

Re-annotation for hypothetical protein CA803_03125 of Methicillin-Resistant *Staphylococcus aureus* strain SO-1977 isolated from Sudan

Sofia B. Mohamed*, Mohamed M. Hassan, KA Abdalla Munir, Nusiba I. Abdalla, Talal A. Adlan, Aisha K. Babiker

¹Bioinformatics and Biostatistics Department, National University Research Institute, National University, Sudan; Sofia B. Mohamed - Email: Sofiabashir2002@gmail.com; * Corresponding author

Received December 12, 2018; Revised December 26, 2018; Accepted December 27, 2018; Published March 15, 2019

DOI:10.6026/97320630015160

Abstract:

This study aims to describe the global detection and functional inference of hypothetical protein CA803_03125 from *Staphylococcus aureus* SO-1977. Computational methods were utilized to study this protein based on sequence similarity and presence of known protein domains. The BLASTp result revealed a significant similarity between the hypothetical protein (CA803_03125) and ADP-ribose hydrolase protein from four *S. aureus* strains (MW2, MRSA252, COL, and N315). Evolutionary tree diagram revealed a close relationship between the hypothetical protein and proteins of MW2 and COL strains. The physicochemical characterization revealed that all proteins were found to be stable, soluble, hydrophilic and acidic in their nature. The Macro domain was found to exist within all proteins. Moreover, the proteins were of pronounced similarity in terms of primary, secondary and tertiary organization. The protein CA803_03125 (SO-1977) is already known and well characterized as ADP-ribose hydrolase; therefore, we would recommend that its NCBI data has to be updated to be submitted under the name of ADP-ribose hydrolase.

Keyword: Hypothetical Protein, MRSA SO-1977 strain, Bioinformatics tool, ADP-ribose hydrolase, Sudan

Background:

Methicillin-resistant *Staphylococcus aureus* (MRSA) is any strain of a bacterium, *Staphylococcus aureus* that has developed resistance to most of the available antibiotics. In the last decades, epidemiological studies showed sound increasing of endemic and epidemic spread while its control has become an important concern worldwide [1]. In Sudan, studies have also shown high MRSA incidence rate in hospitals [2]. During recent years, hundreds of bacterial genomes are available, while their annotation is of interest [3]. However, many of these protein functions are still unknown. For this reason, there is an increasing demand for the annotation of the functions of uncharacterized proteins, called "hypothetical

proteins" [4]. Hypothetical Protein (HP) is a protein that predicted to be expressed from an open reading frame, but for which there is no experimental evidence of translation [5]. About half of the proteins in most genomes are candidates for HPs. This group is of utmost importance to complete genomic and proteomic information. Detection of new HPs not only offers a presentation of new structures but also new functions [6]. Many protein domains have unknown functions; however, these domains participate in the metabolic pathways of organisms and can cause adverse effects. Several approaches have been developed by scientists with the aid of various computational tools to predict protein function. This has been achieved from information derived from sequence similarity,

phylogenetic analysis, conserved domains, motifs and 3D structure [7]. In this study, an extensive insilico analysis was carried out to explain the functional properties of the hypothetical protein CA803_03125 (accession number: OXL90457), of Methicillin-Resistant *S. aureus* strain SO-1977 using available protein structural and functional analysis tools.

Physicochemical parameters

Computations of various physical and chemical parameters for all proteins were predicted using the ProtParam server. ProtParam was used to determine the following parameters; molecular weight (M. Wt), isoelectric point (pI), amino acid composition, charge (positive or negative), atomic composition, extinction coefficient (EC), estimated half-life, instability index (II), aliphatic index (AI) and grand average of hydropathicity (GRAVY). The amino acids and atomic compositions are self-explanatory [11].



Figure 1: Multiple sequence alignment of the hypothetical proteins from MRSA SO-1977 (labelled blue) with ADP-ribose hydrolase from different *S. aureus* strains, red regions indicates similar residues. Black regions show nonsimilar.

Methodology:

The hypothetical protein targeted, MSA, and Phylogenetic tree construction

Hypothetical protein CA803_03125 was extracted from our project: whole genome sequencing of MRSA SO-1977 strain, with the GenBank accession number: NFZY00000000. Searches for similar protein sequences were carried out by using protein-Basic Local Alignment Search Tool (BLASTp) [8], with the default algorithm search parameters against the UniProtKB/Swiss-Prot database. Multiple sequence alignment (MSA) of the hypothetical protein from the SO-1977 strain with different strains was performed by Clustal Omega alignment program (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). Multiple sequence alignments (MSA) result was shaded by BoxShade server [9]. By the same token, the phylogenetic tree of all compared sequences was constructed using MEGA software version.6 [10].

Table 1: The Most significance of the highest similarity sequences (BLAST result)

Protein	Strain	Accession number	Identity %
ADP-ribose hydrolase	<i>S. aureus</i> subsp. <i>aureus</i> MW2	Q8NYB7.1	98%
ADP-ribose hydrolase	<i>S. aureus</i> subsp. <i>aureus</i> MRSA252	Q6GJZ1.1	98%
ADP-ribose hydrolase	<i>S. aureus</i> subsp. <i>aureus</i> COL	Q5HIW9.1	98%
ADP-ribose hydrolase	<i>S. aureus</i> subsp. <i>aureus</i> N315	P67344.1	98%

Table 2: Percentage of each amino acid composition

Amino Acid	HP (OXL90457)	Q8NYB7	Q6GJZ1	Q5HIW9	P67344
Ala (A)	8.3%	8.3%	8.3%	8.3%	8.3%
Arg (R)	6.0%	5.6%	6.0%	5.6%	5.6%
Asn (N)	6.4%	6.8%	6.8%	6.8%	6.8%
Asp (D)	7.9%	7.9%	7.9%	7.9%	7.9%
Cys (C)	2.6%	2.6%	2.6%	2.6%	2.6%
Gln (Q)	4.5%	4.5%	4.9%	4.5%	4.5%
Glu (E)	5.3%	5.3%	5.3%	5.3%	5.3%
Gly (G)	4.1%	4.1%	4.1%	4.1%	4.1%
His (H)	2.3%	2.6%	1.9%	2.6%	2.6%
Ile (I)	6.4%	6.4%	6.8%	6.4%	6.0%
Leu (L)	10.5%	10.5%	10.9%	10.5%	10.9%
Lys (K)	6.8%	6.4%	6.4%	6.4%	6.8%
Met (M)	2.3%	2.3%	1.9%	2.3%	2.3%
Phe (F)	3.4%	3.4%	3.4%	3.4%	3.0%
Pro (P)	3.0%	3.0%	3.0%	2.6%	3.0%
Ser (S)	4.5%	4.5%	4.5%	4.9%	4.5%
Thr (T)	4.5%	4.5%	4.5%	4.5%	4.5%
Trp (W)	0.8%	0.8%	0.8%	0.8%	0.8%
Tyr (Y)	3.0%	3.0%	3.0%	3.0%	3.0%
Val (V)	7.5%	7.5%	7.1%	7.5%	7.5%
Pyl (O)	0.0%	0.0%	0.0%	0.0%	0.0%
Sec (U)	0.0%	0.0%	0.0%	0.0%	0.0%

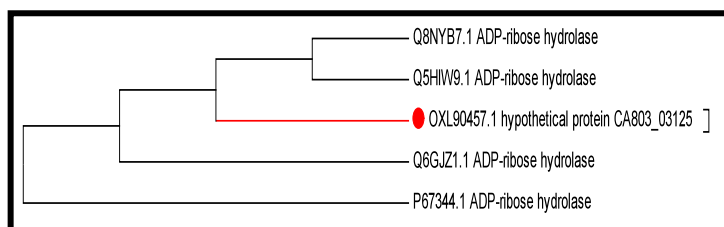


Figure 2: Phylogenetic tree of amino acid sequences of the target hypothetical protein with the ADP-ribose hydrolase of different strains of *S. aureus*.

Prediction of functional sites:

Screening for domain sites of the target (hypothetical) plus compared proteins sequences were predicted using the Pfam (protein families) database [12]. Thereafter, transmembrane domains were predicted by using the SOSUI server [13], which distinguishes between the membrane and soluble proteins from amino acid sequences, and predicts the transmembrane helices for the former.

Secondary structure prediction:

The secondary structures of proteins were estimated by SOPMA [14].

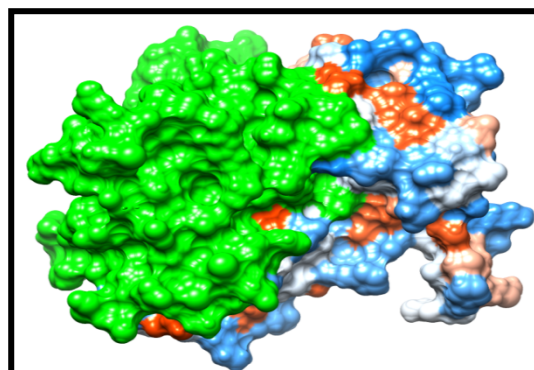


Figure 3: 3D structure of the protein model, the green colour shows the MACRO domain.

Homology Modelling and model validation:

The 3D models of the proteins were constructed using Swiss-Model server [15]. Molecular graphics and analyses were performed with the UCSF Chimera package [16].

Table 3: Atomic composition of HP (OXL90457) and similar proteins

Atomic composition	OXL90457	Q8NYB7	Q6GJZ1	Q5HIW9	P67344
Carbon C	1330	1328	1329	1326	1325
Hydrogen H	2133	2122	2132	2120	2125
Nitrogen N	375	374	374	374	375
Oxygen O	398	399	400	400	399
Sulfur S	13	13	12	13	13
Formula	C ₁₃₃₀ H ₂₁₃₃ N ₃₇₅ O ₃₉₈ S ₁₃	C ₁₃₂₈ H ₂₁₂₂ N ₃₇₄ O ₃₉₉ S ₁₃	C ₁₃₂₉ H ₂₁₃₂ N ₃₇₄ O ₄₀₀ S ₁₂	C ₁₃₂₆ H ₂₁₂₀ N ₃₇₄ O ₄₀₀ S ₁₃	C ₁₃₂₅ H ₂₁₂₅ N ₃₇₅ O ₃₉₉ S ₁₃
Total number of atoms	4249	4236	4247	4233	4237

Table 4: Various physicochemical properties of hypothetical (OXL90457.1) and ADP-ribose hydrolase proteins

Protein ID	Molecular weight	Positive charge	Negative charge	Extinction coefficient	GRAVY	Theoretical isoelectric point	Instability index	Aliphatic index	Estimated half life
OXL90457	30161.62	34	35	22920	-0.271	6.60	35.68	96.05	>10 hours
Q8NYB7	30128.50	32	35	22920	-0.265	6.18	35.30	96.05	>10 hours
Q6GJZ1	30134.53	33	35	22920	-0.263	6.23	37.50	97.89	>10 hours
Q5HIW9	30118.46	32	35	22920	-0.262	6.18	33.85	96.05	>10 hours
P67344	30109.50	33	35	22920	-0.293	6.39	34.61	96.05	>10 hours

Results and discussion:

The protein-BLAST search revealed that the HP CA803_03125 was similar to ADP-ribose hydrolase proteins belong to the other *S. aureus* spp (Table 1). Furthermore, a small number of variations was detected within an MSA result (Figure 1). The phylogenetic tree showed that protein sequences with accession numbers:

Q8NYB7.1 and Q5HIW9.1 were the closest strains (NCBI Taxonomy IDs: 196620 and 93062 respectively) to HP CA803_03125 (Figure 2). According to physicochemical parameters result, the HP CA803_03125 was found to share same physicochemical properties with the ADP-ribose hydrolases. All the proteins seem to be mildly acidic. The values of instability index for all strains were lower than

40 indicating that all the proteins are stable [17]. All strains showed higher aliphatic indices, which suggested that proteins are stable over a wide temperature range. The GRAVY value is negative, which indicates the hydrophilic and the soluble nature of the proteins [18]. Various parameters were arranged in (Tables 3 and 4). Pfam search resulted in identifying a MACRO domain between the HP CA803_03125 and ADP-ribose hydrolase proteins from other strains (Table 5). Macro domains are ancient, highly evolutionarily conserved domains that are widely distributed throughout all kingdoms of life. The 'macro fold' is roughly 25kDa in size and is composed of a mixed α - β fold with similarity to the P loop-containing nucleotide triphosphate hydrolases. They function as binding modules for metabolites of NAD⁺, including poly (ADP-ribose) (PAR), which is synthesized by PAR polymerases (PARPs) (Figure 3) [19]. By the same token, all proteins were classified as soluble proteins and the hydrophobicity range was found from -0.254753 to -0.292857. The secondary structure information may give insights into the higher order structure and functional annotation of the protein [20]. The secondary structure of HP CA803_03125 and ADP-ribose hydrolase proteins was found to be the same. The α -helices were found dominant, followed by random coils, extended strands (β -sheets) and β -turns in all proteins (Table 6). Furthermore, homology modeling for all proteins was built based on a single model template (PDB ID 5kiv) which was the most similar to all of them with varying similarity ranging from 97.74 - 100%. In addition, the GMQE (Global Model Quality Estimation) scores were found to be 0.97 in all proteins, which are very close to 1, as higher number indicates higher reliability (Table 7).

Table 5: Domains identified and their positions for all protein sequences

Protein	Domain	Position
OXL90457	MACRO	103-277
Q8NYB7	MACRO	103-277
Q6GJZ1	MACRO	103-277
Q5HIW9	MACRO	103-277
P67344	MACRO	103-277

Table 6: The percentages of secondary structure prediction

Protein ID	Alpha helix %	Extended strand %	Random coil %	Beta turn %
OXL90457	47.91	15.22	28.52	8.37
Q8NYB7	47.37	15.04	27.82	9.77
Q6GJZ1	47.37	15.04	28.95	8.65
Q5HIW9	47.74	15.04	27.44	9.77
P67344	46.62	14.29	29.70	9.40

Table 7: Homology modelling and models validation

Protein ID	PDB ID	Identity %	GMQE score
OXL90457	5kiv	98.5	0.97
Q8NYB7	5kiv	98.5	0.97

Q6GJZ1	5kiv	98.12	0.97
Q5HIW9	5kiv	97.74	0.97
P67344	5kiv	100	0.97

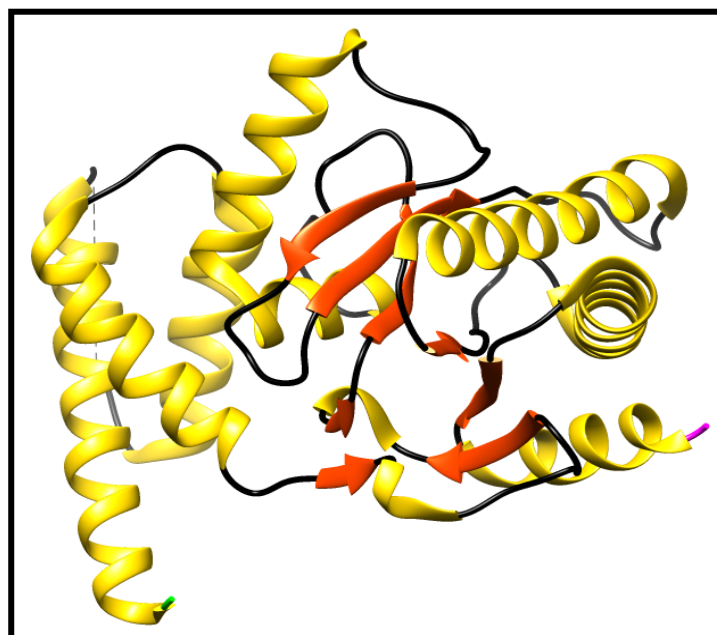


Figure 4: 3D structure of the template protein (PDB ID: 5kiv). Golden color shows alpha helices, red color shows extended strand, black color shows Random coil and Beta turns, green color represents the N terminal and pink color represents the C terminal.

Conclusion:

In this study, the hypothetical protein (OXL90457) from SO-1977 was predicted and identified to be ADP-ribose hydrolase protein. The bacterial MACRO domains are known to influence processes that are crucial for the survival and virulence of bacteria in the host environment. Therefore, MACRO domain would be a subject for further investigations in order to understand the host-pathogen interactions and to explore novel therapeutic routes. The hypothetical protein (OXL90457) should be updated in the NCBI database to be included under the name of ADP-ribose hydrolase.

Acknowledgment:

The authors are grateful to Prof Qurashi M. Ali, the National University (Sudan) president for his encouragement and support.

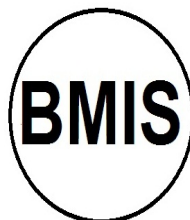
References:

- [1] Voyich J *et al.* J Infect Dis. 2006 **194**(12):1761. [PMID: 17109350].
- [2] Kheder SI *et al.* Pharmacology & Pharmacy. 2012 **12** **3**(01):103.
- [3] Roberts RJ. *PLoS Biol.* 2004 **2**:E42.
- [4] Galperin MY & Koonin EV. *Nucleic Acids Res.* 2004 **32**:5452.
- [5] Lubec G *et al.* Prog Neurobiol. 2005 **77**:90.
- [6] Galperin MY & Koonin EV. *Nucleic Acids Res.* 2004 **32**: 5452.
- [7] P. Bharat Siva Varmaa *et al.* Journal of Infection and Public Health 2015 **8**: 526.
- [8] <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.
- [9] http://www.ch.embnet.org/software/BOX_form.html
- [10] Tamura K *et al.* Molecular biology and evolution. 2013 **30**(12):2725.
- [11] <http://web.expasy.org/protparam/>
- [12] <http://pfam.xfam.org/>
- [13] http://harrier.nagahama-i.bio.ac.jp/sosui/sosui_submit.html
- [14] https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html
- [15] <https://swissmodel.expasy.org/>
- [16] <https://www.cgl.ucsf.edu/chimera/>
- [17] Mohan R & Venugopal S. *Bioinformation.* 2012 **8**: 722.
- [18] Sharma AR *et al.* *Biomed Res Int.* **2014** 1.
- [19] Han W *et al.* *Mutation Research/Reviews in Mutation Research.* 2011 **727**(3):86.
- [20] Mills CL *et al.* *Computational and Structural Biotechnology Journal.* 2015 **1**(13): 182.

Edited by P Kanguane

Citation: Mohamed *et al.* *Bioinformation* 15(3): 160-164 (2019)

License statement: This is an Open Access article which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. This is distributed under the terms of the Creative Commons Attribution License



Biomedical Informatics Society

Agro Informatics Society



Journal