Saudi Journal of Biological Sciences 30 (2023) 103812

HOSTED BY

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

Reconnoitering the sequence and structural analysis of *Staphylococcus aureus* "A" protein

Parthasarathy Ranganathan [a], Akila Varatharajan [b,*], Khaloud Mohammed Alarjani [c], Dunia A Farraj [c], Viji Rajendran [d]

[a] Faculty of Physiotherapy, Meenakshi Academy of Higher Education and Research, West K.K. Nagar, Chennai – 600078, Tamil Nadu, India
[b] Central Research Laboratory, Department of Research, Meenakshi Academy of Higher Education and Research, West K.K. Nagar, Chennai – 600078, Tamil Nadu, India
[c] Department of Botany and Microbiology, College of Science, King Saud University, Riyadh 11451, Saudi Arabia
[d] Department of Marine Science and Convergence Technology, Hanyang University ERICA Campus, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan-si, Gyeonggido 426-791, South Korea

ARTICLE INFO

Article history:
Received 27 July 2023
Revised 24 August 2023
Accepted 9 September 2023
Available online 14 September 2023

Keywords:
Staphylococcus aureus
"A" protein
Sequence analysis
Bioinformatics tools
Computational modeling
Therapeutic strategies

ABSTRACT

Background: The *Staphylococcus aureus* "A" protein plays an essential role in the pathogenicity and virulence of this bacterial species. To gain deeper insights into the protein's characteristics, we conducted an in-depth analysis of its sequence and structure.
Objective: This study aimed to unravel the underlying genetic and structural components that contribute to the protein's functional properties.
Results: Utilizing various bioinformatics tools and techniques, we first examined the protein's primary sequence, identifying key amino acid residues and potential functional domains. Additionally, we employed computational modeling and simulation approaches to determine the tertiary structure of the "A" protein. Through this comprehensive analysis, we discovered novel features and interactions within the protein's structure, shedding light on its potential mechanisms of action. Furthermore, we investigated the protein's evolutionary conservation and compared it with related proteins from other bacterial species.
Conclusions: Overall, our findings provide valuable insights into the sequence and structure of the Staphylococcus aureus "A" protein, which may have implications for understanding its role in pathogenicity and guiding the development of novel therapeutic strategies.

1. Introduction

Staphylococci are typically commensal organisms that are commonly found on the skin and mucosa of humans and animals. Their diameter ranges from 0.5 to 1.5 μm, and they are cocci that are gram-positive, occurring predominantly in groups but also as single cells or in pairs. They are non-motile, and non-sporing and only a few strains have capsules. Their colonies measure around 1 mm in diameter (Milani et al., 2023). Certain strains, particularly those resistant to oxacillin, pose a global threat to both humans and animals. These bacteria are spherical and belong to the bacilli group. Frequently present in the upper respiratory tract and on the skin, they are commonly included in the body's microbiota. *Staphylococci* are often positive for catalase and nitrate reduction, and they are facultative anaerobes capable of growing without oxygen. When grown on mannitol salt agar, a selective medium, *Staphylococcus aureus* produces yellow colonies due to mannitol fermentation and subsequent pH drop. *Staphylococcus aureus* produces at least four types of hemolytic activities α, β, δ, and γ (Sangeetha, 2023).

Staphylococci are bacteria with pyogenic properties, frequently responsible for superlative infections. Minor injuries or immune suppression can increase the likelihood of infection development. They adhere to damaged skin, mucosa, or tissue surfaces and evade the host's defense mechanisms. *Staphylococci* colonize these sites

* Corresponding author at: Central Research Laboratory, Department of Research, Meenakshi Academy of Higher Education and Research, West K.K. Nagar, Chennai 600078, Tamil Nadu, India.
E-mail addresses: parthasarathy@maherfpt.ac.in (P. Ranganathan), akila.researchassociate@maher.ac.in (A. Varatharajan), kalarjani@ksu.edu.sa (K. Mohammed Alarjani), dfarraj@ksu.edu.sa (D.A Farraj), viji86@hanyang.ac.kr (V. Rajendran).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.sjbs.2023.103812
1319-562X/

and cause tissue damage by multiplying and releasing toxins that induce inflammation. These pathogens are clinically significant, causing a wide range of illnesses in both humans and animals worldwide (Chakroborty et al., 2011).

Staphylococcus aureus is responsible for various infections they can cause a range of conditions, from superficial skin infections to severe diseases like respiratory system, intestinal tract, endocarditis, sepsis, and infections affecting the soft tissues, urinary tract, and bloodstream. Additionally, it is a significant foodborne pathogen primarily because it produces enterotoxins that can lead to severe intoxications (Wu et al., 2011, Lu et al., 2014).

In *Staphylococcus aureus*, the peptidoglycan composition consists of N-acetyl muramyl (L-Ala-D-iGln-L-Lys (Gly5)-D-Ala)-(b-4), N-acetyl glucosamine (Mur NAc-Glc NAc), and it generates the glycan strands. By forming amide bonds between penta glycine cross bridges (Gly5) and D-Ala residues of adjacent wall peptides, the creation of a three-dimensional murein network is achieved (De Pedro and Cava, 2015).

The carbohydrate and teichoic acid present in the cell wall of *Staphylococcus* is attached to glycan strands. Conversely, exterior proteins, such as Protein A, are covalently bound and associated with pentaglycine cross bonds. Protein A has its N-terminal portion, which includes five immunoglobulins (Ig) binding domains, situated on the external of *Staphylococcus* (Rajagopal and Walker, 2017). The C-terminal threonine residue of Protein A is amide-linked to the murein sacculus. During synthesis, Protein A is initially produced in the cytoplasm as a pioneer with an N-terminal signal peptide for secretion and a C-terminal cataloging indicator for incorporation into the cell envelope.

The sorting signal of Protein A is cut by sortase A, specifically among the threonine and glycine of its LPXTG motif. The resulting cleaved polypeptide forms a thioester-linked intermediate at the lively site thiol of sortase. A nucleophilic outbreak by a lipid molecule (c55-MUrNAc-(L-Ala-D-iGln-L-Lys(Gly5)-D-Ala-D-Ala)-(b-4)-Glc NAc) at these bond determinations the acyl enzymes. The external protein, now associated with the peptidoglycan initiator, is subsequently fused into the cell wall envelope through transpeptidation and transglycosylation reactions (Mazmanian et al., 1999 Ton-That et al., 1999). It plays a role in scavenging heme (porphyrin) by binding to both Fe(2 +) and Fe(3 +) heme, with the majority being Fe(2 +) heme. This protein utilizes as a high-affinity heme-binding protein and likely participates in transferring heme–iron from cell wall-anchored Isd protein receptors to the probable permease IsdF.

The research primarily aimed to obtain the primary virulent A protein sequence of S. aureus from the SWISS-PROT database. Subsequently, a blast analysis was conducted on the retrieved sequence to identify similar sequences. Multiple alignments were then performed on these homologous sequences to identify functional domains and motifs. An NJ tree was constructed by conducting a phylogenetic analysis of the related sequences. Moreover, the query protein's secondary structure was forecasted. Fingerprints and blocks were created through ungapped alignment. Comparative modeling was carried out using homologous sequences in the SCOP database to predict the folded tertiary structure of the query protein.

2. Materials and methodology

In the current study, computational techniques were employed to analyze amino acid sequences in a biologically significant manner. Insilico tools and resources were utilized to effectively characterize a protein sequence, identify its biochemical functions, and predict its folded tertiary structure. Since even slight changes in structure can greatly impact a protein's function, the utilization of software tools with precision was crucial in these investigations.

2.1. Basic local alignment search tool (BLAST)

BLAST (Basic Local Alignment Search Tool), as defined by Altschul et al. (1990), has emerged as the standard for search and alignment tools. Despite its name, BLAST is far from basic it is an advanced algorithm that has gained popularity because of its availability, speed, and accuracy, a BLAST search primarily involves querying one or more databases, typically hosted by NCBI (https://www.ncbi.nlm.nih.gov/), with the sequence of interest to identify homologous sequences. (McGinnis and Madden, 2004). BLAST is valuable for inferring function and also not only aids in identifying gene family members but also provides insights into the evolutionary relationships between sequences.

2.2. E-value

To restrict the number of hits to the most significant ones, the E-value, also known as the expected value, can be modified. The E-value is influenced by the length of the query sequence and the size of the database. For E-value < 10e − 100, the sequences are considered identical, resulting in a long alignment between both the query sequence and the hit sequence in their entirety. When the E-value falls between 10e − 50 and 10e − 100, the sequences are virtually equal, with a long stretch of the query protein matching the database. For E-value between 10e − 10 and 10e − 50, the sequences are closely related, potentially indicating a domain match or similarity. When the E-value is between 1 and 10e − 6, it becomes a gray area where the sequences could be true homologs. However, for E-value > 1, the proteins are unlikely to be related. Hits with E-value > 10 are generally considered to be insignificant unless the query sequence is very short.

2.3. T blast n analysis

This method involved translating a protein sequence and conducting comparisons to identify matches with other translated sequences. Each frame was systematically analyzed for similarity by introducing gaps. In the case of nucleotides, a list of word lengths eleven was generated and compared against the query sequence. This approach was described by Altschul et al. (1997).

2.4. Multiple sequence alignment

The purpose of the multiple sequence alignment was to generate a concise and informative summary of sequence data, revealing the relationships between sequences and a specific gene family. The formulation of the Clustal W multiple alignment method can be attributed to Thomson et al., 1994. For this study, the pairwise aligned sequences obtained from the previous blast analysis were carefully chosen to assess the levels of homology and relatedness. By utilizing the dynamic programming method, all combinations of the eleven sequences were compared to obtain similarity scores. The sequences showing similarities were subjected to local alignment using dynamic programming, and the resulting consensus sequences from each comparison were aligned with one another. Finally, all sequences were combined using the consensus sequence as a guide. The most similar sequence obtained from this process, along with the conserved regions, was presented along with comparative scores.

2.5. Trees constructed using the neighbor-joining method

The purpose of this analysis was to investigate the evolutionary relationships of the sequence, as outlined by Feng and Doolittle, 1996. To accomplish this, a distance matrix was constructed, representing the genetic distances between sequences, for each pair-

Table 1
"A" protein sequence (A6QG34).

MRIIKYLTIL VISVVILTSC QSSSSQESTK SGEFRIVPTT VALTMTLDKL DLPIVGKPTS 60 YKTLPNRYKD VPEIGQPMEP NVEAVKKLKP THVLSVSTIK DEMQPFYKQL NMKGYFYDFD 120 SLKGMQKSIT QLGDQFNRKA QAKELNDHLN SVKQKsIENKA AKQKKHPKVL ILMGVPGSYL 180 VATDKSYIGD LVKIAGGENV IKVKDRQYIS SNTENLLNIN PDIILRLPHG MPEEVKKMFQ 240 KEFKQNDIWK HFKAVKNNHV YDLEEVPFGI TANVDADKAM TQLYDLFYKD KK 292
Total- 292 Amino acid residues

Table 2
Translated "A" Protein Sequences.

TTGAGAATCA	TAAAGTATTT	AACCATTTTA	GTGATAAGCG	TCGTTATCTT	AACCAGCTGT	60
CAATCTTCCA	GTTCTCAACA	ATCAACTAAA	TCCGGCGAAT	TCAGAATCGT	ACCAACAACT	120
GTTGCATTGA	CAATGACATT	GGACAAATTG	GATTTACCAA	TTGTCGGCAA	ACCCACGTCA	180
TATAAGACAT	TGCCTAATCG	TTATAAAGAT	TACCGGAAAT	TGGTCAACCA	ATGGAGCCGA	240
ATGTTGAAGC	TGTTAAAAAG	TTAAAACCAA	CACATGTTTT	GAGTGTGTCA	ACGATTAAAG	300
ATGAAATGCA	ACCATTTTAC	AAACAATTAA	ATATGAAAGG	CTACTTTTAT	GATTTTGATA	360
GTTTAAAAGG	GATGCAAAAG	TCGATTACAC	AATTAGGTGA	TCAATTTAAT	CGTAAAGCAC	420
AAGCAAAAGA	ATTAAATGAC	CATTTAAATT	CTGTAAAGCA	AAAAATTGAA	AATAAAGCAG	480
CTAAACAAAA	GAAACATCCC	AAAGTATTAA	TATTAATGGG	TGTACCTGGT	AGCTATTTAG	540
TAGCAACTGA	TAAATCATAT	ATTGGTGATT	TAGTTAAAAT	AGCAGGTGGA	GAAAATGTTA	600
TTAAAGTGAA	AGATCGTCAA	TATATTTCGT	CTAATACTGA	AAATTTATTG	AATATCAATC	660
CAGATATTAT	TTTAAGATTA	CCACACGGAA	TGCCTGAAGA	AGTTAAGAAA	ATGTTTCAAA	720
AAGAATTTAA	ACAAAATGAT	ATTTGGAAAC	ATTTTAAAGC	TGTGAAAAAT	AATCATGTTT	780
ATGACTTAGA	GGAAGTGCCA	TTCGGTATTA	CAGCAAATGT	TGATGCTGAT	AAGGCAATGA	840
CTCAATTGTA TGATTTATTT TATAAGGATA AAAAA 875						

Table 3
Theoretical estimation pI: 9.36 Composition of Amino acids.

Ala (A)	12	4.1%	Lys (K) 38	13.0%
Arg (R)	6	2.1%	Met (M) 10	3.4%
Asn (N)	16	5.5%	Phe (F) 10	3.4%
Asp (D)	18	6.2%	Pro (P) 15	5.1%
Cys (C)	1	0.3%	Ser (S) 18	6.2%
Gln (Q)	15	5.1%	Thr (T) 16	5.5%
Glu (E)	15	5.1%	Trp (W) 1	0.3%
Gly (G) 13	4.5%	Tyr (Y)	12	4.1%
His (H) 6	2.1%	Val (V)	23	7.9%
Ile (I) 21	7.2%	Pyl (O)	0	0.0%
Leu (L) 26	8.9%	Sec (U)	0	0.0%
(B)	0	0.0%		
(Z)	0	0.0%		
(X)	0	0.0%		

The total number of negatively charged residues (Asp + Glu) is 33.
The total number of positively charged residues (Arg + Lys)
44.
The N-terminal of the sequence considered is M (Met).
Instability index:
The instability index (II) is computed to be 35.63.
This classifies the protein as stable.
Aliphatic index
89.73.
Molecular weight
33270.71 Daltons.

wise comparison of the proteins. The sequences that were to be aligned acted as the terminal nodes and leaves of the tree. By calculating the distances between these sequences (nodes), a phylogenetic tree was constructed by connecting the most similar sequences.

2.6. Analyzing predictions of pattern or domain

Functional domains and regular expression patterns were identified by detecting conserved regions in the multiple aligned sequences. In the present study, the prosite tool was utilized to discover patterns within the aligned sequences. Initially, the aligned sequences were examined to locate conserved regions. Subsequently, sequence information within these conserved regions was recorded. The recorded information was searched in the database to identify true positive and false positive hits. The false positive matches were re-evaluated, resulting in the creation of optimal patterns for the query sequence.

The given set of multiple aligned sequences was used to create an alignment encoded as hidden Markov models (Sonnhammer et al., 1998). Each position within the alignment was designated as a match, delete, or insertion state, which facilitated the search for sequence conservation within the alignment. The hidden Markov models for the query protein were constructed by searching the aligned sequences in the Pfam database.

2.7. Analysis of fingerprint

The diagnostic patterns construct indicating membership in the protein family followed by according to Attwood et.al, 1998. We examined all the conserved regions within the sequence and transformed the sequence information into matrices that recorded the frequency of residues observed at each position of the motifs. These created fingerprints were then employed to detect distant relatives of the protein within the prints database.

2.8. Block search

In this research, the protein sequence was used to search the block database to identify blocks and discover more members of the protein family. The conserved segment of amino acids, which provides a distinct function or structure to the queried protein, was aligned without any gaps introduced. The resulting blocks were then carefully matched in the correct order to validate the protein's family identity.

2.9. Predicting secondary protein structure

The neural network-based n predicts algorithm was exploited to forecast the secondary structure of the query protein. This prediction relies on the spatial organization of amino acids in the primary structure, which is influenced by hydrophobic interactions, hydrogen bonding involving C = O and N–H groups in the backbone, as well as van der Waals interactions. These interactions give rise to two distinct secondary structures that exhibit no distortion.

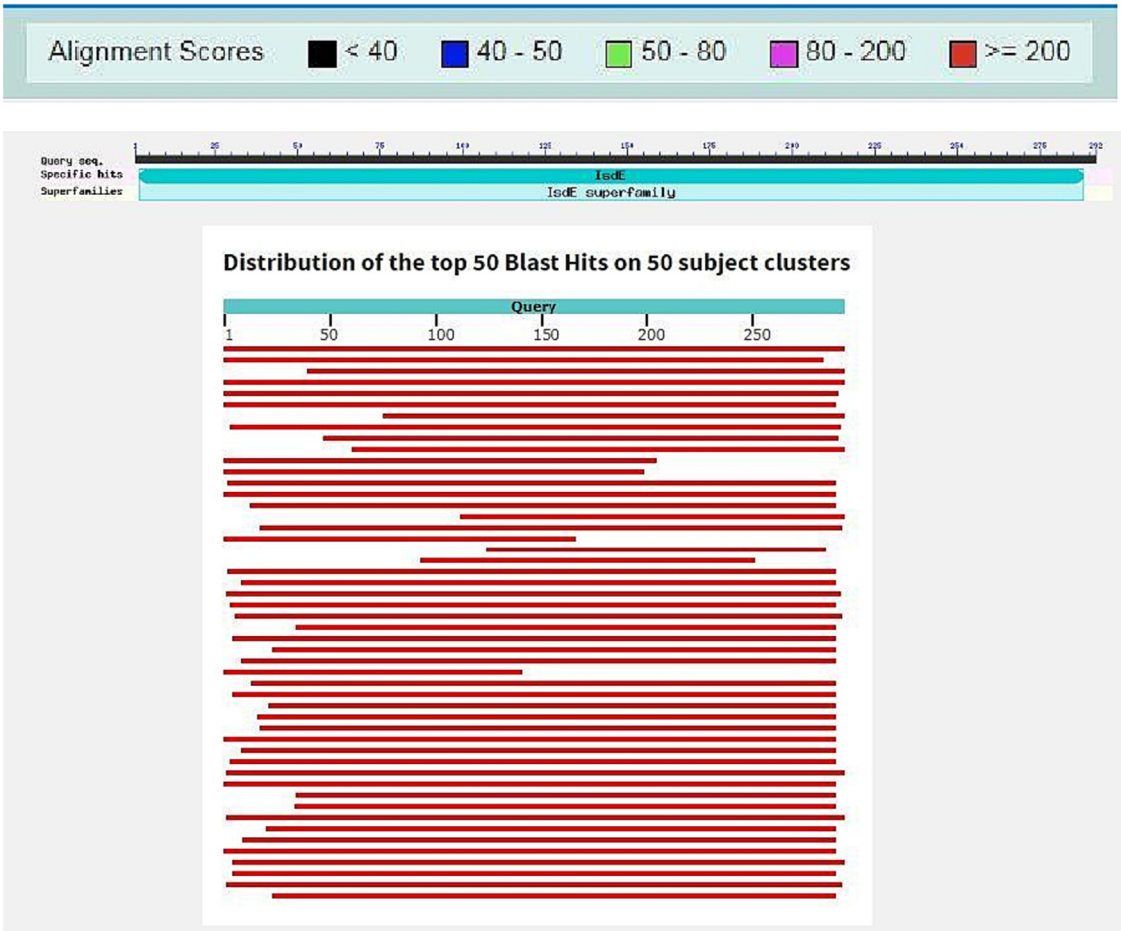


Fig. 1. BLAST query results in Program: blast p Database: swiss Query: (292 letters).

2.10. Transmembrane helices

The TM pred algorithm (Krogh et al., 2001) was employed to predict the transmembrane regions and orientation of the query protein. Initially, a hydrophobic analysis was performed by calculating the proportion of nonpolar amino acids in the protein. The neural network utilizes a dynamic programming algorithm to predict the helical orientation (topology) of the protein, taking into account the percentage of hydrophobic amino acids present.

2.11. Process of homology modeling

To predict the folded structure of the query protein, comparative modeling (Moreira, 2003) was utilized. Sequences from the SCOP database that exhibited>30% similarity and possessed identical backbones were obtained. The template structure, which could be a single or multiple structure within the protein family, was selected to model the query protein. In the prediction process, the first step involved placing the side chains of the query sequence onto the template's backbone. The structurally conserved regions, such as core secondary structure elements, were then modeled, followed by the variable regions (connecting loops) in the protein. Finally, the side chains were modeled, resulting in the prediction of the folded 3D structure for the query protein.

3. Results

The protein sequence of Staphylococcus aureus "A" virulent protein, specifically the Heme uptake affinity protein (A6QG34), was obtained from the Swiss-Prot database, which is an annotated data bank. The details of this protein sequence can be found in Table 1.

The query protein comprises 292 amino acid residues, and the translated sequence of this query protein is provided in Table 2.

3.1. Analysis and characterization of the query protein

The amino acid compositions of the query protein were determined by utilizing the Expasy Protparam tool, and the results are displayed in Table 3. The presence of all essential amino acids in the protein indicates its stability. By calculating the sum of the masses of the amino acid residues, the molecular weight of the protein was determined to be 33270.71 Daltons.

3.2. Blast search results

For *Blast p* analysis of the query protein, color-coded alignments of similar sequences were generated based on their expectation (E) values (Fig. 1). For *T Blast N* results, the translated query protein was analyzed, and the color-coded alignments along with 111 significant matches were obtained from multiple non-redundant databases (Fig. 2).

3.3. Analysis of motif

The conserved regions identified in the query protein through multiple alignments were subsequently searched in different protein secondary databases, such as Pfam, Prosite, and InterPro, to

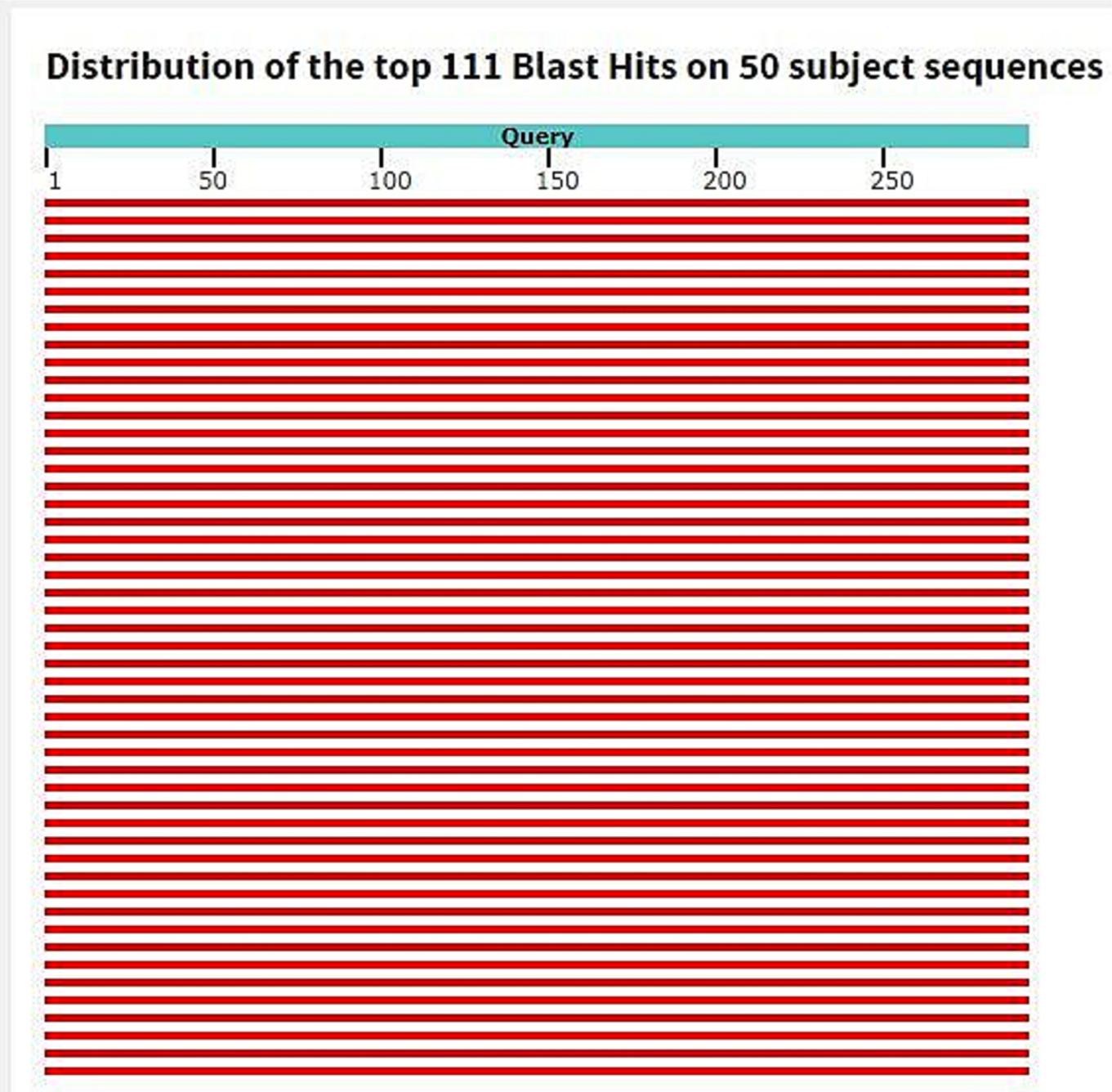


Fig. 2. T BLAST N Query = (292 letters).

establish profiles, patterns, and blocks (Table 4, Table 5, Table 6, and Table 7).

3.4. Analysis of fingerprint

A seed alignment was created in the PRINTS database for the query protein, encompassing a diagnostic signature found in 11 sequences. The fingerprint performance in terms of diagnosis was not particularly high, mainly due to the presence of only two motifs. Motif 1 consists of a highly conserved hexapeptide region, while motif 2 extends over a hydrophobic C-terminal anchor.

3.5. Block search

Weighting schemes were employed to characterize the constituent motifs and align them without introducing any gaps. The previously established conserved set of residues for the query protein was once again searched to identify the blocks with the highest scores. Through a search across the blocks database for the A-

Table 4
Identical sequences to "A" protein (A6QG34).

WP_252597239.1:1–292	
MRVIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
WP_060585232.1:1–292	
MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIVLRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
WP_187787496.1:1–292	
MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLVLMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
HCC5652601.1:1–292	
MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVVKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
WP_224799133.1:1–292	
MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYVGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
WP_160199169.1:1–292	
MRIIKYLTVLVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
WP_130111794.1:1–292	
MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
CAC6283612.1:1–292	
MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
WP_001220199.1:1–292	
MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240
WP_064127868.1:1–292	
MRIIKYLTILVVSVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS	60
YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD	120
SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL	180
VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292	240

protein family, which comprises 11 sequences, two blocks were discovered.

3.6. Secondary structure prediction

Software tools are employed to analyze protein sequences, accurately predict their folding patterns, and interpret their biochemical functions at the molecular level.

3.7. Transmembrane helices

Considering that the query protein is a surface protein, computational techniques (specifically TMpred) were utilized to predict the presence of transmembrane helices. The identification of non-polar amino acid residues and transmembrane helices serves as a confirmation of the query protein's surface protein nature.

3.8. Homology modelling

The query protein was initially searched in different structural protein databases, such as PDB, to find pre-existing folded tertiary structures. However, no matches were found in these structural repositories. Therefore, comparative modeling was employed to predict the protein's structure. Sequences related to the query protein, such as a protein fragment from SCOP and human apilo proteins, were used as templates for the homology modeling process. The resulting folded structure, as predicted earlier by the TMpred algorithms, displayed predominantly helical conformations in both models (Fig. 4).

4. Discussion

In this study, a specific "A" protein from *Staphylococcus aureus* isolate was obtained from the Swiss-Prot database (A6QG34),

Table 5
CLUSTAL O (1.2.4) Multiple Sequence Alignments.

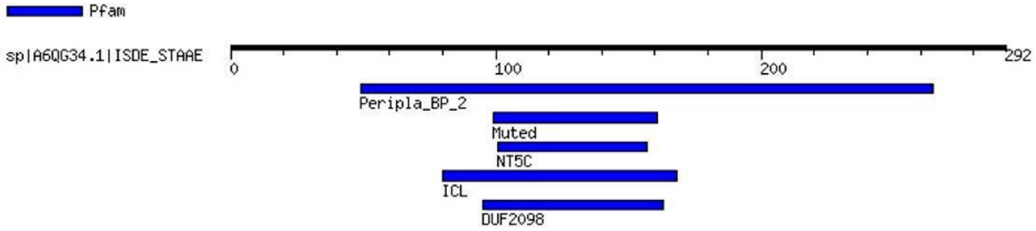
```
WP_252597239.1:1–292 MRVIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_060585232.1:1–292 MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_187787496.1:1–292 MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
HCC5652601.1:1–292 MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_224799133.1:1–292 MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_160199169.1:1–292 MRIIKYLTVLVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_130111794.1:1–292 MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
CAC6283612.1:1–292 MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_001220199.1:1–292 MRIIKYLTILVISVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_064127868.1:1–292 MRIIKYLTILVVSVVILTSCQSSSSQESTKSGEFRIVPTTVALTMTLDKLDLPIVGKPTS 60
WP_252597239.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_060585232.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_187787496.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
HCC5652601.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_224799133.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_160199169.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_130111794.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
CAC6283612.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_001220199.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_064127868.1:1–292 YKTLPNRYKDVPEIGQPMEPNVEAVKKLKPTHVLSVSTIKDEMQPFYKQLNMKGYFYDFD 120
WP_252597239.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_060585232.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_187787496.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLVLMGVPGSYL 180
HCC5652601.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_224799133.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_160199169.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_130111794.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
CAC6283612.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_001220199.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_064127868.1:1–292 SLKGMQKSITQLGDQFNRKAQAKELNDHLNSVKQKIENKAAKQKKHPKVLILMGVPGSYL 180
WP_252597239.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
WP_060585232.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIVLRLPHGMPEEVKKMFQ 240
WP_187787496.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
HCC5652601.1:1–292 VATDKSYIGDLVKIAGGENVVKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
WP_224799133.1:1–292 VATDKSYVGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
WP_160199169.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
WP_130111794.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
CAC6283612.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
WP_001220199.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
WP_064127868.1:1–292 VATDKSYIGDLVKIAGGENVIKVKDRQYISSNTENLLNINPDIILRLPHGMPEEVKKMFQ 240
WP_252597239.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
WP_060585232.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
WP_187787496.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
HCC5652601.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
WP_224799133.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
WP_160199169.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
WP_130111794.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
CAC6283612.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
WP_001220199.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
WP_064127868.1:1–292 KEFKQNDIWKHFKAVKNNHVYDLEEVPFGITANVDADKAMTQLYDLFYKDKK 292
```

which is an annotated primary protein data bank (Bairoch and Apweiller, 1998). Whisstock and Lesk, (2003) reveals the protein's sequence might provide insights into its structural characteristics. Certain regions of the sequence might be responsible for specific functions or interactions (Table 1). Exploring the sequence could involve identifying domains, motifs, or other functional elements. The retrieved protein was then analyzed for its sequence and structure using different bioinformatics tools. To identify homologous sequences, the primary sequence was initially analyzed using NCBI tools, specifically blast P and T blast N. Table 2 translated 'A' Protein Sequences'' encapsulates the conversion of genetic information from nucleic acids to proteins. This process is essential for understanding biological functions, evolutionary relationships, and various applications in science and medicine (Goldman et al., 2023). This analysis yielded 159 blast hits with E values ranging between e across multiple non-redundant databases.

Genomic analysis conducted on the invasive *Staphylococcus aureus* indicated its association with the emm gene pool (Ferretti et al., 2001). During the translated blast analysis of the query protein, it was found to have similarities with other emm gene types. Further examination of the homologous sequence obtained from the pairwise analysis involved using the Clustal W multiple alignment method (Thomson et al., 1994). This analysis resulted in the identification of 10 groups and three motifs. Muhammad et al., (2023) reveals that encompasses the theoretical determination of the isoelectric point (pI) of a protein and the description of the amino acid composition in that protein (Table 3). These concepts are central to understanding the properties and behavior of proteins in various biological contexts.

Signal peptide regions not only indicate homology but also provide insights into the function of a protein. Table 4 discussing identical sequences to the ''A'' protein (A6QG34) involves exploring the conservation, function, and evolutionary significance of proteins with the same amino acid sequence (Ali et al., 2023). It can provide insights into fundamental biological processes and contribute to various fields of research, including molecular biology, evolutionary biology, and biotechnology. When it comes to the query protein, a motif responsible for signal transfer was identified at the beginning of the transmembrane domain. To predict the domains of the protein, the Pfam collection of hidden Markov models

Table 6
Motif Scan For A – Protein In Secondary Databases.



Pfam (5 motifs)

Pfam	Position(Independent E-value)		Description
Peripla_BP_2	49..265(1.5e-39)	Detail	PF01497, Periplasmic binding protein
Muted	99..161(0.0048)	Detail	PF14942, Organelle biogenesis, Muted-like protein
NT5C	101..157(0.026)	Detail	PF06941, 5' nucleotidase, deoxy (Pyrimidine), cytosolic type C protein (NT5C)
ICL	80..168(0.031)	Detail	PF00463, Isocitrate lyase family
DUF2098	95..163(0.065)	Detail	PF09871, Uncharacterized protein conserved in archaea (DUF2098)

Table 7
Characterization of various domains of the query protein in pfam database.

Domain	Start	End
Peripla_BP_2	49	265
Muted	99	161
NT5C	101	157
ICL	80	168
DUF2098	95	163

(HMM) was employed. This analysis grouped strongly matched sequences into common families (Krogh et al., 2001). The HMM analysis indicated that the query protein belongs to the A protein family and is related to other surface protein families within the Pfam database.

The analysis of fingerprints in the database confirmed the presence of the hydrophobic C terminal anchor motif and seed alignment for Protein A. Additionally, the ungapped alignment was utilized to create blocks, allowing for the inclusion of more members into the family of the query protein. This further confirmed the conserved regions within the motifs. Danoy et al., (2023) summarizes the CLUSTAL O (1.2.4) is a version of the CLUSTAL software used for performing multiple sequence alignments (Table 5). This tool is important in various biological analyses, including understanding evolutionary relationships, predicting protein structures, and annotating functional motifs. However, like any alignment tool, it has its limitations and considerations for optimal usage.

The similarity search criteria were then employed to investigate the structure and function of the query protein. The folding of the protein is influenced by steric hindrance and hydrogen bonding patterns, which affect its torsional angles (Creighton, 1993). The protein was found to contain amino acids with higher propensities for helices, such as alanine, glutamic acid, and leucine. On the other hand, proline, which exhibits an index of 35.63%, was also present in the protein. Stable secondary structures were predicted for the protein. Table 6 performing a motif scan for the "A" protein in secondary databases involves searching for conserved sequence patterns that could correspond to functional elements (Li et al., 2023). Secondary databases like InterPro, Pfam, PROSITE, and SMART offer resources for motif scanning, aiding in functional annotation and insights into protein sequences.

According to Krogh et al., 2001, the presence of transmembrane helices indicates that a protein is located in the membrane. The Tampered analysis predicted two transmembrane helices with a strong preference for this orientation in the query protein. This analysis further confirmed the surface protein identity of the query protein. Shinde et al., (2023) stated using the Pfam database for the characterization of various domains in a query protein enables a deeper understanding of its functional and structural components (tabe 7). It provides insights into the protein's potential roles, interactions, and evolutionary context.

Fig. 1 the BLASTp query against the Swiss-Prot database with a 292-letter query sequence provides information about homologous matches, potential functional annotations, and insights into the relationships of the query protein with known sequences. Lai

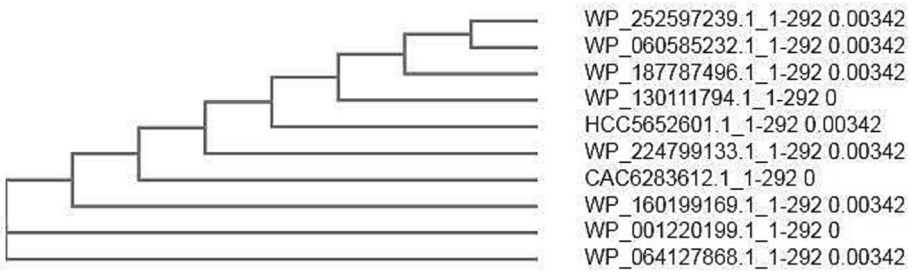


Fig. 3. NJ Tree for multiple aligned sequence.

Fig. 4. Homology Modeling of A Protein (3D Structure).

et al., (2000) a TBLASTN query with a 292-letter translated nucleotide sequence aims to identify potential protein-coding regions within the nucleotide sequence by comparing it to protein sequences in the database (Fig. 2). This approach is valuable for gene prediction, functional annotation, and understanding the potential coding capacity of DNA sequences.

The Neighbor-Joining (NJ) algorithm constructs a phylogenetic tree (Fig. 3) based on pairwise distances calculated from a multiple sequence alignment (Brinkman and Leipe, 2001). The NJ tree provides insights into the evolutionary relationships among the sequences and is valuable for understanding the relatedness and evolutionary history of species or sequences. Fig. 4 homology modeling is a powerful computational tool for predicting the three-dimensional structure of a protein based on the known structure of a related protein. It plays a crucial role in understanding protein structure–function relationships, drug discovery, and molecular studies. However, it's essential to consider its limitations and validate the resulting models rigorously (Kopp and Schwede, 2004).

5. Conclusion

The protein mentioned above, retrieved from SWISS-PROT underwent compositional, domain, fingerprint, secondary structure, and tertiary structure analyses using various software tools to establish its functions. The compositional analysis of the protein unveiled a higher number of aliphatic amino acids, and the instability index was determined to be 35%. Consequently, it was classified as a stable protein. The protein sequence was then subjected to Blast P and T blast n analyses to identify similar sequences.

Through Multiple Sequence Alignment using the Clustal tool, the related sequences revealed the presence of three motifs. Secondary structure analysis and transmembrane analysis indicated that the protein predominantly consists of helices in its secondary structure. By employing homology modeling with similar sequences from the SCOP database, the folded tertiary structure of the query protein was determined.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors extend their appreciation to the Researchers supporting project number (RSP2023R185), King Saud University, Riyadh, Saudi Arabia.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sjbs.2023.103812.

References

Ali, H., Mahmood, M., Adnan, M., Afzal, G., Perween, S., Akhtar, R.W., Jabeen, R., Asif, A.R., Faryal, S., Rahman, A., Muhammad, S.A., 2023. Comparative evolutionary and structural analyses of the TYRP1 gene reveal molecular mechanisms of biological functions in mammals. J. King Saud Univ.-Sci. 35, (6) 102772.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Attwood, G.T., Klieve, A.V., Ouwerkerk, D., Patel, B.K., 1998. Ammonia-hyper producing bacteria from New Zealand ruminants. Appl. Environ. Microbiol. 64 (5), 1796–1804.

Bairoch, A., Apweiler, R., 1998. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res. 26 (1), 38–42.

Brinkman, F.S. and Leipe, D.D., 2001. Phylogenetic analysis. *Bioinformatics: a practical guide to the analysis of genes and proteins*, *2*, p.349.

Chakroborty, D., Sarkar, C., Yu, H., Wang, J., Liu, Z., Dasgupta, P.S., Basu, S., 2011. Dopamine stabilizes tumor blood vessels by up-regulating angiopoietin 1 expression in pericytes and Krüppel-like factor-2 expression in tumor endothelial cells. Proc. Natl. Acad. Sci. 108 (51), 20730–20735.

Creighton, J.R., Parmeter, J.E., 1993. Metal CVD for microelectronic applications: An examination of surface chemistry and kinetics. Crit. Rev. Solid State Mater. Sci. 18 (2), 175–237.
de Pedro, M.A., Cava, F., 2015. Structural constraints and dynamics of bacterial cell wall architecture. Front. Microbiol. 6, 449.
Feng, D.F. and Doolittle, R.F., 1996. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. In *Methods in enzymology* (Vol. 266, pp. 368-382). Academic Press.
Ferretti, J.J., McShan, W.M., Ajdic, D., Savic, D.J., Savic, G., Lyon, K., Primeaux, C., Sezate, S., Suvorov, A.N., Kenton, S., Lai, H.S., 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc. Natl. Acad. Sci. 98 (8), 4658–4663.
Goldman, A.D., Weber, J.M., LaRowe, D.E. and Barge, L.M., 2023. Electron transport chains as a window into the earliest stages of evolution. *Proceedings of the National Academy of Sciences*, *120*(34), p.e2210924120.
Kopp, J., Schwede, T., 2004. Automated protein structure homology modeling: a progress report. Pharmacogenomics 5 (4), 405–416.
Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E.L., 2001. Predicting trans membrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305 (3), 567–580.
Lai, C.H., Chou, C.Y., Ch'ang, L.Y., Liu, C.S., Lin, W.C., 2000. Identification of novel human genes evolutionarily conserved in Caenorhabditis elegans by comparative proteomics. Genome Res. 10 (5), 703–713.
Li, R., Xu, J., Qi, Z., Zhao, S., Zhao, R., Ge, Y., Li, R., Kong, X., Wu, Z., Zhang, X. and He, Q., 2023. High-resolution genome mapping and functional dissection of chlorogenic acid production in Lonicera maackii. *Plant Physiology*, p.kiad295.
Lu, L., Li, G.Y., Swindlehurst, A.L., Ashikhmin, A., Zhang, R., 2014. An overview of massive MIMO: Benefits and challenges. IEEE J. Sel. Top. Signal Process. 8 (5), 742–758.
Mazmanian, S.K., Liu, G., Ton-That, H., Schneewind, O., 1999. *Staphylococcus aureus* sortase, an enzyme that anchors surface proteins to the cell wall. Science 285 (5428), 760–763.
McGinnis, S., Madden, T.L., 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 32 (suppl_2), W20–W25.
Milani, M., Curia, R., Shevlyagina, N.V., Tatti, F., 2023. Staphylococcus aureus. In: Bacterial Degradation of Organic and Inorganic Materials. Springer, Cham. https://doi.org/10.1007/978-3-031-26949-3_1.
Moreira, M.J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71 (4), 1027–1048.
Muhammad Rehman, H., Rehman, H.M., Naveed, M., Khan, M.T., Shabbir, M.A., Aslam, S. and Bashir, H., 2023. In Silico Investigation of a Chimeric IL24-LK6 Fusion Protein as a Potent Candidate Against Breast Cancer. *Bioinformatics and Biology Insights*, *17*, p.11779322231182560.
Rajagopal, M. and Walker, S., 2017. Envelope structures of Gram-positive bacteria. *Protein and sugar export and assembly in Gram-positive bacteria*, pp.1-44.
Sangeetha, D., Sumathi, V., Ishwarya, R., 2023. Biofilm formation-An overview and its control Measures. Curr. Res. Life Sci., 101
Shinde, T.V., Shinde, T.G., Chougule, V.V., Ghorpade, A.R., Utekar, G.V., Jadhav, A.S., Pawar, B.S., Sanmukh, S.G., 2023. Comparative functional genomics studies for understanding the hypothetical proteins in variant Microti 12. Open Bioinform. J. 16 (1).
Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., Durbin, R., 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. 26 (1), 320–322.
Thomson, E., Hanson, T.L., McLanahan, S.S., 1994. Family structure and child well-being: Economic resources vs. parental behaviors. Soc. Forces 73 (1), 221–242.
Ton-That, H., Liu, G., Mazmanian, S.K., Faull, K.F., Schneewind, O., 1999. Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of Staphylococcus aureus at the LPXTG motif. Proc. Natl. Acad. Sci. 96 (22), 12424–12429.
Whisstock, J.C., Lesk, A.M., 2003. Prediction of protein function from protein sequence and structure. Q. Rev. Biophys. 36 (3), 307–340.
Wu, H., Zhang, Y., 2011. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. 25 (23), 2436–2452.