the month of September and October 2015 and another in January 2016. Out of the 10 confirmed cases, 2 and 7 were in the first two outbreaks while only one in the third outbreak respectively. All the cases were of age > 10 years, unimmunized or partially immunized [11]. Indonesia is seeing a recent outbreak of diphtheria, sparking fear among the people and the government. Between January and November 2017, the government has recorded 593 diphtheria cases, spread across 95 regencies in 20 provinces. The death toll has reached 32 [12]. Diphtheria Toxoid was developed around 1921 but was not widely used until the early 1930s. It was incorporated with tetanus toxoid and acellular pertussis vaccine and became routinely used in the 1940s. DTaP is the vaccine of choice for children 6

weeks through 6 years of age [13]. German physician Emil von Behring (1854-1917) and Japanese scientist Shibasaburo Kitasato (1852-1931) were the first researchers to develop an antitoxin vaccine against diphtheria in about 1890. German microbiologist Paul Ehrlich (1854-1915) developed standard dosages for the antitoxin. In 1891 pathologist Anna Wessels Williams found a stronger strain of diphtheria antitoxin. In 1894 French bacteriologist Pierre Roux (1853-1933) developed a diphtheria antitoxin serum that he used to successfully treat more than 300 cases of the disease. In 1913 Hungarian scientist and pediatrician Bela Schick (1877-1967) introduced the Schick test, which tells whether or not a person can get diphtheria if exposed to it. In the test a small amount of diphtheria toxin is injected under the skin. If a red, swollen rash develops, the person is susceptible to diphtheria and should be immunized [14].

3. Pathogenesis

Diphtheriae toxin, which is secreted by toxigenic strains of *C. diphtheriae*, is a single polypeptide of M_r 58,342. Toxigenic strains of *C. diphtheriae* carry the *tox* structural gene found in lysogenic corynebacteriophages β *tox*$^+$, γ *tox*$^+$, and ω *tox*$^+$. Highly toxic strains have two or three *tox*$^+$ genes inserted into the genome [15]. The toxin is made of two joined proteins [16]. The B fragment binds to a receptor on the surface of the susceptible host cell, which proteolytically cleaves the membrane lipid layer enabling segment A to enter. Fragment A inhibits an amino acid transfer from RNA translocase to the ribosomal amino acid chain, thus inhibiting protein synthesis is required for normal host cell functioning [17]. DT causes a catalytic transfer of NAD to diphthamide, which inactivates the elongation factor, resulting in the inactivation eEF2, which results in protein synthesis blockage and subsequent cell death [18]. The best-studied effect is the role of iron in toxinogenesis. Studies during the 1930s demonstrated that DT production is greatest when *C. diphtheriae* is grown under iron-limiting conditions and is severely inhibited under high-iron growth conditions [19].

4. Methodology

1) From the NCBI hompage (https://www.ncbi.nlm.nih.gov/gene) 80 hypothetical proteins have been collected from the genome sequence of corynebacterium diphtheriae in the FASTA format.
2) MOTIF SEARCH (http://www.genome.jp/tool/motif/) and SMART (http://smart.embl.de) are used for motif and domain analysis and predicted 30 characterized proteins and 50 uncharacterized proteins from 80 hypothetical proteins.
3) BLAST (https://blast.ncbi.nlm.nih.gov/blast.cgi?PAGE=Proteins) is used to predict 30 characterized proteins with sequence similarity.
4) TMPRED (https://embnet.vitalit.ch/software/TMPRED_form.html) and CRNPRED (https://pdbj.org/crnpred/) is used to predict 20 transmembrane helices from 30 characterized proteins.
5) STRING (http://string-db.org) is used to predict functional partners of 21 characterized proteins from 30 characterized proteins.
6) SOSUI (http://www.tuat.ac.jp/mitaku/sosui/) is used for solubility analysis of 21 characterized proteins and predicted 10 transmembrane proteins.
7) PROTPARAM (http://www.expasy.org/tools/protparam.html) is used for analysis of physical and chemical parameters of 10 transmembrane proteins and predicted 7 stable proteins.
8) PSORTB (http://www.psort.org/psortb/) is used for protein localization of 7 stable proteins and predicted 5 localized proteins.
9) PHYRE 2 (http://www.sbg.bio.ic.ac.uk/phyre2) is used to predict fold recognition of 5 localized proteins out of which **3 proteins** are predicted with high confidence score (above 98%).

5. Results and Discussion

Domain prediction by using MOTIF SEARCH and SMART:

Table 1: DOMAIN analysis result for hypothetical proteins from MOTIF SEARCH and SMART

Accession No.	Motif Search	E-Value	Region	Smart	Region	Blast Identity Score
1) WP_010933941.1	HiCB	0.16	3-26	Low complexity	29-40	99%
2) WP_014308991.1	Cadherin	0.089	78-122	CA	107-182	99%
3) WP_035100135.1	DUF2964	0.43	12-55	Transmembrane region	10-29	99%
4) WP_010933953.1	Alpha-amylase	4.3e-06	39-75	Alpha-amylase	82-123	99%
5) WP_010933956.1	Phage_GPO	0.08	24-111	Transmembrane region	116-138	99%
6) WP_010933988.1	RNase_T	2.3e-05	33-75	RNase_T	32-82	99%
7) WP_041627721.1	Phage integrase	0.38	182-220	No domain	N/A	99%
8) WP_010933996.1	Condensation	0.25	64-92	No domain	N/A	99%
9) WP_041628104.1	Prok-E2-A	0.17	181-222	Transmembrane region	16-38	99%
10) WP_041627726.1	Spore-YhCN-YLa	0.047	18-69	Low complexity	11-22	-
11) WP_010934025.1	DUF1801	1.1e-10	29-121	DUF1801	28-123	99%
12) WP_010934103.1	HNN	0.28	134-180	Low complexity	247-266	-
13) WP_003850222.1	HNH	1.4e-05	38-91	HNHC	26-87	99%
14)WP_003850227.1	Terminase_1	0.0049	189-301	Low complexity	507-529	99%

15) WP_010934110.1	Phage-tail-1	0.07	56-101	No domain	N/A	99%
16) WP_010934114.1	Phage Min_Tail	3.4e-08	236-406	Transmembrane region	511-530	99%
17) WP_010934116.1	Zf-C4H2	0.0027	156-266	Coiled coil	210-236	99%
18) WP_010934117.1	PTR	0.011	359-414	Low complexity	55-67	99%
19) WP_010934120.1	PRA 1	0.14	33-70	Transmembrane region	27-46	99%
20) WP_010934122.1	CC2_LZ	0.025	32-85	Transmembrane region	4-26	95%
21) WP_010934104.1	Ribosomal_L11	0.13	124-151	Low complexity	383-397	99%
22) WP_010934128.1	DUF3180	0.0014	10-48	Transmembrane region	15-37	99%
23) WP_010934134.1	sdrD_B	0.017	117-149	Transmembrane region	13-35	99%
24) WP_010934140.1	RIP	0.16	75-140	Transmembrane region	7-29	99%
25) WP_010934142.1	No domain	N/A	71-83	Low complexity	71-83	98%
26) WP_010934154.1	SUFU	5.4e-08	58-183	SUFU	40-183	99%
27) WP_085713323.1	Zf_ISL3	2.2e-07	25-65	Zf_ISL3	22-68	99%
28) WP_010934168.1	Mannosyl trans 2	5e-09	25-182	Mannosyl-trans2	1-195	99%
29) WP_003850400.1	No domain	N/A	N/A	Transmembrane region	4-26	64%
30) WP_010934186.1	DUF559	0.00032	209-273	AbrEi_4	29-140	99%

Accession no. WP_010934120.1 has PRA1 protein family with Pfam ID PF03208 is a domain successfully predicted by Motif Finder and SMART database tools having e-value of 0.14 present in the region between 33-70. Prenylated Rab acceptor 1 (PRA1) domain proteins are small transmembrane proteins that regulate vesicle trafficking as receptors of Rab GTPases and vacuolar soluble N-ethylmaleimide-sensitive factor attachment receptor protein VAMP2. In humans, PRAF1 is a general Rab protein regulator required for vesicle formation from the Golgi complex. It may control vesicle docking and fusion by mediating the action of Rab GTPases to the SNARE complexes. It inhibits the removal of Rab GTPases from the membrane by GDI. PRAF2 is a new 19 kDa protein with four putative transmembrane (TM) domains, which plays a role in the regulation of intracellular protein transport. Recently, PRAF2 was found to interact with the chemokine receptor CCR5. It also plays a pro-apoptic role in cerulenin-induced neuroblastoma apoptosis. Strong PRAF2 protein expression was also found in human tumor tissues of the breast, colon, lung, and ovary, with a weaker staining pattern in normal tissues of the same patient. PRAF3 negatively modulates SLC1A1/EAAC1 glutamate transport activity by decreasing its affinity for glutamate. It plays a important role in vesicular trafficking, lipid transport and cell migration. 2 transmembrane helices having α/β protein structure are present in it predicted by TMPred and CRNPRED protein secondary structure prediction tools. This hypothetical protein contains the neighbourhood partners which are predicted by using STRING interacting protein partners prediction tool. This is a stable protein containing 107 amino acids with the molecular weight of 11336.08 determined by PRotParam. Its subcellular localization is cytoplasmic membrane predicted by Psortb. The 3D structure of protein is predicted with **66.1%** confidence score by Phyre 2.

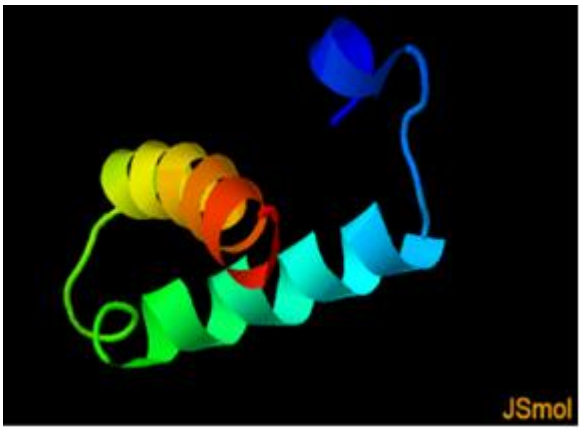

Figure 1: Accession no. WP_010934120.1 [PRA1 protein family]

Prediction of No. and Type of Secondary Structure present by using TM PRED and CRNPRED databases:

Table 2: Secondary Structure result for hypothetical proteins from TM PRED and CRNPRED

Accession No.	TM PRED No. of Secondary Structure Present	CRNPRED Type of Secondary Structure Present
1)WP_035100135.1	3	α,β
2)WP_010933953.1	1	α,β
3)WP_010933956.1	1	α,β
4)WP_010933996.1	3	α
5)WP_041628104.1	12	α,β
6)WP_041627726.1	1	α
7)WP_010934025.1	1	α,β
8)WP_010934103.1	6	α,β
9)WP_003850227.1	5	α,β
10)WP_010934114.1	18	α,β
11) WP_010934116.1	3	α,β
12) WP_010934117.1	2	α
13) WP_010934120.1	2	α,β
14) WP_010934122.1	2	α,β
15)WP_010934128.1	1	α,β
16) WP_010934134.1	2	α,β
17) WP_010934140.1	2	α,β

18) WP_010934154.1	1	α,β
19)WP_010934168.1	11	α,β
20) WP_003850400.1	3	α,β

Accession no. WP_003850400.1 has MrpF_PhaF protein superfamily with Pfam ID PF04066 is a domain successfully predicted by Motif Finder and SMART database tools having e-value of 5.04e-19 present in the region between 1-88. Putative monovalent cation/H^+ antiporter subunit F; reviewed. Members of the PhaF/MrpF family are predicted to be integral membrane proteins with 3 transmembrane regions, involved in regulation of pH. PhaF ia a part of a potassium efflux system involved in pH regulation. It is also involved in symbiosis in Rhizobium meliloti. MrpE is part of a Na^+/H^+ antiporter complex, also involved in pH homeostasis. MrpE is thought to be an efflux system for Na^+ and cholate. The Mrp system in Bacilli may also have primary energization capacities. 3 transmembrane helices having α/β protein structure are present in it predicted by TMPred and CRNPRED protein secondary structure prediction tools. This hypothetical protein contains the neighbourhood and cooccurence partners which are predicted by using STRING interacting protein partners prediction tool. This is a stable protein containing 88 amino acids with the molecular weight of 9567.58 determined by PRotParam. Its subcellular localization is cytoplasmic membrane predicted by Psortb. The 3D structure of protein is predicted with **62.1%** confidence score by Phyre 2.

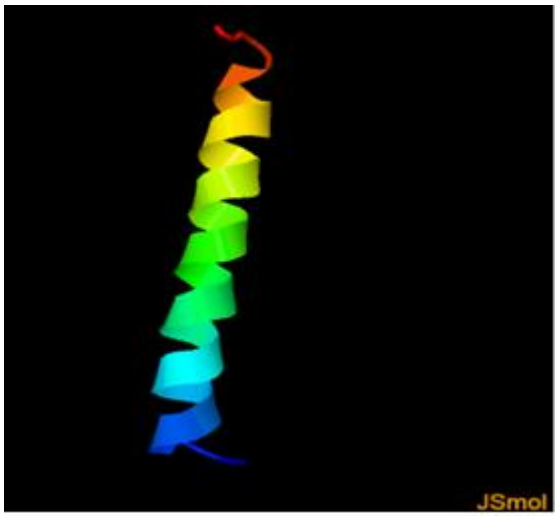

Figure 2: Accession no. WP_003850400.1 [MrpF_PhaF protein superfamily]

*Prediction of Interacting partners by using STRING database***:**

Table 3: Interacting Partners searching results for hypothetical proteins from STRING

Serial No.	Accession No.	Name	Score	Interacting Partners
1.	WP_010933941.1	VN94_00535	0.859	Neighbourhood
2.	WP_010933953.1	CDPW8_0013	0.970	Neighbourhood, Co-Occurrence, Experiments, Database, Text Mining
3.	WP_010933956.1	VN94_00625	0.648	Neighbourhood
4.	WP_010934025.1	VN94_00965	0.819	Neighbourhood, Gene Fusion
5.	WP_010934103.1	BioA	0.999	Neighbourhood, Gene Fusion, Co-Occurrence, Co-Expression, Database, Text Mining
6.	WP_003850227.1	CDPW8_0159	0.829	Neighbourhood, Co-Occurrence
7.	WP_010934110.1	CDPW8_0157	0.728	Neighbourhood
8.	WP_010934114.1	VN94_01445	0.845	Neighbourhood, Co-Occurrence
9.	WP_010934116.1	CDPW8_0172	0.859	Neighbourhood
10.	WP_010934117.1	CDPW8_0173	0.859	Neighbourhood
11.	WP_010934120.1	CDPW8_0176	0.859	Neighbourhood
12.	WP_010934122.1	CDPW8_0178	0.859	Neighbourhood
13.	WP_010934104.1	CDPW8_0157	0.728	Neighbourhood
14.	WP_010934134.1	SpaE	0.919	Neighbourhood, Text Mining
15.	WP_010934140.1	CDPW8_0260	0.537	Neighbourhood
16.	WP_010934142.1	VN94_01600	0.802	Neighbourhood, Co-Occurrence
17.	WP_010934154.1	VN94_01655	0.859	Neighbourhood
18.	WP_085713323.1	CDPW8_2148	0.527	Neighbourhood
19.	WP_010934168.1	VN94_01730	0.700	Neighbourhood, Co-0occurence
20.	WP_003850400.1	MnhF	0.859	Neighbourhood
21.	WP_010934186.1	VN94_01830	0.702	`Co-Occurrence

Accession no. WP_ 010934114.1 has Phage Min_Tail protein family with Pfam ID PF10145 is a domain predicted by Motif Finder and SMART database tools having e-value of 3.4e-08 present in the region between 236-406. The initial interaction between a tailed phage and its bacterial host is mediated by the distal part of the phage tail, and studies of the distal tail structures thus contribute to a fundamental understanding of phage-host interactions. Moreover, determinations of the underlying mechanisms of phage-host interactions at the level of single proteins depend on a prior knowledge of the interacting part of the phage structure Phage –related minor tail protein family members contain 2 putative structural gene, orf tmp (tape measure protein) & orf bpp (baseplate protein), which were examined by introduction of specific mutations in the prophage state. Specific mutations in orf tmp resulted in the production of mostly phage head structures without tails and a few wild type looking phages. Furthermore , construction of an inframe deletion or duplication of 29% in orf tmp was shown to shorten or lengthen the phage tail by approximately 30% respectively. The orf tmp is proposed to function as a tape measure protein, TMP is important for assembly of the phage tail & involved in tail length determination. Minor tail structures, such as fibers and baseplates, are also found among members of these families. A long phage tail is usually assembled from the distal end proceeding towards the head-proximal end until a mature tail structure is formed, and the tail is ultimately joined to a mature and independently assembled head, thereby forming an infectious phage virion. 18 transmembrane helices having α/β protein structure are present in it predicted by TMPred and CRNPRED protein secondary structure prediction tools. This hypothetical protein contains the neighbourhood and cooccurence partners which are predicted by using STRING interacting protein partners prediction tool. This is a stable protein containing 1880 amino acids with the molecular

weight of 196716.42 determined by PRotParam. Its subcellular localization is extracellular predicted by Psortb. The 3D structure of protein is predicted with **98.7%** confidence score by Phyre 2.

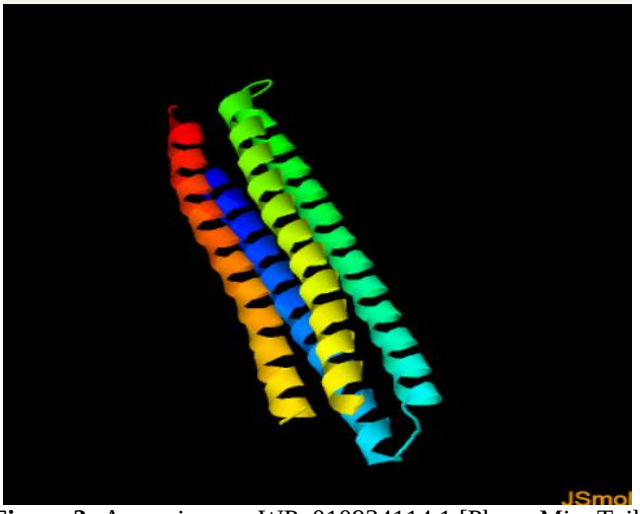

Figure 3: Accession no. WP_010934114.1 [Phage Min_Tail protein family]

Prediction of Average of Hydrophobicity score and Solubility by using SOSUI database:

Table 4: Average of Hydrophobicity score and Solubility results for hypothetical proteins from SOSUI

Serial No.	Accession No.	Average Of Hydrophobicity	Solubility
1.	WP_035100135.1	0.547701	Transmembrane Protein
2.	WP_010933953.1	-0.293220	Soluble Protein
3.	WP_010933956.1	-0.758606	Transmembrane Protein
4.	WP_010933996.1	-0.188509	Soluble Protein
5.	WP_041628104.1	0.686755	Transmembrane Protein
6.	WP_041627726.1	-0.862602	Soluble Protein
7.	WP_010934025.1	-0.407910	Soluble Protein
8.	WP_010934103.1	-0.155943	Soluble Protein
9.	WP_003850227.1	-0.303130	Soluble Protein
10.	WP_010934114.1	-0.095265	Transmembrane Protein
11.	WP_010934116.1	-0.445036	Soluble Protein
12.	WP_010934117.1	-0.329207	Soluble Protein
13.	WP_010934120.1	0.042282	Transmembrane Protein
14.	WP_010934122.1	-0.319481	Transmembrane Protein
15.	WP_010934128.1	-0.475258	Transmembrane Protein
16.	WP_010934134.1	-0.227541	Transmembrane Protein
17.	WP_010934140.1	-0.305970	Soluble Protein
18.	WP_010934154.1	-0.084889	Soluble Protein
19.	WP_010934168.1	0.790956	Transmembrane Protein
20.	WP_003850400.1	0.366154	Transmembrane Protein

Accession no. WP_010934168.1 has Mannosyl-trans2 and EpSG protein family with Pfam ID PF04188 and PF14897 is a domain successfully predicted by Motif Finder and SMART database tools having e-value of 5e-09 and 0.0031 present in the region between 25-182 and 222-318. Mannosyltransferase (PIG-V) is a family of eukaryotic ER membrane proteins that are involved in the synthesis of glycosylphosphatidylinositol (GPI), a glycolipid that anchors many proteins to the eukaryotic cell surface. Proteins in this family are involved in transferring the second mannose in the biosynthetic pathway of GPI. Protein *O*-mannosyltransferases (Pmt proteins) plays an important role in the secretion, localization, and function of many proteins, as well as in cell wall integrity and virulence. EpSG protein are likely glycosyl transferases belonging to the membrane protein GT-C clan. EpSG (transmembrane protein) may be involved in the production of the exopolysacchaaride (EPS) component of the extracellular matrix during biofilm formation. Bacterial exopolymeric substances (EPS) are molecules released in response to the physiological stress encountered in the natural environment. EPS are structural components of the extracellular matrix in which cells are embedded during biofilm development. The chemical nature and functions of these EPS are dependent on the genetic expression of the cells within each biofilm. EPS is responsible for the adhesion of the chains of cells into bundles. Required for biofilm maintenance. 11 transmembrane helices having α/β protein structure are present in it predicted by TMPred and CRNPRED protein secondary structure prediction tools. This hypothetical protein contains the neighbourhood and cooccurence partners which are predicted by using STRING interacting protein partners prediction tool. This is a stable protein containing 345 amino acids with the molecular weight of 36814.23 determined by PRotParam. Its subcellular localization is cytoplasmic membrane predicted by Psortb. The 3D structure of protein is predicted with **99.4%** confidence score by Phyre 2.

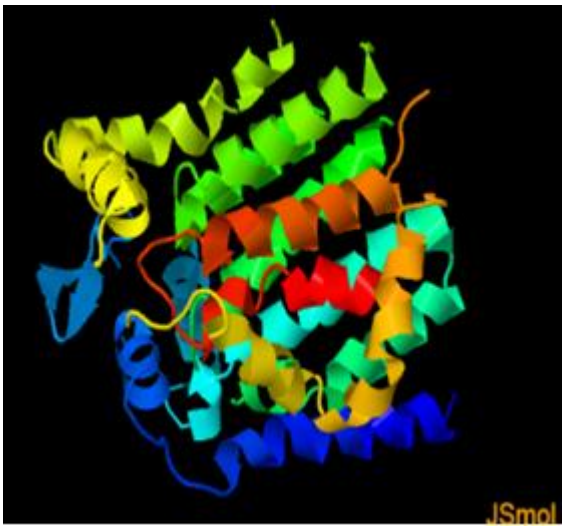

Figure 4: Accession no. WP_010934168.1 [Mannosyl-trans2 and EpSG proteins]

Prediction of Stability by using PROTPARAM database:

Table 5: Stability results for hypothetical proteins from PROTPARAM

Serial No.	Accession No.	No. of Amino Acids	Molecular Weight	Isoelectric Point (pI)	Instability Index	Grand Average of Hydrophobicity
1.	WP_010933956.1	202	21982.47	8.47	29.21 (STABLE)	-0.745
2.	WP_010924114.1	1880	196716.42	6.59	25.16 (STABLE)	-0.079
3.	WP_010934120.1	107	11336.08	5.44	35.9211 (STABLE)	0.382
4.	WP_010934122.1	112	12166.08	8.12	32.63 (STABLE)	-0.130
5.	WP_010934134.1	263	27084.86	9.15	24.49 (STABLE)	-0.132
6.	WP_010934168.1	345	36814.23	9.84	27.55 (STABLE)	0.988
7.	WP_003850400.1	88	9567.58	9.36	31.20 (STABLE)	0.934

Accession no. WP_010934134.1 has SdrD-B, carboxypep D-reg and ATP1G1_PLM_MAT8 like domain with Pfam ID PF17210, PF13620 and PF02038 respectively are successfully predicted by Motif Finder and SMART database tools having e-value of 0.017, 0.014 & 0.29 present in the region between 117-149, 98-148 & 233-260. SdrD-B domain has 3 calcium binding sites within a greek key beta sandwich fold. Biofilms of live bacteria forming on medical devices and implants contribute significantly to bacterial blood dissemination and to the spread of nasocomial infections cell surface SdrD protein plays a key role in the attachment of bacteria to the extracellular matrix (ECM) and in the formation of biofilm. SdrD binds calcium ions using its B_1-B_5 region bearing EF- hand Ca-binding sites, leading to conformational changes in the structure of SdrD. This alters the distance between the bacterial surface and the ECM-interacting domain of SdrD in a spring-like fashion, participating in bacterial attachment. Carboxypeptidase regulatory-like domain has a regulatory protein within termed as beta-sandwich, comprising 7 strands in 2 sheets in a greek-key topology. Some family members have additional 1-2 strands to the common fold. It is a putative outer membrane protein probably involved in nutrient binding and it may also function as cellular signal transducers. ATP1G1/PLM/MAT8 or FXYD protein family is a family containing ion transport regulators. Initial functional characterization suggested that FXYD proteins act as a channels or as modulators of ion channels. Another function include they act as tissue-specific regulatory subunits of the Na,K-ATPase. Each of theses auxillary subunits produces a distinct functional effect on the transport characteristics of the Na,K-ATPase that is adjusted to the specific functional demands of the tissue in which the FXYD protein is expressed. 2 transmembrane helices having α/β protein structure are present in it predicted by TMPred and CRNPRED protein secondary structure prediction tools. This hypothetical protein contains the neighbourhood and textmining partners which are predicted by using STRING interacting protein partners prediction tool. This is a stable protein containing 263 amino acids with the molecular weight of 27084.86 determined by PRotParam. Its subcellular localization is cell wall predicted by Psortb. . The 3D structure of protein is predicted with **99.7%** confidence score by Phyre 2.

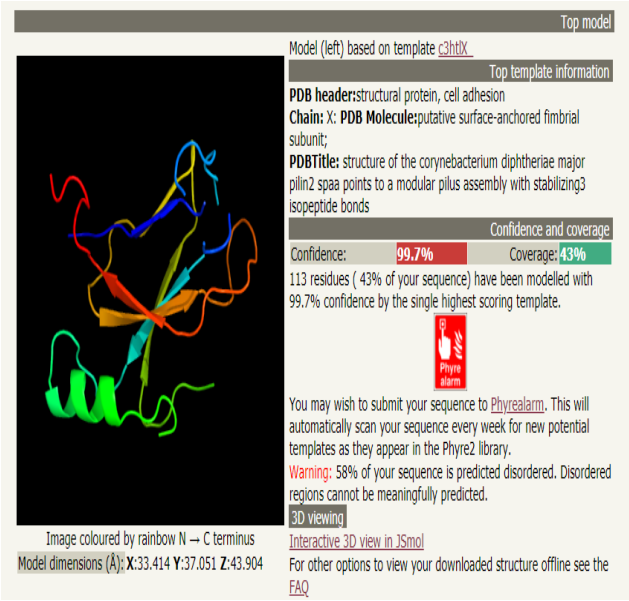

Figure 5: Accession no. WP_010934134.1 [SdrD-B, carboxypep D-reg and ATP1G1_PLM_MAT8 like domain]

Prediction of Subcellular Localization by using PSORTB database:

Table 6: Subcellular Localization results for hypothetical proteins from PSORTB

Serial No.	Accession No.	Subcellular Localization
1.	WP_010924114.1	Extracellular
2.	WP_010934120.1	Cytoplasmic Membrane
3.	WP_010934134.1	Cell Wall
4.	WP_010934168.1	Cytoplasmic Membrane
5.	WP_003850400.1	Cytoplasmic Membrane

6. Conclusion

The current approach had been used for genome annotation of uncharacterized proteins in *Corynebacterium diphtheriae*. The uncharacterized protein sequences showed good sequence relationship with other proteins. The current PHYRE 2 study proved that most sequences show good similarity. 5 uncharacterized proteins in *Corynebacterium diphtheriae* had been found, which showed dominant helix-turn-helix domains and structural class regions. First protein was found in PRA1 protein family, PRAF1, PRAF2 and PRAF3 played an important role in vesicular trafficking, lipid transport and cell migration. Strong PRAF2 protein expression was also found in human tumor tissues of the breast, colon, lung and ovary. Second protein was found in MrpF_PhaF superfamily, proteins from this family were involved in regulation and homeostasis of pH. The Mrp system in Bacilli might also have primary energization

capacities. Third protein was found in phage Min_Tail family, which was involved in phage host interactions and tail length determination. Therefore, it played a role in the formation of an infectious phage virion. Fourth protein was found in Mannosyl-trans2 and EpSG protein family. Mannosyl-trans2 played an important role in the secretion, localization, and function of many proteins, as well as in virulence. EpSG protein was involved in biofilm formation. Fifth protein was found in SdrD-B, carboxypep D-reg and ATP1G1_PLM_MAT8 protein families. SdrD protein played a key role in the attachment of bacteria to the ECM and in the formation of biofilm, Carboxypep D-reg protein played a role in cellular signal transduction and FXYD protein family was a family containing ion transport regulators. Hypothetical proteins were annotated by using some bioinformatics tools i.e. using NCBI, MOTIF SEARCH, SMART, BLAST, TM PRED, CRNPRED, STRING, SOSUI, PROTPARAM, PSORTB and PHYRE2. Further functional annotation and methods can be used for more analysis. The current work may be used as a preliminary study to find the novel drug targets against *Corynebacterium diphtheriae*.

References

[1] John, R.M. and Baron, S. (1996). University of Texas Medical Branch at Galveston. Chapter 32 Corynebacterium Diphtheriae, 4th edition, *Medical Microbiology*.
[2] Weinbaum, C. (2015). Centers for Disease Control and Prevention Epidemiology and Prevention of Vaccine-Preventable Diseases. *Diphtheria*, 13th Edition.
[3] Spradley, R. (2011). A Microbial Biorealm page on the genus *Corynebacterium diphtheria*, Microbewiki.
[4] Sangal, L. et. al. (2017). Resurgence of Diphtheria in north Kerala, india, 2016: laboratory supported case-Based surveillance Outcomes, Frontiers in Public Health, 5 (218).
[5] English, P. (1985). Diphtheria and theories of infectious disease. *Centennial appreciation of the critical role of diphtheria in the history of medicine*, Pediatrics, 76:1-9.
[6] Medindia Health Network Pvt Ltd. (2007). *Diphtheria*. Available: https://microbewiki.kenyon.edu/index.php/Corynebacterium_diphtheriae.
[7] Chung; Thom, K.; Behring, E.V. (1917). Department of Biology, University of Memphis, Encyclopedia Britannica. *Diphtheriae*.
[8] Osborne, R. Diptheria. Available: http://www.austincc.edu/microbio/2704v/cd1.htm.
[9] Frassetto, L.A. (2006). Corynebacterium infections. Available: http://www.austincc.edu/microbio/2704v/cd1.htm.
[10] Parande, M.V.; Parande, A.M.; Lakkannavar, S.L.; Kholkute, S.D. and Roy, S. (2014). Diphtheria outbreak in rural North Karnataka, India. Department of Health Research, Govt. of India, Belgaum, Karnataka 590010, India.
[11] Das, P.P.; Patgiri, S.J.; Saikia, L. and PauL, D. (2016). Recent Outbreaks of Diphtheria in Dibrugarh District, Assam, India. *Journal of Clinical and Diagnostic Research*, 10(7):DR01-DR03.
[12] Sariadji, K. (2017). Indonesia's diphtheria outbreak: problems in vaccination and antibiotics efficacy, Researcher in Bacteriology, National Institute of Health Research and Development Ministry of Health Indonesia.
[13] World Health Organisation (2014). Information Sheet Observed Rate Of Vaccine Reactions Diptheria , Pertussis , Tetanus Vaccines. Global Vaccine Safety Essential Medicines & Health Products 20, Avenue Appia, CH-1211 Geneva 27 Switzerland.
[14] DPT vaccine, Medical Discoveries powered by JRank. Available: http://www.discoveriesinmedicine.com/Com-En/DPT-Vaccine.html.
[15] Hadfield, T.L. et. al. (2000). The Pathology of Diphtheria. *The Journal of Infectious Diseases*, 181:S116–S120.
[16] Dass, J. F. and Verma, D. (2008). Implications from predictions of HLA-DRB1 binding peptides in the membrane proteins of Corynebacterium diphtheriae. *Bioinformation*, 3(3):111-3.
[17] Chen, R.T.; Broome, C.V. and Weinstein, R.A. (1985). Diphtheria in the United States. *Am J Public Health*, 75(12):1393-7.
[18] Bruce, M.L. and Taylor, P.T. (2017). Diphtheria.Medscape.
[19] Pappenheimer, A.M. and Johnson, S. (1936). Studies on diphtheria toxin production. I. The effect of iron and copper. *Br J Exp Pathol*, 17:41-335.

Author Profile

Diksha Verma received the Msc Biotechnology from Guru Nanak Girls College, Model Town Ludhiana in 2018.