Genome Annotation and Structure Predictions for Hypothetical Proteins in Agrobacterium Fabrum Str. C58 Plasmid At

Azeem Uddin Siddiqui
Indian School of Mines (ISM), Dhanbad - 826004, Jharkhand, India.

Mohd. Ahmad
Jamia Hamdard, Mehrauli Badarpur Road, Hamdard Nagar, New Delhi, Delhi, 110062,India.

Archis Pandya
Ashok and Rita Patel Institute of Integrated Study and Research in Biotechnology and Allied Sciences, Sardar Patel University New Vallabh Vidya Nagar-388121, Anand, Gujarat.

Swapnil Sanmukh
National Environmental Engineering Research Institute (NEERI)
CSIR-Complex, Chennai-600113,
Tamil Nadu (India)

Krishna Khairnar
National Environmental Engineering Research Institute (NEERI), Nagpur-440020
Maharashtra (India)

ABSTRACT
The in-silico approach was utilised for prediction of structure, function and sub-cellular localization of the hypothetical proteins in *Agrobacterium fabrum* str. C58 plasmid At. In *Agrobacterium fabrum* str. C58 plasmid At out of 209 genes screened for hypothetical proteins, structures, functions and sub-cellular localization were predicted for 84 hypothetical protein. The Bioinformatics web tools like CDD-BLAST, INTERPROSCAN and PFAM were used for the functional annotations of hypothetical proteins; Cello v 2.5 was used to determine the sub- cellular localization of annotated hypothetical proteins whereas, PS^2 Server-Protein Structure Prediction server was used for generating 3-D structures of the identified proteins by searching protein databases for the presence of conserved domains and templetes. This Insilico study revealed much helpful information regarding the understanding of functional characteristics of hypothetical proteins in *Agrobacterium fabrum* str. C58 plasmid At as well as their role in the life cycle of the bacterium.

Keywords
Unknown proteins; Bioinformatics web tools, protein databases, tertiary structures, functional characteristics.

1. INTRODUCTION
We ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material Agrobacterium tumefaciens is a plant pathogen which has a unique ability to transfer a defined segment of DNA to eukaryotes, which finnaly integrates into the eukaryotic genome. Such ability has extensively used to transfer and integrate DNA for random mutagenesis and is well known as a powerful tool for production of transgenic plants [3]. A. tumefaciens is known as the causal agent of crown gall disease in plants [28]. This strain has been intensively studied and is the parent of many strains used for the genetic transformation of plants [6,9].

The genome structure of A. tumefaciens C58 5.67-Mb genome isolated from a cherry tree (Prunus) tumor, consists of a circular chromosome, linear chromosome, and two plasmids: the tumor-inducing plasmid pTiC58 and a second plasmid, pAtC58 [12,3,11]. Plasmid pAtC58 contains the attachment (att) genes involved in initial specific attachment of the bacterium to plant cells, as well as a second, partial att locus [6,9]. The pAtC58 plasmid is reportedly dispensable for virulence [14]. It also encodes a protein with strong similarity to fungal tannases. This enzyme might allow the bacterium to use tannins as nutrients or to defend itself against the antimicrobial activities of many tannins [14,29]. The genome contains 5419 predicted protein-coding genes [14], 1236 conserved hypothetical genes (22.8%) and 708 hypothetical genes (13.1%) with no significant matches in the sequence databases.

The online bioinformatics tools and servers have ability for searching databases by choosing standard parameters for revealing the function of a particular gene (protein), determine the presence of the enzymatic conserved domain/s in the sequences (which may assist in the categorizing protein into specific family) and predict the three dimensional structures for protein sequences. Bioinformatics web tools like CDD-BLAST, INTERPROSCAN, and PFAM can be very useful for understanding the function of targeted protein sequence and Cello can be used to find the location of protein or enzyme within the cell [8,13,15-27]. The 3-D structure prediction of such proteins can be carried out by using Protein Structure Prediction Server (PS2 server) [8].

Our main objective behind this work is to identify the functional as well as structurally predictable unidentified hypothetical proteins restricted to the Plasmid pAtC58 of Agrobacterium tumefaciens using bioinformatics tools and online servers [10].

2. METHODOLOGY

2.1 Sequence Retrieval
The whole genome sequences for *Agrobacterium fabrum* str. C58 plasmid At was retrieved from the KEGG database (http://www.genome.jp/kegg/). [14, 29]

2.2 Functional Annotation and Categorization

The hypothetical proteins from *Agrobacterium fabrum str. C58 plasmid At* were screened and analysed for the presence of conserved functional domains and 3-D structures using the web-tools. The four bioinformatics web tools CDD-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) [1, 4, 5, 7], INTERPROSCAN (http://www.abi.ac.uk/interpro) [32] Pfam (http://www.pfam.sanger.ac.uk/) [2] and Cello [30] were used. CDD-Blast, Interproscan and Pfam have the ability to search the defined conserved domains in the sequences and assist in the classification of proteins in appropriate family depending upon the information available in databases. The Cello server predicts the possible sub-cellular localization for the identified protein or enzyme.

2.3 Protein Structure Prediction

The determinations of 3-D structures of the protein sequences under consideration were carried out by using PS2 Protein Structure Prediction Server (http://www.ps2.life.nctu.edu.tw/) [7, 31]. This online server accepts the protein (query) sequences in FASTA format to generate resultant proteins 3D structures. The structure determination is based on the template detected in the functional annotations and which must be available in the structure alignment for modeling purpose.

3. RESULTS AND DISCUSSION

The insilico studies for characterizing 209 genes for hypothetical proteins from the whole genome sequences for *Agrobacterium fabrum* str. C58 plasmid At were carried out. Out of 209 genes, probable function annotation, characterization, subcellular localization and 3-D structure prediction was successfully for 84 gene sequences of protein. The online-automated bioinfromatics tools like CDD- Blast, Interproscan, Pfam, Cello and PS^2server were used for structurural and functional characterization of screened hypothetical proteins. The results obtained after analysis of are represented in Table 1. The 3-D structures built by using best scoring templates are represented in the order as Template ID, Identity, Score and E-value in structure column of respective *Agrobacterium fabrum* str. C58 plasmid At specific gene in Table 1.

4. CONCLUSION

These studies have sorted 84 functionally as well as structurally important hypothetical proteins from Agrobacterium fabrum str. C58 plasmid At, which suggest that many probable functional proteins are available in the Agrobacterium fabrum str. C58 plasmid At. We have successful characterized the 84 unknown proteins from 209 screened hypothetical protein sequences from Agrobacterium fabrum str. C58 plasmid At for verifying their structure and functions of the gene products. This predicted functions and three dimensional structures may assist in establishing their role in the life cycle of the bacterium. This computationally generated data can also be used for developing new protocols for transgenic plant production or for modification of the existing protocols through computational docking studies [3].

5. ACKNOWLEDGMENTS

S.S (Ph.D. Research Scholar) wants to thanks A. S. (M.Tech trainee), M. A (M.Tech trainee) & A. P (M.Sc trainee) for carrying out extensive bioinformatics analysis. S.S and K.K did the preparation of Manuscript, referencing and critical editing. The authors also want to thanks S.S and K.K for support and suggestions for carrying out this work.

6. REFERENCES

[1] Alejandro AS, Aravind L, Thomas LM, Sergei S, John LS, Yuri IW, Eugene VK, Stephen F A. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 29(14), 2994-3005, (2001).

[2] Alex B, Lachlan C., Richard D, Robert DF., Volker H, Sam GJ, Ajay K, Mhairi M, Simon M, Erik LLS., David JS., Corin Y, Sean RE. The Pfam families' database. Nucleic Acids Research, Vol. 32, D138-D141, (2004).

[3] Allardet-Servent A, Michaux-Charachon S, et al. (1993).Presence of one linear and one circular chromosome in the Agrobacterium tumefaciens C58 genome. Journal of bacteriology 175(24): 7869-7874.

[4] Altschul SF., Madden TL., Schaffer AA., Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389-402, (1997).

[5] Aron MB, John BA, Myra KD, Carol DS, Noreen RG, Marc G, Luning H, Siqian H, David IH, John DJ, Zhaoxi K, Dmitri K, Christopher JL, Cynthia AL, Chunlei L, Fu L, Shennan L, Gabriele HM, Mikhail M, James SS., Narmada T, Roxanne AY., Jodie JY, Dachuan Z, Stephen HB. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Research, Vol. 35, D237–D240, (2006).

[6] Binns, AN and Thomashow MF. Cell biology of Agrobacterium infection and transformation of plants. Annual Reviews in Microbiology 42(1): 575-606 (1988).

[7] Cédric N, Desmond GH, Jaap H. T-coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205-217, (2000).

[8] Chih-Chieh C, Jenn-Kang H, Jinn-Moon Y (PS)2: protein structure prediction server Nucl. Acids Res. 34, W152-W157, (2006).

[9] Dessaux Y, Petit A, Farrand SK, Murphy PJ, in The Rhizobiaceae: Molecular Biology of Model Plant-Associated Bacteria, H. P. Spaink, A. Kondorosi, P. J. J. Hooykaas, Eds. (Kluwer Academic, Dordrecht, Netherlands, 1998), pp. 173Ð197.

[10] Edward E, Gary LG., Osnat H, John M, John O, Roberto JP, Linda B, Delwood R., Andrew J H. Biological function made crystal clear- annotation of hypothetical proteins via structural genomics. Current Opinion in Biotechnology 11, 25-30, (2000).

[11] Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A , Liu F, Wollam C, Allinger M, Doughty D , Scott C , Lappas C, Markelz B , Flanagan C, Crowell C, Gurson J, Lomo,C., Sear C, Strub G, Cielo C and Slater S. Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58. Science 294 (5550), 2323-2328 (2001)

[12] Hamilton R and Fall M. The loss of tumor-initiating ability inAgrobacterium tumefaciens by incubation at high temperature. Cellular and Molecular Life Sciences **27**(2): 229-230 (1971).

[13] Paunikar WN, Sanmukh SG, and Ghosh TK. Exploring the hypothetical proteins in Rizhophages and their role in influencing Rhizobium species in soil. CiiT International Journal of Artificial Intelligent systems and Machine Learning (2011).

[14] Rosenberg C. and Huguet T. The pAtC58 plasmid of Agrobacterium tumefaciens is not essential for tumour induction. Molecular and General Genetics MGG **196**(3): 533-536 (1984).

[15] Sanmukh SG and Paunikar WN. Deciphering unknown proteins in Human Herpes Viruses. CiiT International Journal of Automation and Autonomous System (2012).

[16] Sanmukh SG and Paunikar WN. Study of hypothetical proteins in Shigella phages. CiiT International Journal of fuzzy Systems (2011).

[17] Sanmukh SG and Paunikar WN. Study of prophages in Lactobacillus species. CiiT International Journal of Automation and Autonomous System (2012).

[18] Sanmukh SG and Paunikar WN. Understanding Mycobacteriophages through their unrevealed proteins. CiiT International Journal of Fuzzy Systems (2012).

[19] Sanmukh SG and Paunikar WN. Yersinia Phages and their Novel Proteins. CiiT International Journal of Data Mining and Knowledge Engineering (2012).

[20] Sanmukh SG, Meshram DB, Paunikar WN, Ghosh TK. Computational characterizations for structure and function of unclassified proteins in Ictalurus punctatus. CiiT International Journal of Artificial Intelligent Systems and Machine Learning (2011).

[21] Sanmukh SG, Paunikar WN and Ghosh TK and Chakrabarti T. Structural and functional prediction of hypothetical proteins in Bacteriophages against Halophilic bacteria - An Insilico Approach. International Journal of Pharma and Biosciences (2011).

[22] Sanmukh SG, Paunikar WN and Ghosh TK. Study of Hypothetical Proteins in Salmonella Phages and Predicting their Structural and Functional Relationship. CiiT International Journal of Biometrics and Bioinformatics (2011).

[23] Sanmukh SG, Paunikar WN, Ghosh TK and Chakrabarti T. Structure and Function Predictions of Hypothetical Proteins in Vibrio Phages. International Journal of Biometrics and Bioinformatics. 4(5), pp 161-175 (2010).

[24] Sanmukh SG, Paunikar WN, Ghosh TK. Computational approach for structure and functionality search for hypothetical proteins in Mycobacterium leprae CiiT International Journal of Data Mining and Knowledge Engineering (2011).

[25] Sanmukh SG, Paunikar WN, Meshram DB and Ghosh TK. Functionality search in hypothetical proteins of Halobacterium salinarum CiiT International Journal of fuzzy Systems (2011).

[26] Sanmukh SG, Paunikar WN, Meshram DB and Ghosh TK. Insilico function prediction for hypothetical proteins in Vibrio parahaemolyticus Chromosome II. CiiT International Journal of Data Mining and Knowledge Engineering (2011).

[27] Sanmukh SG, Rahman M and Paunikar WN. Comparative Genomic Studies of hypothetical proteins in Cyanophages International Journal of Computer Applications. 45(15):16-33 (2012).

[28] Smith EF and Townsend C. A plant-tumor of bacterial origin. Science **25**(643): 671-673(1907).

[29] WoodDW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF Jr, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D Sr, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV and Nester EW. The genome of the natural genetic engineer Agrobacterium tumefaciens C58. Science 294 (5550), 2317-2323 (2001).

[30] Yu CS, Lin CJ, et al. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science 13(5): 1402-1406(2004).

[31] Zafer A, Yucel A, Mark B. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, BMC Bioinformatics ,7, 178, (2006).

[32] Zdobnov EM, Rolf A. Interproscan- an integration platform for the signatures recognition methods in InterPro. Bioinformatics 17,847-848, (2001).

Table 1. Predicted Structures, annotations and sub-cellular localization of hypothetical proteins in Agrobacterium fabrum str. C58 plasmid At through comparative genomic approach

Gene ID	CDD-Blast	Interproscan	Pfam	Cello	PS^2 structure
1136781	Uncharacterized conserved protein [Function unknown]	no	Protein of unknown function DUF72	Cytoplasmic 3.275 *	1vpqA- 24 -221 -1e-58
1136783	Flavodoxin_2 super family[cl00438], Flavodoxin-like fold; This family consists of a domain with a flavodoxin-like fold. The family includes ..	Flavodoxin_2	Flavodoxin-like fold	Cytoplasmic 2.899 *	1f9zA- 30 -162 -1e-41
1136792	TatD like proteins; E.coli TatD is a cytoplasmic protein, shown to have magnesium dependent DNase activity.	DNase_TatD	TatD related Dnase	Cytoplasmic 3.837 *	1zzmA -25-258- 8e-70
1136793	Adenine nucleotide alpha hydrolases superfamily including N type ATP PPases, ATP .	no	Queuosine biosynthesis protein QueC	Cytoplasmic 2.654 *	2pg3A -15 -37- 0.007
1136795	KAP_NTPase[pfam07693], KAP family P-loop domain,Predicted P-loop ATPase	KAP_NTPase	KAP family P-loop domain	Cytoplasmic 2.429 * InnerMembrane 2.093 *	3ctpA -15 -34 -8e-04
1136796	topoisomerase-primase (TOPRIM) nucleotidyl transferase/hydrolase domain,ATP-binding cassette domain of Rad50; The catalytic domains of Rad50 are similar to the ATP-binding cassette of ABC, Predicted ATP-dependent endonuclease of the OLD family [DNA replication, recombination	SSF52540	AAA domain,AAA ATPase domain	Cytoplasmic 4.670 *	2o5vA -15- 67 -9e-14
1136798	super family[cl17451]Homeodomain-like domain;RNA polymerase sigma factor, sigma-70 family	G3DSA:1.10.10.60	Homeodomain-like domain	Cytoplasmic 2.496 *	1u78A -12- 40- 2e-04
1136799	no	Acyl_CoA_acyltransferase,G3DSA:3.40.630.30	no	Cytoplasmic 3.287 *	1ro5A -18 -46 -4e-06
1136813	ParB-like nuclease domain,RepB plasmid partitioning protein,sporulation protein J (antagonist of Soj) containing ParB-like nuclease domain	ParBc,ParBc	ParB-like nuclease domain,RepB plasmid partitioning protein	Cytoplasmic 4.263 *	1vz0B -20- 181 -2e-46
1136822	Ku-core domain, Ku-like subfamily;Ku-homolog [Replication, recombination, and repair]	Ku	Ku70/Ku80 beta-barrel domain	Cytoplasmic 3.683 *	1jeqA -14 -185- 1e-47
1136823	Ku-core domain, Ku-like subfamily,Ku-homolog [Replication, recombination, and repair]	Ku	Ku70/Ku80 beta-barrel domain	Cytoplasmic 4.240 *	1jeqA -16 -175 -1e-44
1136839	methyl indole-3-acetate methyltransferase,Alpha/beta hydrolase family	Abhydrolase_6, PTHR10992:SF201,PTHR10992	Alpha/beta hydrolase family	Periplasmic 2.518 *	1va4A -20 -125 -9e-30
1136849	Uncharacterized protein conserved in bacteria	DUF1537	Protein of unknown function, DUF1537	Cytoplasmic 3.510 *	1yzyA -20 -297- 2e-81
1136845	BNR repeat-like domain	Sialidase	BNR repeat-like domain	Periplasmic 2.816 *	1w8oA -16 -160 -4e-40

1136859	, KaiC is a circadian clock protein, RecA-superfamily ATPases implicated in signal transduction	AAA	KaiC	Cytoplasmic 3.983 *	2gblA -23 -317 -5e-87
	Uncharacterized conserved protein	no	no	Cytoplasmic 1.803 * Periplasmic 1.315 *	1ofuY -14 -35 -0.004
1136956	NADP oxidoreductase coenzyme,Predicted dinucleotide-binding enzymes	G3DSA:3.40.50.720	NADP oxidoreductase coenzyme F420-dependent	InnerMembrane 2.925 *	2vq3A- 25 -154 -6e-39
1136962	Uncharacterized enzyme involved in biosynthesis of extracellular polysaccharides,	ABM	Antibiotic biosynthesis monooxygenase	Cytoplasmic 2.771 *	3bm7A -20 -59 -4e-10
1136963	Alpha/beta hydrolase family; This family contains alpha/beta hydrolase enzymes of diverse specificity	Abhydrolase_6	Alpha/beta hydrolase family	Cytoplasmic 4.760 *	2vf2A -18 -151 -1e-37
1136964	Oxidoreductase family, NAD-binding Rossmann fold	GFO_IDH_MocA	Oxidoreductase family, NAD-binding Rossmann fold	Cytoplasmic 3.842 *	2nvwA -24 -174 -2e-44
1136982	SnoaL-like domain	SSF54427,G3DSA:3.10.450.50	SnoaL-like domain	Cytoplasmic 2.534 *	2bngC -11 -51 -1e-07
1136985	Serine hydrolase,Alpha/beta hydrolase family	Abhydrolase_6, SSF53474	Alpha/beta hydrolase family,Domain of unknown function (DUF4350)	Periplasmic 3.119 *	1xklA-21 -164 -1e-41
1136996	Amino acid synthesis	AA_synth	Amino acid synthesis	Cytoplasmic 3.046 *	3byqB -31 -232- 5e-62
1136997	NIPSNAP; Members of this family include many hypothetical proteins	Dimer_A_B_barrel	NIPSNAP	Cytoplasmic 3.132 *	1vqyB -100 -135 -2e-33
1137003	PucR C-terminal helix-turn-helix domain, GAF domain	GAF	PucR C-terminal helix-turn-helix domain,GAF domain	Cytoplasmic 3.155 *	3ci6B- 19 -97- 1e-20
1137011	M14_ASTE_ASPA_like_2[cd06252], Peptidase M14 Succinylglutamate desuccinylase (ASTE)/aspartoacylase (ASPA)-like,Predicted deacylase	ASP,AstE_AspA	Succinylglutamate desuccinylase / Aspartoacylase family,Uncharacterized protein conserved in archaea (DUF2119)	Cytoplasmic 3.264 *	3cdxA -32- 224- 2e-59
1137025	Aminoglycoside 3'-phosphotransferase (APH) and Choline Kinase (ChoK) family,Phosphotransferase enzyme family	APH	Start End From To APH Phosphotransferase enzyme family	Cytoplasmic 4.344 *	1zylA -18- 204 -2e-53
1137051	YciF bacterial stress response protein, ferritin-like iron-binding domain	Ferritin/RR_like	Domain of unknown function (DUF892),RAM signalling pathway proteinEndoplasmic Reticulum Oxidoreductin 1 (ERO1),Protein of unknown function (DUF1488)	Cytoplasmic 4.095 *	2gyqA- 43 -171 -1e-43
1137066	Uncharacterized protein family (UPF0093)	(UPF0093)	Uncharacterised protein family (UPF0093)	InnerMembrane 4.505 *	2p9iB -20 -36 -0.006
1137086	Siderophore-interacting protein	SIP	Siderophore-interacting FAD-binding domain,Siderophore-interacting protein	Cytoplasmic 4.497 *	2gpjA -24 -229- 3e-61
1137095	arsenical resistance protein ArsH,NADPH-dependent FMN reductase;	FMN_red	NADPH-dependent FMN reductase	Cytoplasmic 1.808 * Periplasmic 1.771 *	2q62D -84- 277 -2e-75

1137096	MFS[cd06174], The Major Facilitator Superfamily (MFS) ,MFS_1[pfam07690], Major Facilitator Superfamily;	MFS_1	Major Facilitator Superfamily	InnerMembrane 4.973 *	1pw4A -12 -104 -4e-23
1137102	Demethylmenaquinone methyltransferase,[PRK06201], hypothetical protein	Methyltransf_6, RNaseE_inh/diMeMenaQ_MeTrfase	Demethylmenaquinone methyltransferase	Cytoplasmic 3.882 *	2c5qE -24 -120- 1e-28
1137103	Uncharacterized protein conserved in bacteria	YflP,TctC	Tripartite tricarboxylate transporter family receptor	Cytoplasmic 2.570 *	2qpqA -28 -367 -1e-102
1137104	Tripartite tricarboxylate transporter TctB family	TctB	Tripartite tricarboxylate transporter TctB family	InnerMembrane 4.502 *	2b5uA- 15 -39 -0.001
1137105	Uncharacterized protein conserved in bacteria	TctA	Tripartite tricarboxylate transporter TctA family	InnerMembrane 4.970 *	2b5uA- 18 -32 -0.003
1137109	KduI/IolB family	EutQ	Ethanolamine utilisation protein EutQ	Cytoplasmic 1.849 * Extracellular 1.533	2pytA -22 -138 -2e-34
1137115	Uncharacterized conserved protein	DUF849	Prokaryotic protein of unknown function (DUF849)	Cytoplasmic 3.939 *	3c6cA -51 -347 -1e-96
1137122	PAS domain; PAS motifs appear in archaea, eubacteria and eukarya,PAS fold	PAS_3,SSF55785	PAS fold	Cytoplasmic 3.309 *	2v0uA- 18 -35 -0.005
1137132	Uncharacterized conserved protein	no description,DUF1508	Domain of unknown function (DUF1508),Domain of unknown function (DUF4480)	Cytoplasmic 1.709 * Extracellular 1.531 * Periplasmic 1.084 *	3bidE -31 -86 -2e-18
1137138	EAL domain	EAL,SSF141868 ,G3DSA:3.20.20 .450	EAL domain	Cytoplasmic 2.073 * Periplasmic 1.452	2r6oA -37 -119 -1e-28
1137139	Transposase; Transposase proteins are necessary for efficient DNA transposition,Transposase and inactivated derivatives [DNA replication, recombination, and repair]	HTH_Tnp_1,G3DSA:1.10.10.60, Homeodomain_like	Transposase	Cytoplasmic 3.456 *	2jn6A -21 -43 -4e-05
1137145	Diguanylate-cyclase (DGC) or GGDEF domain,c-di-GMP synthetase (diguanylate cyclase, GGDEF domain)	GGDEF	GGDEF domain	InnerMembrane 4.762 *	1w25A -35 -207 -2e-54
1137160	SnoaL-like domain, Uncharacterized protein conserved in bacteria	DUF1348,G3DSA:3.10.450.50,SSF54427	Protein of unknown function (DUF1348)	Cytoplasmic 2.916 *	2imjA- 74- 276 -2e-75
1137180	Xylose isomerase-like TIM barrel	Xyl_isomerase-like_TIM-brl,AP_endonuc_2	Xylose isomerase-like TIM barrel	Cytoplasmic 4.777 *	2hk0A- 15- 91- 5e-19
1137188	Conserved protein/domain typically associated with flavoprotein oxygenases	Flavin_Reduct	Flavin reductase like domain	Cytoplasmic 3.826 *	1ejeA- 33 -157 -1e-39
1137199	Uncharacterized protein conserved in bacteria	SSF69118	DEK C terminal domain	Cytoplasmic 3.155 *	2prrD- 15- 101 -7e-23
1137200	Carboxymuconolactone decarboxylase family	CMD	Carboxymuconolactone decarboxylase family	Cytoplasmic 4.792 *	2prrD -27- 203 -2e-53
1137207	AP endonuclease family 2,Sugar phosphate isomerases/epimerases	Xylose isomerase-like,AP_endonuc_2	Xylose isomerase-like TIM barrel	Cytoplasmic 4.280 *	3cnyA -23 -241 -8e-65
1137212	Putative glycolipid-binding	Glycolipid_bind	Putative glycolipid-binding,Ataxin 2 SM domain	Cytoplasmic 4.344 *	2h1tA -33 -171 -8e-44

1137221	WGR domain of molybdate metabolism regulator and related proteins	WGR	WGR domain	Cytoplasmic 2.842 *	2eocA -25- 86 -3e-18
1137227	Predicted secreted hydrolase	CrtC	Hydroxyneurosporene synthase (CrtC)	Periplasmic 3.163 *	2ichB -36 -350 -4e-97
1137228	Members of the SGNH-hydrolase superfamily	Lipase_GDSL_2,Esterase_SGNH_hydro-type	GDSL-like Lipase/Acylhydrolase family	Cytoplasmic 1.234 * OuterMembrane 1.199 *	1fxwF -20 -145- 4e-36
1137257	Crp-like helix-turn-helix domain;effector domain of the CAP family of transcription factors, cAMP-binding proteins	cNMP	Crp-like helix-turn-helix domain,Cyclic nucleotide-binding domain	Cytoplasmic 3.448 *	1zybA -13 -98 -2e-21
1137248	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	Glyoxalase-like domain	Cytoplasmic 3.579 *	2c21A -22 -58 -5e-10
1137261	Crp-like helix-turn-helix domain;effector domain of the CAP family of transcription factors,cAMP-binding proteins	CNMP_BINDING_3,cNMP	Cyclic nucleotide-binding domain,Crp-like helix-turn-helix domain	Cytoplasmic 4.280 *	1ft9A -17 -109 -5e-25
1137267	YciF bacterial stress response protein, ferritin-like iron-binding domain	Ferritin/RR_like	Domain of unknown function (DUF892),Endoplasmic Reticulum Oxidoreductin 1 (ERO1)	Cytoplasmic 3.189 *	2gyqA -50 -184 -6e-48
1137270	XdhC and CoxI family	XdhC_CoxI,XdhC_C	XdhC and CoxI family,XdhC Rossmann domain	Cytoplasmic 4.427 *	2gqwA -17- 37 -0.006
1137274	GT_2_like_f is a subfamily of the glycosyltransferase family 2 (GT-2) with unknown function	NTP_transf_3,SSF53448	MobA-like NTP transferase domain	Cytoplasmic 3.503 *	1e5kA -24 -128 -7e-31
1137275	Type 1 glutamine amidotransferase (GATase1)-like domain	UCP016642,BPL_N	Biotin-protein ligase, N terminal	Cytoplasmic 2.555 *	1q7rA -24 -50- 8e-07
1137283	KaiC is a circadian clock protein	AAA	KaiC	Cytoplasmic 4.722 *	2gblA -27 -402- 1e-115
1137291	FAD-NAD(P)-binding;	NAD_binding_8	FAD-NAD(P)-binding	Cytoplasmic 2.632 *	1w4xA- 15- 45 -5e-07
1137295	Type I periplasmic ligand-binding domain of uncharacterized ABC,Periplasmic binding protein;	Peripla_BP_6	Periplasmic binding protein	Periplasmic 3.124 *	2e4uB -12 -75 -2e-14
1137311	Predicted membrane protein	no	Domain of unknown function (DUF4126),Glycine-zipper containing OmpA-like membrane domain	InnerMembrane 4.035 *	1ciiA -39 -39 -6e-04
1137314	Fusaric acid resistance protein-like;Predicted membrane protein	FUSC	Fusaric acid resistance protein family	InnerMembrane 4.516 *	2qr4A -16 -30 -0.009
1137317	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily;		Glyoxalase-like domain	Periplasmic 2.026 * Cytoplasmic 1.527 *	3ct8A -19 -83 -3e-17
1137321	5'-phosphate oxidase	Pyridox_oxidase	Pyridoxamine 5'-phosphate oxidase	Cytoplasmic 3.605 *	2asfA- 17 -38- 0.002
1142192	Predicted hydrolase of the alpha/beta superfamily	SSF53474,G3DSA:3.40.50.1820	Prolyl oligopeptidase family	Cytoplasmic 2.010 *	2qm0B -32 -78 -3e-16
1142198	Bacterial SH3 domain;	SH3_3	Bacterial SH3 domain	Periplasmic 2.053 *	2eyzA- 14 -35- 0.010

1142208	beta-Keto acyl carrier protein reductase (BKR), involved in Type II FAS, classical (c) SDRs;	G3DSA:3.40.50.720	Enoyl-(Acyl carrier protein) reductase	Cytoplasmic 2.623 *	1o5iA -55 -40 -1e-04
1142561	Domain of unknown function (DUF892);	Ferritin/RR_like, DUF892	Domain of unknown function (DUF892)	Cytoplasmic 3.046 *	2gyqA- 48 -53 -1e-08
1142619	Cold shock proteins [Transcription]	no	Cold-shock' DNA-binding domain	Periplasmic 2.014 * Cytoplasmic 1.913 *	1mjcA -45 -37 -0.001
1136861	Uncharacterized conserved protein	no	no	Cytoplasmic 1.803 * Periplasmic 1.315 *	1ofuY -14 -35- 0.004
1136862	Uncharacterized conserved protein	PROKAR_LIPOPROTEIN	HAMP domain	OuterMembrane 2.969 *	3c8cA -15 -63- 1e-10
1136869	Uncharacterized conserved protein	SSF143081,DUF159	Uncharacterised ACR, COG2135	Periplasmic 2.660 * Cytoplasmic 2.108 *	2aegA -100 -263 -3e-71
1136871	Uncharacterized conserved protein	UCP034285,SSF52540	no	Cytoplasmic 3.580 *	2cvhA- 17 -40- 4e-04
1136872	Nucleotidyltransferase/DNA polymerase involved in DNA repair,DNA Polymerase Y-family	IMS,SSF56672	impB/mucB/samB family	Cytoplasmic 3.714 *	1t94B -17 -233 -3e-62
1136874	Protein of unknown function (DUF1419)	DUF1419	Protein of unknown function (DUF1419)	Cytoplasmic 3.868 *	2hzqA -18- 36 -0.003
1136880	WGR domain of molybdate metabolism regulator and related proteins;	WGR	WGR domain	Cytoplasmic 1.563 * Extracellular 1.391 * Periplasmic 1.155 *	2cr9A 24 51 4e-08
	Higher Eukarytoes and Prokaryotes Nucleotide-binding domain;Nucleotidyltransferase (NT) domain of Staphylococcus aureus kanamycin nucleotidyltransferase, and similar protei	NTP_transf_2,HEPN	Nucleotidyltransferase domain,M26 IgA1-specific Metallo-endopeptidase C-terminal region,FtsK/SpoIIIE family	Cytoplasmic 4.876 *	2rffA -19 -56 -2e-08
1136891	7-keto-8-aminopelargonate synthetase and related enzymes,hypothetical protein;	Aminotran_1_2, G3DSA:3.40.640.10	Aminotransferase class I and II	Cytoplasmic 3.960 *	1fc4A -26 -499 -1e-142
	Predicted secreted hydrolase	CrtC,no description	Hydroxyneurosporene synthase (CrtC)	Extracellular 2.273 * Periplasmic 1.901 *	2ichB- 34 -347 -3e-96
1136913	S-adenosylmethionine-dependent methyltransferases (SAM or AdoMet-MTase)	Methyltransf_12	Methyltransferase domain	Cytoplasmic 2.832 *	1im8A- 20- 203 -1e-53
1136921	Organic hydroperoxide reductase	OsmC	OsmC-like protein	Cytoplasmic 3.172 *	2ql8A -22- 66 -4e-12