BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(15)

Computational structural and functional analysis of hypothetical proteins of *Staphylococcus aureus*

Ramadevi Mohan & Subhashree Venugopal*

Division of Biomolecules and Genetics, School of Biosciences and Technology, VIT University, Vellore-632014, Tamil Nadu, India; Subhashree Venugopal – Email: vsubhashree@vit.ac.in; *Corresponding author

Received July 17, 2012; Accepted July 23, 2012; Published August 03, 2012

Abstract:
Genome sequencing projects has led to an explosion of large amount of gene products in which many are of hypothetical proteins with unknown function. Analyzing and annotating the functions of hypothetical proteins is important in *Staphylococcus aureus* which is a pathogenic bacterium that cause multiple types of diseases by infecting various sites in humans and animals. In this study, ten hypothetical proteins of *Staphylococcus aureus* were retrieved from NCBI and analyzed for their structural and functional characteristics by using various bioinformatics tools and databases. The analysis revealed that some of them possessed functionally important domains and families and protein-protein interacting partners which were ABC transporter ATP-binding protein, Multiple Antibiotic Resistance (MAR) family, export proteins, Helix-Turn-helix domains, arsenate reductase, elongation factor, ribosomal proteins, Cysteine protease precursor, Type-I restriction endonuclease enzyme and plasmid recombination enzyme which might have the same functions in hypothetical proteins. The structural prediction of those proteins and binding sites prediction have been done which would be useful in docking studies for aiding in the drug discovery.

Keywords: Hypothetical proteins, *Staphylococcus aureus*, Functional analysis, Bioinformatics tools

Background:
Staphylococcus aureus is a gram-positive pathogen that is a major cause of multiple types of infections both in and outside of the hospital setting. These infections range from superficial skin infections to deeper infections of hair follicles, abscesses, and deep tissue infections, and even to systemic infections including those of the heart, lungs, bones, and blood **[1]**. Although the organism is part of the normal human flora, it can cause infection when there is a break in the skin or mucous membrane that grants it access to the surrounding tissues **[2-4]**. In the pre-antibiotic era, these infections were often life threatening, and even today they may give rise to death despite treatment with antibiotics. S. aureus strains can produce a number of different components that may contribute to virulence, including surface-associated adhesins, capsular polysaccharides, exoenzymes, and exotoxins. *Staphylococcus aureus* carries a large repertoire of virulence factors, including over 40 secreted proteins and enzymes that it uses to establish and maintain infections **[5, 6].** Some of these virulence factors are known to cause or be associated with specific diseases, for example, endocarditis and osteomyelitis, septic arthritis, and septicemia, toxic shock syndrome toxin (TSST) and toxic shock syndrome; Panton-Valentine leukocidin (PVL) and necrotizing pneumonia and skin diseases **[7, 8]**; the exfoliative toxins A and B (ETA and ETB) and scalded skin syndrome, impetigo, skin infections, and atopic dermatitis **[8, 9]**; and the family of staphylococcal enterotoxins A and B(SEA and SEB) and food poisoning **[6]**.

There is a growing need for the automatic annotation of proteins of unknown functional, termed ''hypothetical proteins'' **[10]**, the structures of which are known **[11]**. Structural genomics initiatives provide ample structures of hypothetical proteins at an ever increasing rate. However without function annotation, this structural goldmine is of little

use to biologists who are interested in particular molecular systems. The structures of many hypothetical proteins are solved in pipelines at structural- genomics centers, which usually lack the resources to engage in thorough functional characterization of each of the solved structures. Moreover, some of the proteins, which are considered to be well annotated, may have additional functions beyond their listed records. About half the proteins in most genomes are candidates for HPs **[12].** This group is of utmost importance to complete genomic and proteomic information. Detection of new HPs not only offers presentation of new structures but also new functions. There will be new structures with so far unknown conformations and new domains and motifs will be arising. A series of additional protein pathways and cascades will be revealed, completing our fragmentary knowledge on the mosaic of proteins per se. The network of protein–protein interactions will be increasing logarithmically. New HPs may be serving as markers and pharmacological targets.

Last not least, detection of HP would be of benefit to genomics enabling the discovery of so far unknown or even predicted genes **[10]**. Hypothetical protein is a protein that is predicted to be expressed from an open reading frame, but for which there is no experimental evidence of translation. Hypothetical proteins constitute a substantial fraction of proteomes of human as well as of other eukaryotes. With the general belief that the majority of hypothetical proteins are the product of pseudogenes, it is essential to have a tool with the ability of pinpointing the minority of hypothetical proteins with a high probability of being expressed **[13]**. There is so far no classification of hypothetical proteins (HPs) and working terms are replacing definitions of hypothetical proteins. In the strict sense, HPs are predicted proteins, proteins predicted from nucleic acid sequences and that have not been shown to exist by experimental protein chemical evidence. Moreover, these proteins are characterized by low identity to known, annotated proteins. Conserved hypothetical proteins are defined as a large fraction of genes in sequenced genomes encoding those that are found in organisms from several phylogenetic lineages but have not been functionally characterized and described at the protein chemical level **[14, 15]**. These structures may represent up to half of the potential protein coding regions of a genome.

Methodology:

Sequence retrieval

Randomly retrieved 10 hypothetical protein sequences of *Staphylococcus aureus* from NCBI **[16]** and were used in this study. The sequence IDs of those 10 hypothetical proteins were gi|166409299, gi|166409303, gi|166409302, gi|166409301, gi|166409300, gi|390516769, gi|166409293, gi|390516759, gi|390516760 and gi|166409294. To analyze the hypothetical proteins and assign their physicochemical and structural and functional properties, various bioinformatics tools and databases were used.

Physicochemical and functional characterization

For physicochemical characterization, theoretical Isoelectric point (pI), molecular weight, total number of positive and negative residues, extinction coefficient **[17]**, instability index **[18]**, aliphatic index **[19]** and grand average hydropathy (GRAVY) **[20]** were computed using the Expasy's Protparam server **[21]**.

PFAM

Pfam **[22, 23]** is a collection of multiple protein-sequence alignments and HMMs, and provides a good repository of models for identifying protein families, domains and repeats. There are two parts to the pfam database: Pfam A, a set of manually curated and annotated models; PfamB, which has higher coverage but is fully automated (with no manual curation). Pfam B HMMs are created from alignments generated by ProDom in the automatic clustering of the protein sequences in SWISS-PROT and TrEMBL.

CDD-BLAST

CD-Search **[24]** is NCBI's interface to searching the Conserved Domain Database with protein query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query **[25]**.

Protein-Protein interactions prediction

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins,) **[26]** is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: Genomic Context, High-throughput Experiments, (Conserved) Co-expression and Previous Knowledge. STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 5,214,234 proteins from 1133 organisms **[27]**.

Prediction of transmembrane proteins

SOSUI server is used to characterize whether the protein is soluble or transmembrane in nature **[28]**.

Stability of proteins

DISULFIND **[29, 30]** is a server for predicting the disulfide bonding state of cysteines and their disulfide connectivity starting from sequence alone. Disulfide bridges play a major role in the stabilization of the folding process for several proteins. Prediction of disulfide bridges from sequence alone is therefore useful for the study of structural and functional properties of specific proteins. In addition, knowledge about the disulfide bonding state of cysteines may help the experimental structure determination process and may be useful in other genomic annotation tasks.

Protein structure prediction

Online server PS^2 (PS Square) Protein Structure Prediction Server **[31]** was used **[32-36]** which accepts the protein query sequences in FASTA format and uses the strategies of Pair-wise and multiple alignment by combining powers of the programs PSI-BLAST, IMPALA and T-COFFEE in both target – template selection and target–template alignment and finally it constructs the protein 3D structures using integrated modeling package of PS2 using best scored orthologous template.

Q-Site Finder

Q-Site Finder **[37]** is a new method of ligand binding site prediction. It works by binding hydrophobic (CH_3) probes to the protein, and finding clusters of probes with the most favorable binding energy. These clusters are placed in rank order of the likelihood of being a binding site according to the

sum total binding energies for each cluster. Q-Site Finder was shown to identify sites with high precision. The advantage of this is that putative binding sites are identified as closely as possible to the actual binding site. It uses the interaction energy between the protein and a simple Vander Waal's probe to locate energetically favorable binding sites. Energetically favorable probe sites are clustered according to their spatial proximity and clusters are then ranked according to the sum of interaction energies for sites within each cluster **[38]**.

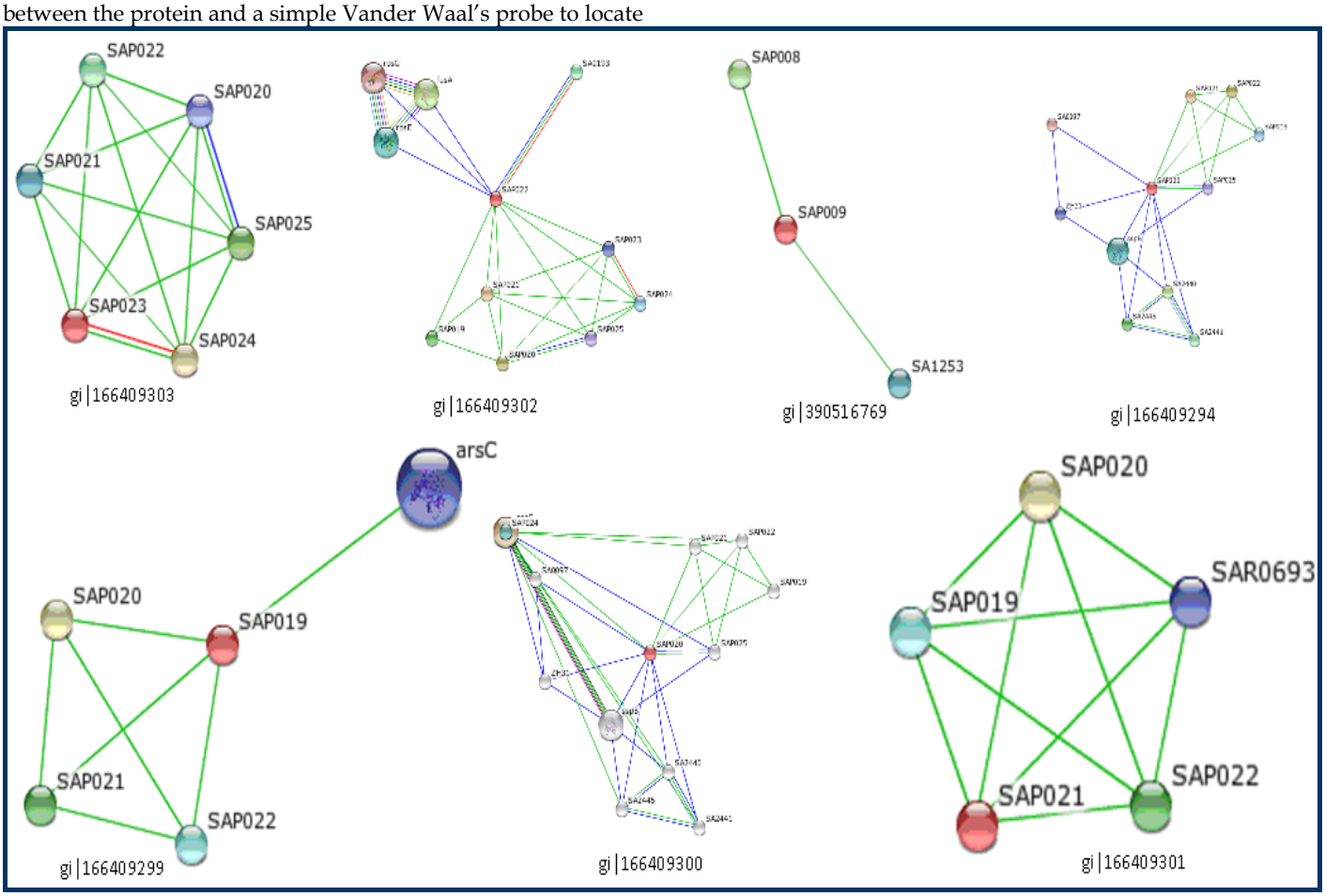

Figure 1: Protein-protein interactions of hypothetical proteins predicted by STRING tool

Discussion:

The physicochemical properties of hypothetical proteins were tabulated in **Table 1 (see supplementary material).** The calculated isoelectric point (pI) will be useful because at pI, solubility is least and mobility in an electro focusing system is zero. Isoelectric point (pI) is the pH at which the surface of protein is covered with charge but net charge of protein is zero. At pI, proteins are stable and compact. The computed isoelectric point (pI) will be useful for developing buffer system for purification by isoelectric focusing method. Although Expasy's Protparam computes the extinction coefficient for 276, 278, 279, 280 and 282 nm wavelengths, 280 nm is favored because proteins absorb light strongly there while other substances commonly in protein solutions do not. Extinction coefficient of hypothetical proteins homologue at 280 nm is ranging from 1490 to 77825 M cm with respect to the concentration of Cys, Trp and Tyr. The high extinction coefficient of hypothetical proteins indicates presence of high concentration of Cys, Trp and Tyr. The computed extinction coefficients help in the quantitative study of protein–protein and protein–ligand interactions in solution. The instability index provides an estimate of the stability of protein in a test tube. There are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones. This method assigns a weight value of instability. Using these weight values it is possible to compute an instability index (II). A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable **[18]**. The instability index value for the hypothetical proteins was found to be ranging from 7.98 to 64.89. The stable proteins were gi|166409302, gi|166409301, gi|166409300, gi|390516769, gi|166409293, gi|390516759 and gi|166409294 and the other proteins were unstable. The aliphatic index (AI) which is defined as the relative volume of a protein occupied by aliphatic side chains (A, V, I and L) is regarded as a positive factor for the increase of thermal stability of globular proteins. Aliphatic index for the hypothetical proteins sequences ranged from 65.36 to 138.39. The very high aliphatic index of the protein sequences indicates that these proteins may be stable for a wide temperature range. The lower thermal stability of proteins was indicative of a more flexible structure when compared to other protein. The Grand Average hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence. GRAVY indices of hypothetical proteins are ranging from -0.172. This low range of value indicates the possibility of better interaction with water.

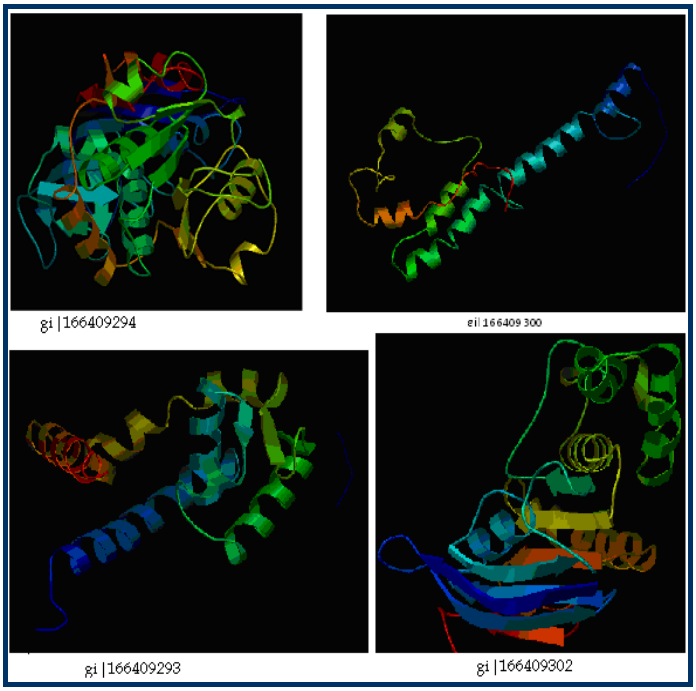

Figure 2: Structures of *S. aureus* hypothetical proteins modeled by PS SQUARE server

Functional analysis of these proteins includes protein domains and family prediction and prediction of trans-membrane regions, disulfide bond and identification of sub-cellular localization sites. Domains can be thought of as distinct functional and/or structural units of a protein. These two classifications coincide rather often, as a matter of fact, and what is found as an independently folding unit of a polypeptide chain also carries specific function. Domains are often identified as recurring (sequence or structure) units, which may exist in various contexts. In molecular evolution such domains may have been utilized as building blocks, and may have been recombined in different arrangements to modulate protein function **[24].** The proteins were classified into particular family based on the presence of specific domain in the sequence. Out of 10 hypothetical proteins, 7 proteins possessed specific domains in them which were lactococcin_972, Mob_Pre, L_ocin_972_ABC, DUF 1093 & 1430, COG4652, ABC_MJ0796_Lo1CDE_FtsE, HTH, HTH_MARR, oxido_YhdH and MDR_yhdh_yhfp domains and they were classified as super-families accordingly. Most of these possessed functionally important domains in them except the sequences with id gi|166409301 and gi|166409300 which had domains of unknown function. There were no domains in the other 3 proteins. The presence of these domains in the hypothetical proteins reveals that the proteins might be involved in performing the same function. The domains of the hypothetical proteins and their super-family descriptions were given in **Table 2 & Table 3 (see supplementary material).**

The study of subcellular localization is important for elucidating protein functions involved in various cellular processes. Knowledge of the subcellular localization of a protein can significantly improve target identification during the drug discovery process. The localization site of the hypothetical proteins selected in this study was predicted by PSORTB and they were tabulated in **Table 4 (see supplementary material).** Cytoplasmic membranes were found to be preferred site for performing functions in these proteins as they were seen in most of the proteins involved in this study. Multiple localization sites were found in sequences with id gi|166409303, gi|166409301 and gi|390516759 in which the targeting sites might be of anyone of Cytoplasmic, Cytoplasmic membrane, extracellular and cell wall.

Pfam database search made to identify domains and families present in hypothetical proteins **Table 5 & Table 6 (see supplementary material).** They were Zinc-binding dehydrogenase family, MAR family, TMEM9 and ABC-transporters. SOSUI distinguishes between membrane and soluble proteins from amino acid sequences, and predicts the transmembrane helices for the former. The server SOSUI classified 3 hypothetical proteins as transmembrane proteins having transmembrane helices atleast one in each and maximally six transmembrane regions were found in the protein, gi|166409300. The transmembrane regions type and their length were tabulated in **Table 7 (see supplementary material).** All the seven other proteins were soluble ones. The transmembrane regions are rich in hydrophobic amino acids. Thus there was higher number of hydrophobic amino acid residues in the transmembrane proteins. When those hypothetical proteins were analyzed for disulphide bridges by DISULPHIND server to predict the thermo stability of the proteins, it revealed no disulphide bonds in any of those proteins which revealed that they were thermally unstable.

Protein-protein interactions (PPI) are essential for almost all cellular functions. Proteins often interact with one another in a mutually dependent way to perform a common function. As an example, the transcription factors interact among themselves to bring about transcription. It is therefore possible to infer the functions of proteins based on their interaction partners. Proteins seldom carry out their function in isolation; rather, they operate through a number of interactions with other biomolecules. Experimental elucidation and computational analysis of the complex networks formed by individual protein-protein interactions (PPIs) are one of the major challenges in the post-genomic era. Protein-protein interaction databases have become a major resource for investigating biological networks and pathways in cells **[39].** The protein with ID gi|166409301 was found to have interaction with ABC transporter ATP binding protein. Gi|166409299 had interaction with arsenate reductase protein which reduces arsenate [As (V)] to arsenite [As (III)] and dephosphorylates tyrosine and thus the protein might involve in the enzymatic function of the protein. Gi|166409302 showed interactions with three proteins which were A) Elongation factor G which promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome and also has Vitronectin-binding activity; B) 30S ribosomal protein S5 which plays an important role in translational accuracy with S4 and S12; C) 30S ribosomal protein S7; being one of the primary rRNA binding proteins, it binds directly to 16S rRNA where it nucleates assembly of the head domain of the 30S subunit; it is located at the subunit interface close to the decoding center, probably blocks exit of the E-site tRNA. Gi|166409300 was found to interact with cysteine protease precursor; Cysteine protease is able to degrade elastin, fibrogen, fibronectin and kininogen. It exhibits a strong preference for substrates where arginine is

preceded by a hydrophobic amino acid and also promotes detatchment of primary human keratinocytes. Along with other extracellular proteases the protein is involved in the colonization and infection of human tissues.

Gi|166409293 had interaction with carboxy-terminal processing proteinase CtpA. Gi|166409294 showed interaction with quinone oxidoreductase putative YhdH/YhfP. The other interacting proteins were hypothetical proteins. The protein-protein interacting networks of the hypothetical proteins were given in **(Figure 1) & Table 8 (see supplementary material).** Thus those hypothetical proteins could have the functions of their interacting proteins. The three dimensional structures of the hypothetical proteins were modeled by PS Square server **(Figure 2).** Of the eleven hypothetical proteins, PS Square server could model only four proteins. Since there was low sequence identity, the remaining six proteins could not be modeled. The templates used by the server to model those proteins were tabulated in **Table 9 (see supplementary material).** Identifying the location of ligand binding sites on a protein is of fundamental importance for a range of applications including molecular docking, de novo drug design and structural identification and comparison of functional sites. Active site residues of the hypothetical proteins were given in **Table 10 (see supplementary material).** The active binding site residues would be helpful for docking with specific ligand to study the binding interactions between them.

Conclusion:

There is a need to annotate and find the structural and functional properties of hypothetical proteins in the pathogenic bacteria *Staphylococcus aureus* which produce many virulence factors and cause serious infections and disease. We retrieved 10 hypothetical proteins from NCBI database and characterized its physicochemical properties and identified domains and families using various bioinformatics tools and databases. The structures were modeled and their ligand binding sites were identified. The analysis revealed functionally important domains and families which were involved in inducing protein synthesis and multiple antibiotic resistances in the bacteria and also perform enzymatic functions. This also would provide useful solution for drug discovery for those proteins which were involved in multiple antibiotic resistance and disease mechanisms.

References:

[1] Lowy FD, *N Engl J Med*. 1998 **339**: 520 [PMID: 9709046]
[2] Cheung AL *et al. FEMS Immunol Med Microbiol.* 2004 **40**: 1 [PMID: 14734180]
[3] Lindsay JA & Holden MT, *Trends Microbiol*. 2004 **12:** 378 [PMID: 15276614]
[4] Goldenberg DL & Reed JI, *N Engl J Med*. 1985 **312**: 764 [PMID: 3883171]
[5] Diep BA *et al. J Infect Dis*. 2006 **19**: 1495
[6] Arbuthnott JP *et al. Soc Appl Bacteriol Symp Ser*. 1990 **19**: 107S [PMID: 2119059]
[7] Gillet Y *et al. Lancet*. 2002 **359**: 753 [PMID: 11888586]
[8] Lina G *et al. Clin Infect Dis*. 1997 **25**: 1369 [PMID: 9431380]
[9] Capoluongo E *et al. J Dermatol Sci*. 2001 **26**: 145
[10] Lubec G *et al. Prog Neurobiol*. 2005 **77**: 90 [PMID: 16271823]
[11] Friedberg I, *Brief Bioinform*. 2006 **7**: 225 [PMID: 16772267]
[12] Minion FC *et al. J Bacteriol*. 2004 **186**: 7123 [PMID: 15489423]
[13] Claus D *et al. BMC Bioinformatics*. 2009 **10**: 289
[14] Galperin MY & Koonin EV, *Nucleic Acids Res*. 2004 **32**: 5452 [PMID: 15479782]
[15] Galperin MY, *Funct Genomics.* 2001 **2**: 14 [PMID: 18628897]
[16] http://www.ncbi.nlm.nih.gov/.
[17] Gill SC & Von Hippel PH, *Anal Biochem*. 1989 **182**: 319 [PMID: 2610349]
[18] Guruprasad K *et al. Prot Eng*. 1990 **4**: 155
[19] Ikai AJ, *J Biochem*. 1980 **88**: 1895 [PMID: 7462208]
[20] Kyte J & Doolottle RF, *J Mol Biol.* 1982 **157**: 105 [PMID: 7108955]
[21] http://us.expasy.org/tools/protparam.html.
[22] Bateman A *et al. Nucleic Acids Res*. 2000 **28**: 263 [PMID: 10592242]
[23] http://pfam.sanger.ac.uk/.
[24] http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi/.
[25] Marchler BA *et al. Nucleic Acids Res*. 2011 **39:** 225
[26] http://string.embl.de/.
[27] Szklarczyk D *et al. Nucleic Acids Res*. 2011 **39**: D561 [PMID: 21045058]
[28] http://bp.nuap.nagoya-u.ac.jp/sosui/sosui_submit.html
[29] Ceroni A *et al. Nucleic Acids Res*. 2006 **34**: W177 [PMID: 16844986]
[30] http://disulfind.dsi.unifi.it/.
[31] http://www.ps2.life.nctu.edu.tw/.
[32] Chen-CC *et al. Nucleic Acids Res*. 2006 **34**: W152 [PMID: 16844981]
[33] Altschul SF *et al. Nucleic Acids Res*. 1997 **25**: 3389 [PMID: 9254694]
[34] Schaffer AA *et al. Nucleic Acids Res*. 2001 **29**: 2994 [PMID: 11452024]
[35] Notredame C *et al. J Mol Biol*. 2000 **302**: 205
[36] Wendy B *et al. Nucleic Acids Res*. 2000 **28**: 19
[37] Laurie AT & Jackson RM, *Bioinformatics.* 2005 **21**: 1908 [PMID: 15701681]
[38] http://www.modelling.leeds.ac.uk/qsitefinder/.
[39] Peri S *et al. Genome Res*. 2003 **13**: 2363 [PMID: 14525934]

Edited by P Kangueane
Citation: Mohan & Venugopal, Bioinformation 8(15): 722-728 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Physicochemical properties of hypothetical proteins by Protparam tool

Sequence ID	No of aa	MW	pI	(-) R	(+ R)	EC	II	AI	GRAVY
gi|166409299	97	10407.8	10.11	1	12	25440	43.32	65.36	-0.182
gi|166409303	129	15748.3	10.14	13	31	22920	41.52	83.8	-0.877
gi|166409302	208	23392.4	9.29	27	34	7450	22.22	115.72	-0.148
gi|166409301	103	12163.4	9.73	11	20	11920	22.96	113.4	-0.287
gi|166409300	644	75501.5	9.14	60	73	77825	35.81	119.1	0.128
gi|390516769	31	3677.5	9.3	3	5	1490	7.98	138.39	0.726
gi|166409293	139	15938.3	9.37	12	18	13075	36.4	95.4	-0.443
gi|390516759	209	24226.7	9.25	23	32	19495	23.33	91.87	-0.612
gi|390516760	80	9250.4	4.76	15	11	4470	64.89	101.12	-0.611
gi|166409294	323	35653	6.25	36	34	40340	30.12	105.82	0.004

Table 2: Identification of domains by CDD-BLAST

Sequence ID	Domains
gi|166409299	Lactococcin_972 super family
gi|166409303	Mob_Pre super family
gi|166409302	ABC_MJ0796_Lo1CDE_FtsE, L_ocin_972_ABC
gi|166409301	DUF1093 super family
gi|166409300	DUF1430 super family, COG4652
gi|166409293	HTH_MARR, HTH super family
gi|166409294	MDR_yhdh_yhfp, oxido_YhdH

Table 3: Functional description of superfamilies of hypothetical proteins

Superfamily	Description
Lactococcin_972	Represent bacteriocins related to lactococcin. Associates with a seven transmembrane putative immunity protein.
Mob_Pre	With some plasmids, recombination can occur in a site specific manner that is independent of RecA. In such cases, the recombination event requires another protein called Pre. Pre is a plasmid recombination enzyme. This protein is also known as Mob (conjugative mobilisation).
L_ocin_972_ABC	putative bacteriocin export ABC transporter, lactococcin 972 group; A gene pair with a fairly wide distribution consists of a polypeptide related to the lactococcin 972 and multiple-membrane-spanning putative immunity protein. This model represents a small clade within the ABC transporters that regularly are found adjacent to these bacteriocin system gene pairs and are likely serve as export proteins.
ABC_MJ0796_Lo1CDE_FtsE	This family is comprised of MJ0796 ATP-binding cassette, macrolide-specific ABC-type efflux carrier (MacAB), and proteins involved in cell division (FtsE), and release of liporoteins from the cytoplasmic membrane (LolCDE). They are clustered together phylogenetically. MacAB is an exporter that confers resistance to macrolides, while the LolCDE system is not a transporter at all. An FtsE null mutants showed filamentous growth and appeared viable on high salt medium only, indicating a role for FtsE in cell division and/or salt transport. The LolCDE complex catalyses the release of lipoproteins from the cytoplasmic membrane prior to their targeting to the outer membrane.
DUF1093	Proteins of unknown function
DUF1430	This family represents the C-terminus (approximately 120 residues) of a number of hypothetical bacterial proteins of unknown function. These are possibly membrane proteins involved in immunity.
COG4652	Uncharacterized protein conserved in bacteria [Function unknown]
HTH	Helix-turn-helix domains; A large family of mostly alpha-helical protein domains with a characteristic fold; most members function as sequence-specific DNA binding domains, such as in transcription regulators. This superfamily also includes the winged helix-turn-helix domains.
HTH_MARR	helix_turn_helix multiple antibiotic resistance protein
oxido_YhdH	putative quinone oxidoreductase, YhdH/YhfP family; a superfamily in which some members are zinc-binding medium-chain alcohol dehydrogenases while others are quinone oxidoreductases with no bound zinc. This subfamily includes proteins studied crystallographically for insight into function: YhdH from Escherichia coli and YhfP from Bacillus subtilis. Members bind NADPH or NAD, but not zinc.
MDR_yhdh_yhfp	Yhdh and yhfp-like putative quinone oxidoreductases; QOR catalyzes the conversion of a quinone + NAD(P)H to a hydroquinone + NAD(P)+. Quinones are cyclic diones derived from aromatic compounds. Membrane bound QOR actin the respiratory chains of bacteria and mitochondria, while soluble QOR acts to protect from toxic quinones (e.g. DT-diaphorase). QOR reduces quinones through a semi-quinone intermediate via a NAD(P)H-dependent single electron transfer. QOR is a member of the medium chain dehydrogenase/reductase family, but lacks the zinc-binding sites of the prototypical alcohol dehydrogenases of this group. NAD(P)(H)-dependent oxidoreductases are the major enzymes in the interconversion of alcohols and aldehydes, or ketones.

Table 4: Prediction of Subcellular localization sites in hypothetical proteins by PSORTB

Sequence ID	Localization
gi|390516760	Cytoplasmic
gi|390516759	Unknown
gi|166409293	Cytoplasmic
gi|166409299	CytoplasmicMembrane
gi|166409303	Unknown
gi|166409302	CytoplasmicMembrane
gi|390516769	CytoplasmicMembrane
gi|166409301	Unknown
gi|166409300	CytoplasmicMembrane
gi|166409294	CytoplasmicMembrane

Table 5: Families found by PFAM database

Sequence ID	Pfam-A	Pfam-B	Domain
gi|166409294	ADH_zinc_N	-	ADH_N
gi|390516760	DUF885	-	YolD
gi|390516759	-	Pfam-B 7040	-
gi|166409293	MarR	Pfam-B_33	-
gi|390516769	DUF1430	Pfam-B_2627	
gi|166409300	DUF1430		
gi|166409301	Tmemb_9	Pfam-B_15664	
	DUF1312		
	DUF1093		
gi|166409302		Pfam-B_16450	ABC_tran
gi|166409303	Mob_Pre		
gi|166409299	Lactococcin_972		

Table 6: Descriptions of Pfam families of hypothetical proteins

Sequence ID	Description
gi|166409294	Alcohol dehydrogenase GroES-like domain
	Zinc-binding dehydrogenase family
gi|390516760	YolD-like protein
gi|166409293	MAR family; multiple antibiotic resistance
gi|166409301	TMEM9; widely expressed and localised to the late endosomes and lysosomes
gi|166409302	ABC transporters belong to the ATP-Binding Cassette superfamily

Table 7: SOSUI result of hypothetical proteins

Sequence ID	N-terminal	Transmembrane region	C-terminal	Type	Length
gi|166409299	42	FVSSCIASTILFGTLLGVTYKAE	64	PRIMARY	23
gi|166409301	41	TLIIVTIVLLIIILSLLLVRN	61	PRIMARY	21
gi|166409300	184	ILFFELVIDNDLLVPFIFLGVLY	206	SECONDARY	23
	242	KIIYWLVLTTIFFIANILIIHIA	264	PRIMARY	23
	279	VLLFVFACITTLLWLISYSLLLK	301	PRIMARY	23
	317	SIFMGTFTKCIMVLILSFLLAQN	339	SECONDARY	23
	575	IITIVTIIINVFILLIATVFEII	597	PRIMARY	23
	633	IMLAYTTHILFGSKVLLFIIMSI	655	PRIMARY	23

Table 8: Hypothetical proteins interacting with functionally important proteins

Sequence ID	Interacting proteins
gi|166409299	arsenate reductase
gi|166409302	elongation factor G
	30S ribosomal protein S7
	30S ribosomal protein S5
gi|166409301	ABC transporter ATP-binding protein
gi|166409300	cysteine protease precursor
gi|390516769	type I restriction-modification system endonuclease
gi|166409293	carboxy-terminal processing proteinase CtpA
gi|166409294	regulatory protein MarR

Table 9: Templates used by PS2 server for modeling

Sequence ID	Templates
gi|166409300	1rh1A
gi|166409294	1tt7A
gi|166409293	1z91A
gi|166409302	2it1A

Table 10: Residues involved in ligand binding sites predicted by QSITE finder

Sequence ID	Site volume	Residues
gi|166409293	209	SER 14, ARG 17, VAL 18, PHE 19, HIS 21, PHE 22, LEU 111, ASN 112, PHE 116, ALA 117, ASN 118, LEU 119.
gi|166409294	419	ILE 40, ASN 41 & 243, TYR 42, ALA 128, GLY 156, 159,160 & 162, 242, 313 & 317, VAL 161 & 244, THR 219, CYS 241, GLN 309 & 314, LEU 310 & 311, LYS 312, HIS 315, ARG 318, PHE 152, SER 153.
gi|166409300	169	GLY 56, LEU 57 & 58, LYS 60, ASN 61, ILE 64, TYR 90, ASN 148, VAL 149, PHE 152.
gi|166409302	241	GLU 157, THR 159, GLY 160, LEU 162, ASP 163, THR 164, GLY 167, LYS 168, ILE 171, THR 188, HIS 189, ASP 190, GLU 192, LEU 193.