Research Article

Abdulmajeed Alqurashi[#], Waqar Ahmad[#], Ziaur Rahman*, Javed Nawab, Muhammad Faisal Siddiqui, Ali Akbar, Ayman Ahmad Alkraiem, Muhammad Latif*

Identification of a novel drug target in *Porphyromonas gingivalis* by a computational genome analysis approach

https://doi.org/10.1515/chem-2024-0037
received February 14, 2024; accepted April 21, 2024

Abstract: This study applied a subtractive genomics approach to identify a potential drug target in the *Porphyromonas gingivalis* strain (ATCC BAA-308/W83). The aim was to characterize the whole proteome and hypothetical proteins (HPs) through structural, functional, and pathway predictions. The proteome was systematically reduced to identify essential proteins (EPs), non-homologous proteins (NHPs), and non-paralogous proteins (NPPs) while excluding those that were similar to the human proteome. Out of 1,836 proteins, the cluster database at high identity with tolerance algorithm identified 36 sequences as paralogous, having 80% identity. The resulting 1,827 proteins were compared to the human proteome using BLASTp (*e*-value 10^{-3}), resulting in 1,427 NHPs. These were then aligned with the DEG database using BLASTp (*e*-value of 10^{-5}), identifying 396 NHPs essential for pathogen survival. CELLO predicted the sub-cellular localization, and KEGG Automated Annotation Server identified potential metabolic pathways using a BLASTp similarity search of NHPs and EPs against the infrequently updated KEGG database. A total of 79 HPs essential for *P. gingivalis* were selected, and their molecular weights were determined. HPs were screened for metabolic pathway prediction, and the 3D structures of the proposed HPs were determined using homology modeling, and validation was performed. Only one HP (putative arginine deiminase) was qualified and found to be involved in the arginine and proline metabolic pathway.

Keywords: metabolic pathway prediction, periodontal disease, hypothetical proteins, drug target identification, *Porphyromonas gingivalis*, dental disease, drug resistance

1 Introduction

Porphyromonas gingivalis is a Gram-negative anaerobic bacterium that causes black pigmentation in billions of individuals worldwide. The bacterium is asaccharolytic in nature, extremely proteolytic, and is frequently discovered in deep periodontal pockets in humans [1]. *P. gingivalis* has been reported to be an etiological substance in the pathogenesis of periodontal disease, which leads to inflammatory events [2]. The subgingival plaque samples from patients revealed 88% involvement of this bacterium in chronic periodontitis [3]. *P. gingivalis* has also been found to be involved in many systemic diseases [4,5]. Infection caused by various strains of *P. gingivalis* has revealed that the bacterium is one of the causes of triggering varying degrees of cardiac disorders. The most frequent syndromes include endothelial dysfunction, proliferation of vascular smooth muscles, aortic aneurysms, and atherosclerosis [6–11]. Several antibiotics including Minocin, tetracycline HCl, and doxycycline are usually employed in the treatment of periodontal diseases. *P. gingivalis* is resistant to several of the known antibiotics. Rams et al. showed that *P. gingivalis* resistance to amoxicillin has increased (from

[#] Both authors equally contributed.

*** Corresponding author: Ziaur Rahman**, Department of Microbiology, Abdul Wali Khan University, Mardan, 23200, Khyber Pakhtunkhwa, Pakistan, e-mail: Zrahman@awkum.edu.pk
*** Corresponding author: Muhammad Latif**, Centre for Genetics and Inherited Diseases (CGID), Taibah University, Madinah, Kingdom of Saudi Arabia; Department of Biochemistry and Molecular Medicine, College of Medicine, Taibah University, Madinah, Kingdom of Saudi Arabia, e-mail: latifmayo@yahoo.com
Abdulmajeed Alqurashi, Ayman Ahmad Alkraiem: Department of Biology, College of Science, Taibah University, Medinah, Kingdom of Saudi Arabia
Waqar Ahmad: Department of Microbiology, Abdul Wali Khan University, Mardan, 23200, Khyber Pakhtunkhwa, Pakistan
Javed Nawab: Department of Environmental Sciences, Kohat University of Science and Technology, 26000, Kohat-Khyber, Pakhtunkhwa, Pakistan
Muhammad Faisal Siddiqui: Department of Microbiology, Hazara University, Hazara, 21120, Khyber Pakhtunkhwa, Pakistan
Ali Akbar: Center for Microbiology and Biotechnology, University of Swat, 19130, Swat, Khyber Pakhtunkhwa, Pakistan

0.1 to 2.8%), whereas resistance to clindamycin was low in 1999 but has increased to 9.3% by 2020 in periodontitis patients [12]. This bacterium contributes to the development of drug resistance in periodontal disease and is associated with periodontitis. The acquired resistance in patients is often anticipated by the excessive intake of antibiotics and self-medications, and not taking a proper dose at the proper time. Resistance in microorganisms causes treatment challenges in the community and poses health risks [13,14]. To overcome this challenge, there is an unmet need to investigate the genome and proteome of resistant organisms to identify novel drug targets. The discovery of new drugs against pathogens is time-consuming, tedious, and expensive and also associated with a high degree of uncertainty in terms of success rate. The use of *in silico* tools and highly sensitive genome sequencing are well-established alternative approaches to speed up the drug discovery process. Several bioinformatics tools are usually employed to explore the wide-ranging proteomic data obtained from microbial pathogens. In this regard, the use of subtractive genomic methodology has shown excellent potential for the prediction of possible drug targets in several virulent pathogens [15–17]. To date, a total of 19 genome sequences of *P. gingivalis* have been reported including 8 with complete sequences (strains W83, ATCC 33277, TDC60, HG66, A7436, AJW4, 381, and A7A1-28), and 11 high-coverage draft sequences (JCVI SC001, F0185, F0566, F0568, F0569, F0570, SJD2, W4087, W50, Ando, and MP4504). These sequences have been compiled into fewer than 300 contigs, but ~60–80% of these genes can be anticipated for their potential function with sufficient reliability. The rest of the genes are either hypothetical, well-conserved hypothetical, uncharacterized, or unexplored [18]. The hypothetical proteins (HPs) are those entities that are encoded by an established open-reading frame, but due to the lack of experimental evidence, their function has not yet been confirmed. The purpose of this investigation was to identify and characterize HPs and putative drug targets among the available pool of HPs in the *P. gingivalis* strain by employing *in silico* subtractive genome analysis. To conduct this study, a previously sequenced genome (strain ATCC BAA-308/W83) was used to identify novel drug targets in the HPs with higher accuracy by employing well-optimized

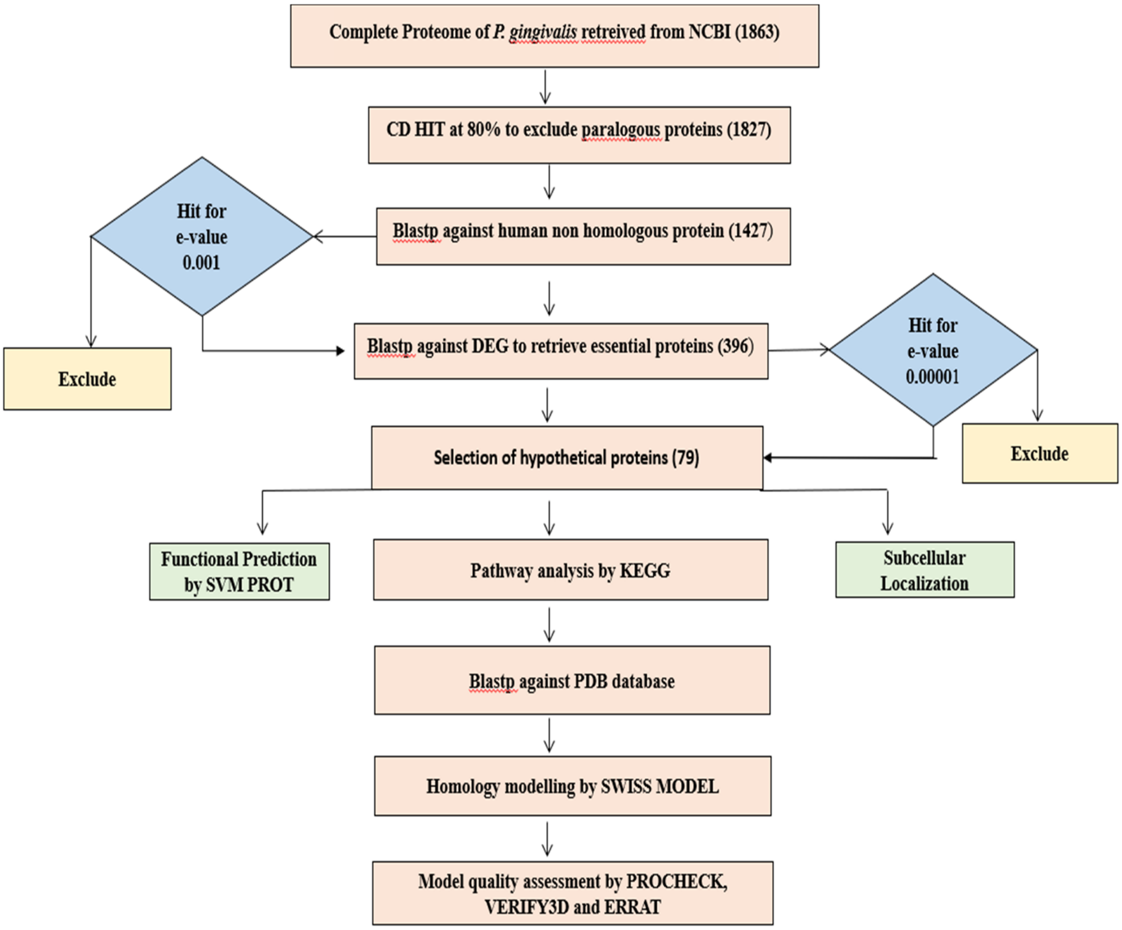

Complete Proteome of P. gingivalis retreived from NCBI (1863)
CD HIT at 80% to exclude paralogous proteins (1827)
Blastp against human non homologous protein (1427)
Hit for e-value 0.001
Exclude
Blastp against DEG to retrieve essential proteins (396)
Hit for e-value 0.00001
Exclude
Selection of hypothetical proteins (79)
Functional Prediction by SVM PROT
Pathway analysis by KEGG
Subcellular Localization
Blastp against PDB database
Homology modelling by SWISS MODEL
Model quality assessment by PROCHECK, VERIFY3D and ERRAT

Figure 1: Schematic description of a workflow and the outcome of an individual step involved in computational subtractive genomics-based target identification in *P. gingivalis.*

bioinformatics tools. The HPs were also analyzed using 3D structural prediction methods. To the best of our understanding, this study is the first report on strain (ATCC BAA-308/W83) using computational approaches to explore and provide an avenue to identify potential novel drug targets in *P. gingivalis,* a known causative agent of periodontal disease in humans.

2 Materials and methods

2.1 Sequence retrieval

The complete genomic strain of *P. gingivalis* (ATCC BAA-308/W83) comprising 1,863 protein sequences was retrieved from the NCBI database [19]. A schematic description of the workflow with the number of permissible genes at an individual screening phase is shown in Figure 1. The proteome of *Homo sapiens* was downloaded from the well-known UniProt database [20]. The UniProt database is among the world's most extensively annotated protein sequence databanks comprising over one million proteins [21]. The UniProt IDs with their locations and the resulting number of retrieved sequences for *P. gingivalis* at each step are shown in Table 1.

2.2 Elimination of paralogous sequences

A cluster database at high identity with tolerance (CD-HIT) was used to identify paralogous (duplicate protein) sequences that precisely clustered the proteins based on sequence identity. By applying the CD-HIT algorithm at 80% identity, the CD-HIT categorized the sequences in descending order. This operation generated the longest sequence of a representative of the first and foremost clusters. The rest of the sequences were compared with the resulting representatives of all clusters obtained by applying the CD-HIT (http://cd-hit.org). A sequence identity of 80% was chosen as a threshold value during the analysis. For a comparison purpose of every sequence, a short word filter was applied to the sequences to establish whether the resemblance was lower than the clustering threshold [22]. The total proteins of the subjected proteome were clustered and proteins with 80% identical were scrutinized to be paralogous.

2.3 Retrieving NHP sequences

The non-paralogous sequences (non-PS) obtained from the above operation were analyzed using BLASTp against the protein group of *H. sapiens* by applying a threshold expectation value (*e*-value) of 0.001. The sequence obtained through this operation comprised homologous sequences (HS) and non-HS (no hits were observed). Notably, the HS had a considerable resemblance (similarity) to the human host. The sequences that exhibited significant resemblance to the human host were eliminated, and the non-HS were subjected to further analysis [23].

2.4 Recognition of non-homologous essential genes

The Database of Essential Genes (DEG) is a source for predicting essential genes on the basis of homologous sequence searches [24]. DEG 6.8 was downloaded from the DEG official website (http://www.essentialgene.org/). Homology with sequences present in the DEG database provides a basis for the existence of non-homologous proteins (NHPs). To do this, the NHP sequences of the ATCC BAA-308/W83 were submitted to BLASTp against DEG with an *e*-value of 0.00001. The sequences retrieved from this approach included all potential NHPs (1,827) and non-homologous essential proteins (EPs) (396),

Table 1: Steps of subtractive genomic with resulting number of retrieved sequences

Sr. No.	Description of step	Retrieved sequences *P. gingivalis*
1	Total counts of proteins in the genome	1,863
2	Eliminated paralogous (>80% identical) in CD-HIT	1,827
3	Obtained proteins against *H. sapiens* using BLASTp (*e*-value 10^{-3})	1,427
4	EPs in DEGG (*e*-value 10^{-5})	396
5	HPs (essential, non-homologous, non-paralogue)	79
6	Functional prediction of HPs	75
7	HPs involve in metabolic pathway	01

which are essential genes for *P. gingivalis* but found in the host. These identified sequences were determined as essential and viable for the survival of the pathogen.

2.5 Selection of HPs

The HPs are those proteins whose functions are not known and are translated by an open-reading frame of the genome. A total of 79 HPs that were non-homologous, non-paralogous, and essential for *P. gingivalis* were selected from known functional proteins for further studies.

2.6 Sub-cellular localization prediction

The CELLO has been reported to be a useful tool for predicting sub-cellular localizations of the proteins present in proteomic data sets. The CELLO predicts the sub-cellular localizations of a query protein using an updated and well-trained model. The CELLO predicts four sub-cellular localizations in Gram-positive bacteria and five in Gram-negative bacteria [25]. A study demonstrated that cytoplasmic proteins are the most frequent drug targets [26]; however, membrane proteins are vaccine targets [27]. Using the CELLO v.2.5., we subjected the non-homologous genes, non-paralogous genes, and essential genes of *P. gingivalis* to prognosticate the sub-cellular localization.

2.7 Functional and family prediction of all NHPs

Support vector machine (SVM-Prot) is another web-based server applied for the categorization of proteins into functional families using the primary sequences [28]. SVM-Prot has a precise degree of capability to classify the distantly linked proteins and HPs of different functions and is applied as a tool for predicting protein functions. The non-PS, non-homologous sequences, and essential sequences of *P. gingivalis* were added to the online server of SVM-Prot to predict their functional family classification.

2.8 Proteins molecular weight

The Protein Molecular Weight database accepts the protein sequence and calculates their molecular weight. In this study, all the HPs were submitted to the Protein Molecular Weight database to evaluate their molecular weight.

2.9 Metabolic pathway analysis

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a popular tool for systematic screening of functions of a gene and associating genomic data with a higher level of functional annotations [27,29]. To identify potential metabolic pathways using KEGG, the widely used server KAAS (KEGG Automated Annotation Server) was used to perform a BLASTp similarity search of NHPs and EPs against the infrequently updated KEGG database [30]. Moreover, the KAAS not only provides key information related to the metabolic pathways but also suggests distinguishing features of the information that comprise KEGG Ontology (KO) designation lists. Moreover, information regarding the alternative pathways along with the corresponding enzymes and enzyme commission numbers can also be obtained via KASS. In the initial step, the BLAST score was computed between the request sequence and the standard sequence set. The reference or standard sequence was obtained from the KEGG GENE database. This process was used for the identification of homologs from the available reference set. In the second step, homologs ranked higher than the threshold level were carefully chosen as ortholog candidates. The selection was performed in accordance with the obtained BLAST score followed by bi-directional hit rate. Third, the potential ortholog contenders were distributed into various KO groupings corresponding to the annotations of the KEGG GENES database. In the last step, the level of assignment score was computed using probability and heuristics. The obtained results supported the prediction of possible metabolic pathways for potential drug targets.

2.10 Homology modeling

The HPs involved in the metabolic pathway and recognizable as possible targets for druggability were examined for existing crystal structures using BLAST and compared to the Protein Data Bank (PDB) database [31,32]. If the crystal structure was unavailable, then the protein sequence was used in homology modeling with the SWISS-MODEL [33]. The prototype structure was determined through the alignment of sequence compared to the available crystal structures in the PDB database. The prototype of the query protein (UniProt ID: Q7MXM8) was discovered as a crystal structure of putative arginine deaminase (PDB ID: 1ZBR. A). The SWISS-MODEL generated a modeled structure for protein and from this only one was carefully chosen. The obtained modeled structure was verified for its stereochemical quality using PROCHECK [34], VERIFY3D [35], and ERRAT servers [36].

2.11 Identification of domains and motifs

The biological functions and crucial roles of proteins are highly related to the domains and motifs of a given protein sequence. These motifs remain the main structural features that are well-conserved in the group of diverse types of proteins. Generally, proteins contain either single or more domains that act as crucial facilitators for protein functions [37]. The domain component is a functional and structural part of eukaryotic proteins that offers recommendations for the annotation of new proteins [38]. Examination of domain and motif of essential and druggable HP sequences was conducted using online tools of bioinformatics such as (i) Simple Modular Architecture Research Tool (SMART) [39], (ii) GenomeNet Motif search, and (iii) ScanProsite [40]. SMART is equipped with an end-user interface (http://www.bork.embl-heidelberg.de/Modules/sinput.shtml) to furnish a fast and automated annotation of the signaling domain arrangement of the sequence of the query protein. The outcome is described by a graphical display that shows the domain positions in a query sequence. Consequently, the resulting SMART sets of various signaling domains are meticulously annotated through hyperlinks to Medline and the Molecular Modeling Database, which can be easily accessed through Entrez [41]. This operation generated easy access to the information correlated with a sequence, homology, composition, and function.

3 Results and discussion

This study aimed to functionally characterize HPs and discover putative drug targets in *P. gingivalis* strain (ATCC BAA-308/W83). The putative drug targets were carefully screened by taking into consideration the drug-target-like benchmarks that demonstrated that they must be non-homologous to the host and crucial to the existence of bacteria while displaying participation in essential metabolic pathways of the bacteria. To this end, the two most important criteria for recommending a potential pharmacological target were that it be vital to the pathogen but not identical to the human host. There has been no experimental analysis to characterize the HPs present in *P. gingivalis* strain (ATCC BAA-308/W83) sequenced previously; hence, an effort was made to annotate the function of these HPs using an *in silico* approach. To describe and develop a relationship, a phylogenetic analysis of hypothetical proteins of *P. gingivalis* was constructed (Figure S1). With the purpose of identifying a potential drug target in *P. gingivalis* strain (ATCC BAA-308/W83), a computational subtractive genomics approach was applied that is a well-accepted and acknowledged approach used for the recognition and prioritization of diverse types of targets for druggability against many pathogens [42–45]. The complete workflow of this study is shown in Figure 1, and the resulting number of possible sequences at every step is shown in Table 1.

3.1 Identification of EPs, NHPs, and non-paralogous proteins (NPPs)

The full proteome of *P. gingivalis* strain (ATCC BAA-308/W83) was retrieved from the NCBI database. The examination of the downloaded proteome of *P. gingivalis* was found to comprise 1,863 proteins. The CD-HIT is a useful tool [46] that is widely employed to cluster the nucleotide or protein sequences, thereby reducing redundancy and manual efforts and enhancing the performance of other sequence analyses. After implementing the CD-HIT algorithm with an identity threshold of 80% (the used criterion), a total of 36 sequences were discovered as paralogous among the 1,836 proteins in the selected strain. It was noted that CD-HIT generated clustered the paralogous protein sequences consequently reducing the total number of sequences in the strain. The resulting 1,827 non-PS of proteome were further subjected to BLASTp (*e*-value 10^{-3}) against the human proteome (*H. sapiens* database) to remove the HS to humans. The output provided 1,427 NHPs.

3.2 Determination of essential genes

The genes that are considered essential for a sustainable life cycle of certain bacterium are known as essential genes. The DEG comprised a comprehensive listing of bacterial genes with a complete information on corresponding sequences that are indispensable for the existence of bacterial lifecycle [47]. Using DEG analysis, we aimed to identify the essential sequences of the bacterial pathogen in *P. gingivalis* but not present in the host organism. A total of 1,427 NHPs were compared to the DEG databank using

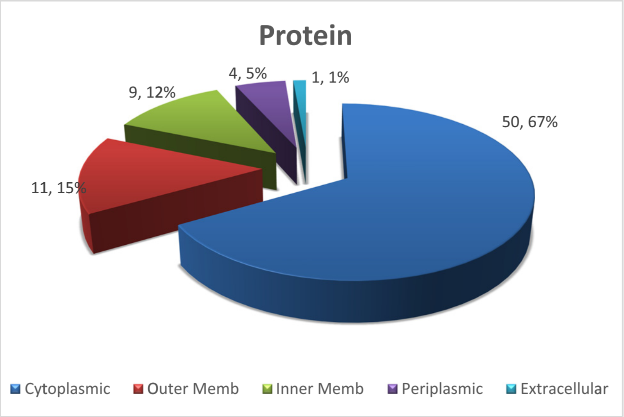


Figure 2: Schematic representation of sub-cellular localization of non-homologous EPs of *P. gingivalis* strain (ATCC BAA-308/W83).

the BLASTp by keeping the *e*-value of 10^{-5}. The outcome of this operation generated a total of 396 NHPs essential for the survival of pathogens which could have hypothetical or uncharacterized proteins. This approach conveniently screened the essential and HPs of the pathogen for subsequent analysis.

3.3 Sub-cellular localization prediction

Identifying the location of EPs is a crucial element in determining the important function of proteins in their specific cellular compartments. Recognizing the localization of a potential drug target is crucial in order to adjust the drug's mode of action towards the target. The CELLO has been reported to be a useful tool for the prognostication of sub-cellular localization of the proteins found in proteomic data. CELLO was used to categorize drug targets into three types: the membrane, cellular, and surface proteins [48]. Figure 2 shows the localization of 75 common drug targets in this study. The findings demonstrated that out of 75 HPs, 50 drug targets were resided in cytoplasmic (67%), 11 distributed in the outer membrane (15%), 9 found in the inner membrane (12%), 1 existed in the extracellular (1%), and 4 were found in the periplasmic proteins (5%).

3.4 Functional identification of shortlisted HPs

The understanding of protein function is a key parameter for investigating the biological phenomenon, disease mechanistic pathways, and discovering therapeutic targets [49]. Numerous bioinformatic tools have been used to characterize the HPs into their functional classes from many species. One such tool is SVMprot, which functionally classifies HPs to envision their functions on the basis of similarity and has shown good predictive performance. In this study, a total of 79 HPs were submitted to SVMprot. Only 4 sequences were found to fail because of their short amino acid length and the remaining 75 HPs were classified into several functional families containing transporters, zinc-binding proteins, and enzymes. We assigned the functional annotation to the proteins with strong confidence and concluded with the function of 75 HPs with high confidence. The outcome obtained from SVMport is presented in Table 2.

Table 2: Uniprot IDs with their potential locations

Cytoplasmic	Cytoplasmic	Cytoplasmic	Inner membrane	Outer membrane	Extracellular	Periplasmic
Q7MX20	Q7MW96	Q7MVV5	Q7MUT8	Q7MWV2	Q7MTZ1	Q7MT93
Q7MT62	Q7MXF2	Q7MVP6	Q7MW64	Q7MXN2	—	Q7MW66
Q7MVU6	Q7MWQ5	Q7MWS6	Q7MX12	Q7MXX1	—	Q7MXM5
Q7MX14	Q7MW58	Q7MXC7	Q7MT47	Q7MUS9	—	Q7MWE1
Q7MTY0	Q7MV76	Q7MV28	Q7MUH7	QTMXE2	—	—
Q7MXB9	Q7MW57	Q7MUW0	Q7MXN6	Q7MWS4	—	—
Q7MUC7	Q7M789	Q7MU54	Q7MVV6	Q7MSY4	—	—
Q7MXV1	Q7MU05	Q7MTV7	Q7MX69	Q7MWK8	—	—
Q7MTP4	Q7MXF9	Q7MWA4	Q7MTKB	Q7MVL8	—	—
Q7MTQ1	Q7MTH1	Q7MXM8	—	Q7MXB5	—	—
Q7MTH0	Q7MT58	Q7MW65	—	Q7MTU1	—	—
Q7MW19	Q7MW18	Q7MUL4	—	—	—	—
Q7MVS8	Q7MUX5	Q7MVZ8	—	—	—	—
Q7MWX9	Q7MT29	Q7MX05	—	—	—	—
Q7MV71	Q7MVB2	Q7MV50	—	—	—	—
Q7MWM9	Q7MUE7	Q7MWW5	—	—	—	—
Q7MVH1	Q7MVB9	—	—	—	—	—

Table 3: HPs Uniprot IDs with their predicted function and molecular weight

Sr. No.	Uniprot IDs	Predicted function	Molecular weight (kDa)
1	Q7MWW5	ATP-binding cassete family	10.11
2	Q7MX20	Transporter	11.91
3	Q7MT62	Enzyme	10.09
4	Q7MTK8	Transporter	48.9
5	Q7MUT8	Transporter	19.51
6	Q7MWE1	Transporter	11.32
7	Q7MWV2	DNA replication	37.54
8	Q7MW64	Transporter	17.17
9	Q7MXI2	ATP-binding cassette family	28.21
10	Q7MVU6	Transporter	27.41
12	Q7MX14	mRNA slicing	13.52
13	Q7MTY0	Lipid binding	7.64
14	Q7MXB9	Metal binding protein	10.01
15	Q7MT47	Transporter	123.36
16	Q7MT93	mRNA slicing	15.75
17	Q7MUC7	Enzyme	14.36
18	Q7MX25	ATP-binding cassette family	11.1
19	Q7MXV1	Transporter	64.01
20	Q7MTP4	ATP-binding cassette family	74.43
21	Q7MTQ1	Lipo protein	27.57
22	Q7MXN2	Transporter	26.25
23	Q7MTH0	ATP-binding cassette family	8.5
24	Q7MWI9	Transporter	20.87
25	Q7MVS8	Lipid binding protein	12.76
26	Q7MUH7	Metal binding	10.97
27	Q7MWX9	Metal binding	27.08
28	Q7MXX1	Transporter	43.27
29	Q7MV71	DNA repair	11.86
30	Q7MWM9	DNA repair	18.11
31	Q7MXN6	ATP-binding cassette family	39.33
32	Q7MVH1	Lipid binding protein	22.2
33	Q7MVV5	Transporter	15.61
34	Q7MWS6	DNA repair	34.42
35	Q7MUS9	Transporter	108.05
36	Q7MXE2	DNA repair	53.51
37	Q7MXC7	Enzyme	46.38
38	Q7MWS4	DNA replication	58.53
39	Q7MV28	DNA repair	47.44
40	Q7MUW0	Enzyme	23.67
41	Q7MW66	DNA repair	18.1
42	Q7MU54	ATP-binding cassette family	14.75
43	Q7MTV7	ATP-binding cassette family	6.09
44	Q7MXM5	DNA repair	20.19
45	Q7MSY4	Transporter	82.84
46	Q7MWK8	Enzyme	129.26
47	Q7MTZ1	Enzyme	32.33
48	Q7MWA4	Lipid binding protein	8.27
49	Q7MW96	ATP-binding cassette family	11.09
50	Q7MXF2	ATP-binding cassette family	18.44
51	Q7MWQ5	Zinc binding	11.36
52	Q7MVL8	Enzyme	54.95
53	Q7MVV6	Transporter	8.23
54	Q7MW58	rRNA binding protein	8.82
55	Q7MV76	Lipid metabolism	47.08
56	Q7MW57	Transporter	100.68
57	Q7M789	ATP-binding cassette family	6.09
58	Q7MU05	Enzyme	24.6
59	Q7MXF9	Transporter	48.36
60	Q7MTH1	DNA replication	55.99
61	Q7MT58	DNA repair	16.65
62	Q7MXB5	Transporter	56.55
63	Q7MUX5	ATP-binding cassette family	15.36
64	Q7MX69	Transporter	21.94
65	Q7MVB2	DNA condensation	132.15
66	Q7MUE7	Transporter	36.16
67	Q7MVP6	Transporter	14.62
68	Q7MXM8	DNA repair	38.3
69	Q7MTU1	Type II secretory pathway family	35.09
70	Q7MTB1	DNA repair	15.39
71	Q7MW65	DNA replication	23.28
72	Q7MUL4	Metal binding	49.9
73	Q7MVZ8	Metal binding	40.91
74	Q7MX05	ATP-binding cassette family	9.27
75	Q7MV50	Transporter	25.95

3.5 Metabolic pathway analysis

The KEGG database is a useful platform that furnishes a network of metabolic pathways along with their complete annotations. KEGG facilitates the screening of protein sequences that are necessary for serving a distinctive role in the process of metabolism. The metabolic pathway evaluation predicts a potential drug target on the basis of the pathogen's unique metabolism. Uddin et al. identified preferred drug targets in a *Mycobacterium avium* [50], we followed similar protocols and focused on the identification of crucial, hypothetical, and non-HS involved in the pathways of metabolism of the *P. gingivalis* strain using the KEGG database [51]. The KAAS utilized BLASTp for the assessment of proteins of interest against the KEGG databank and then annotated the associated

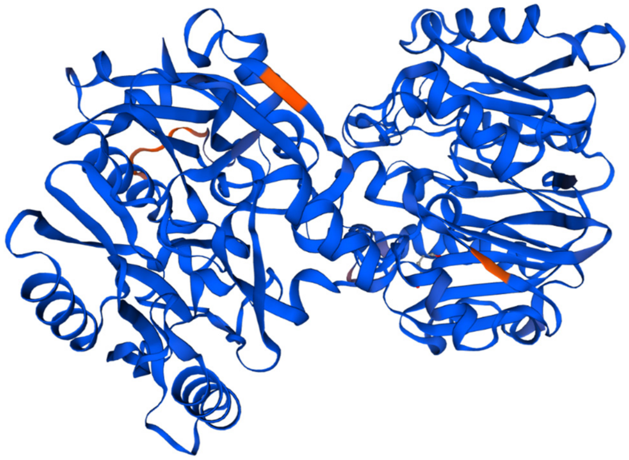
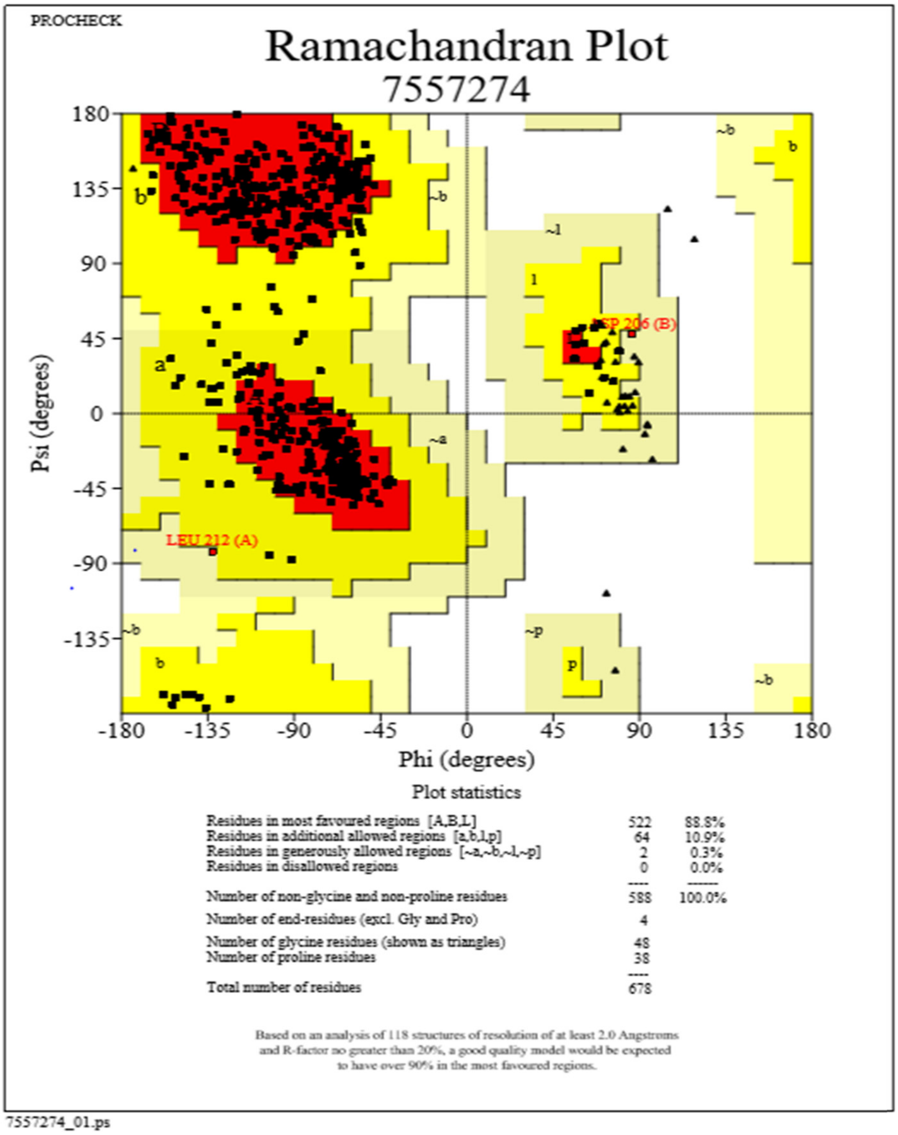


Figure 3: Developed 3D structure of query protein Q7MXM8.

potential role of the proteins. The obtained SVMprot results were subjected to the KEGG database analysis via the KAAS server. A total of 75 sequences of proteins of the *P. gingivalis* strain were screened via the KAAS server. As a result, out of 75 proteins, only one protein was passed by the KAAS and was discovered to be participate in the arginine and proline metabolic pathway (protein sequence with Id Q7MXM8 with KAAS server Ids: KO01100 metabolic pathways, KO00330). The distribution of HPs Uniprot IDs with their predicted function and molecular weight involved in different metabolic pathways is shown in Table 3.

3.6 Homology modeling

The SWISS-MODEL approach was employed for carrying out homology modeling to construct the three-dimensional



Plot statistics

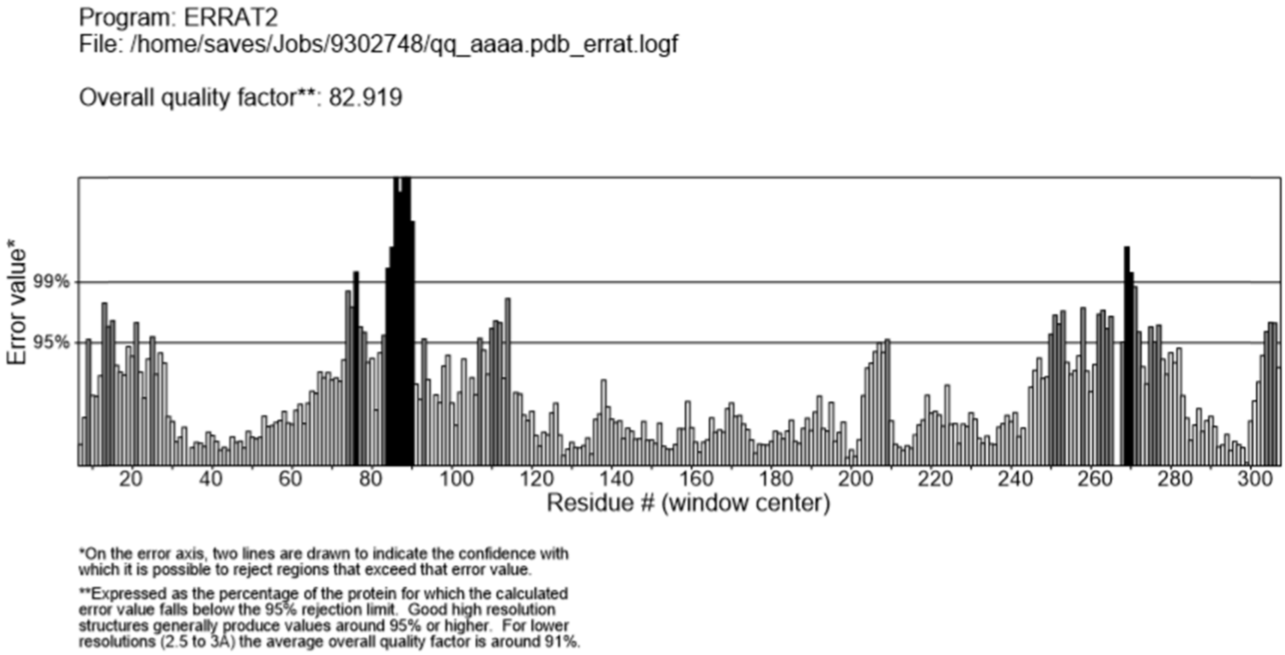
```
Residues in most favoured regions  [A,B,L]              522     88.8%
Residues in additional allowed regions  [a,b,l,p]        64     10.9%
Residues in generously allowed regions  [~a,~b,~l,~p]     2      0.3%
Residues in disallowed regions                            0      0.0%
                                                        ----   ------
Number of non-glycine and non-proline residues          588    100.0%

Number of end-residues (excl. Gly and Pro)                4

Number of glycine residues (shown as triangles)          48
Number of proline residues                               38
                                                        ----
Total number of residues                                678
```

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

Figure 4: Ramachandran plot of query protein Q7MXM8.



```
Program: ERRAT2
File: /home/saves/Jobs/9302748/qq_aaaa.pdb_errat.logf

Overall quality factor**: 82.919
```

*On the error axis, two lines are drawn to indicate the confidence with which it is possible to reject regions that exceed that error value.

**Expressed as the percentage of the protein for which the calculated error value falls below the 95% rejection limit. Good high resolution structures generally produce values around 95% or higher. For lower resolutions (2.5 to 3A) the average overall quality factor is around 91%.

Figure 5: The resulting output of ERRAT shows the quality factor.

structure of the HP [33]. Uddin et al. constructed homology modeling for a protein (WP_003899216.1); however, the used model was different [52]. In this study, for a query Q7MXM8, the 3D structure of a known arginine deaminase (PDB I.D: 1ZBR. A) was adopted as a modular structure. The three-dimensional structure had with 97% identity, 100% query coverage, and 50% positives. The developed protein motif segment is represented in Figure 3 which was verified by VERIFY3D, PROCHECK, and ERRAT. The Ramachandran plot is depicted in Figure 4 and determined by PROCHECK revealed that 88.8% of the residues were found within the most favored region, while 10.9% were located in the additional allowed region, 0.3% were located in the generously permitted region, and 0% were found in the disallowed region. The ERRAT revealed an overall quality factor of 82.9% (Figure 5), whereas VERIFY3D approved the structure, and at least 97.35% of the amino acids score was greater than 0.2 in the 3D/1D profile (Figure 6). Furthermore, the underlying secondary structural features were forecasted by the PSIPRED package as described in Figure 7. The findings revealed that the initial β strand was located near the *N*-terminus of the structure and consisted of amino acid residues 13 to 19. Conversely, the first α helix was also found to be located near the *N*-terminus and contained amino acid residues 28 to 46.

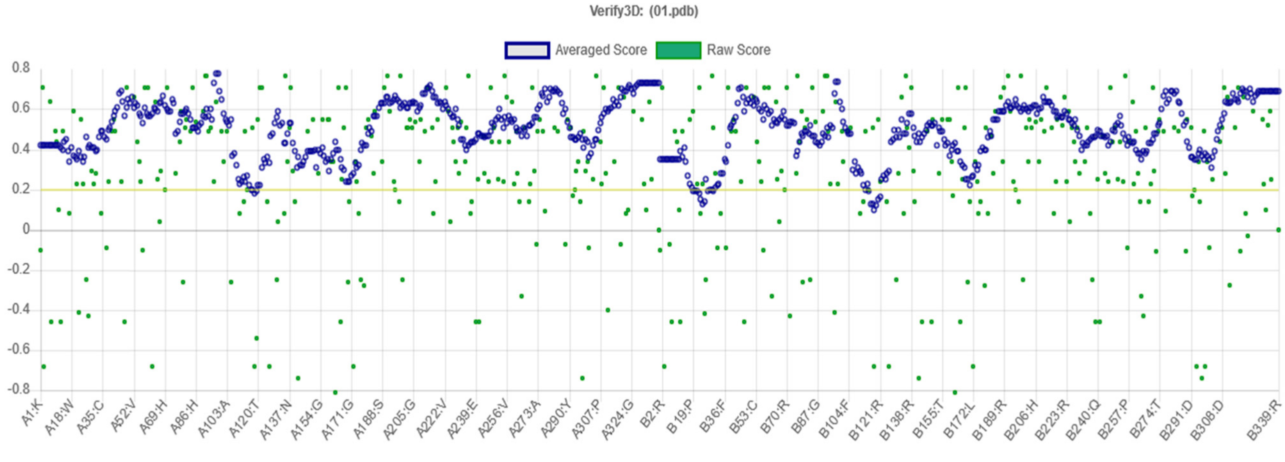


Figure 6: The result output of VERIFY3D shows 97% residues greater than 0.2.

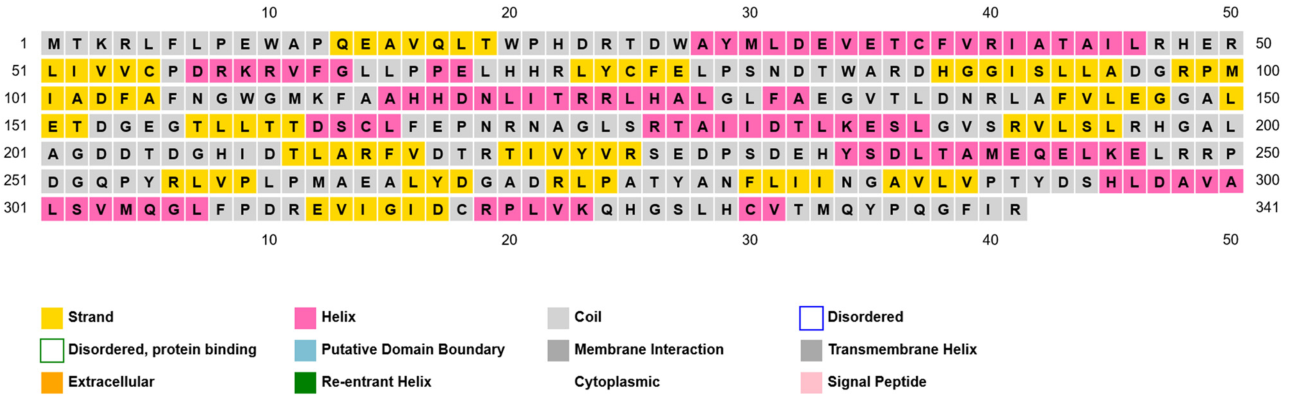

10
20
30
40
50
1 M T K R L F L P E W A P Q E A V Q L T W P H D R T D W A Y M L D E V E T C F V R I A T A I L R H E R 50
51 L I V V C P D R K R V F G L L P P E L H H R L Y C F E L P S N D T W A R D H G G I S L L A D G R P M 100
101 I A D F A F N G W G M K F A A H H D N L I T R R L H A L G L F A E G V T L D N R L A F V L E G G A L 150
151 E T D G E G T L L T T D S C L F E P N R N A G L S R T A I I D T L K E S L G V S R V L S L R H G A L 200
201 A G D D T D G H I D T L A R F V D T R T I V Y V R S E D P S D E H Y S D L T A M E Q E L K E L R R P 250
251 D G Q P Y R L V P L P M A E A L Y D G A D R L P A T Y A N F L I I N G A V L V P T Y D S H L D A V A 300
301 L S V M Q G L F P D R E V I G I D C R P L V K Q H G S L H C V T M Q Y P Q G F I R 341
Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide

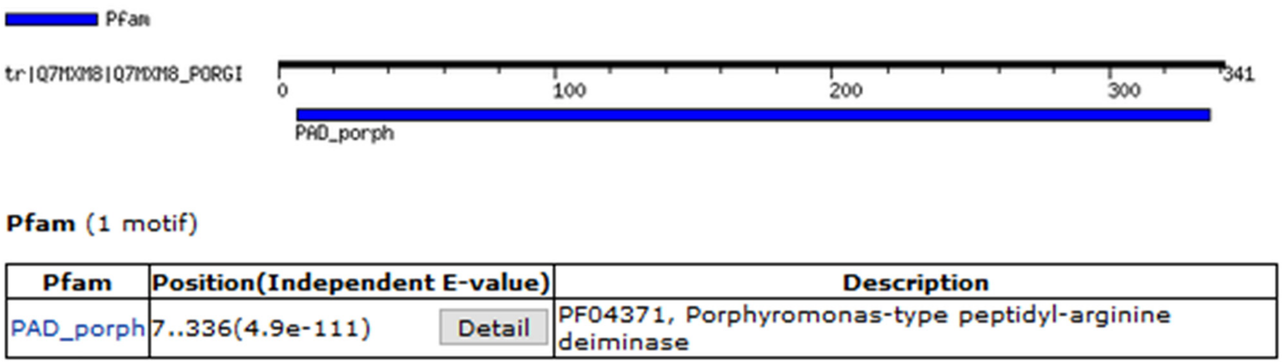
Figure 7: The graphical output from the PSIPRED program for the prediction of secondary structure of protein sequence Q7MXM8.

3.7 Identification of domain and motif

In this study, only one protein was identified by the KAAS server as a component of the essential metabolic pathways that are vital to the bacterial life cycle. The above-mentioned steps also indicated that the selected protein is unique and non-homologous to the human host; therefore, this protein can be considered as a potential drug target against the *P. gingivalis* strain. Additionally, a domain scan was also executed using SMART as described in section 2.12, which disclosed that the potential domain protein of interest (Q7MXM8) was peptidyl agmatine deiminase. The implementation of the ScanProsite tool distinguished the signature fits in the protein deposited as the query. In the output, no hit was observed in this study. The GenomeNet Motif Finder identified motifs against Pfam motif libraries [53] as peptidyl-arginine deiminase (pfam database accession, PF04371, peptidyl-arginine deaminase, https://www.ebi.ac.uk/interpro/entry/pfam/PF04371/) from position 7-336 with an *e*-value (4.9×10^{-111}) are shown in Figure 8.

4 Conclusions

A subtractive genomics methodology was utilized to the complete proteome of the *P. gingivalis* strain (ATCC BAA-308/W83) which shows resistance to several antibiotics. Sub-cellular localization prediction revealed that out of 75 HPs, cytoplasmic, outer membrane, inner membrane, extracellular, and periplasmic proteins had 50 (67%), 11 (15%), 9 (12%), 1 (5%), and 4 (1%) drug targets, respectively. Out of 75 protein sequences of the *P. gingivalis* strain, KAAS led to the qualification of only one HP (putative arginine deiminase) which was found to be involved in the arginine and proline metabolic pathway (protein sequence with Id Q7MXM8 with KAAS server Ids: KO01100 metabolic pathways, KO00330). Putative arginine deiminase was found as a part of an essential metabolic pathway of the bacterial life cycle. This study reports this HP as a prospective drug target for investigation. The suggested drug target can potentially be explored further through the application of structure-based methods to identify novel molecular



Pfam (1 motif)

Pfam	Position(Independent E-value)	Description
PAD_porph	7..336(4.9e-111) Detail	PF04371, Porphyromonas-type peptidyl-arginine deiminase

Figure 8: Identification of full-length peptide motif of the identified target protein (arginine deiminase).

entities as potential drug contenders against the drug targets. This study further demonstrates that the domain and motif in the proposed protein are needed for function and its blocking can inhibit the growth of *P. gingivalis*. This study offers valuable insights in discovering potential drug targets; however, wet laboratory experimental validation remains imperative to elucidate the precise interactions and efficacy of the target.

Acknowledgement: The authors would like to extend sincere appreciation to Dr. Haleema Masud of the Institute of Advanced Studies (IAS) at the University of Warwick, Coventry, UK, for her valuable contribution to language editing assistance during the revision stage.

Funding information: The authors are grateful to the Deputyship for Research & Innovation, Ministry of Education (MoE) Saudi Arabia for supporting funds for this research work (Project No. 442-61). The authors would also like to extend their gratitude to the Deanship of Scientific Research (DSR), Taibah University for its supervision and kind support.

Author contributions: Conceptualization, Z.A., and J.N.; methodology, W.A., A.A., and M.F.S.; validation of study, W.A., A.M.A., and A.A.; formal analysis, Z.A, M.L., A.M.A., A.A.A., and M.F.S.; data curation, Z.A., J.N., A.A.A., and W.A.; writing-original draft preparation, A.A.A., Z.A., and M.L.; writing-review and editing, A.M.A., M.L., A.A.A., and Z.A. Funding Acquisition, M.L., and Z.A; Supervision, Z.A., and M.L. All authors have read and agreed to the published version of the manuscript. *Waqar Ahmad (W.A.), Javed Nawab (J.N.), Abdul Majeed Alqurashi (A.M.A.), Muhammad Faisal Siddiqui (M.F.S.), Ayman Ahmad Alkraiem (A.A.A.), Ali Akbar (A.A.), Ziaur Rahman (Z.A.), and Muhammad Latif (M.L.).

Conflict of interest: The authors declare that they have no conflicts of interest.

Ethical approval: This study is not related to the involvement of either humans or animals.

Data availability statement: All data generated or analyzed during this study are included in this published article and its supplementary information file.

References

[1] Ruan J. Bergey's manual of systematic bacteriology (second edition) Volume 5 and the study of Actinomycetes systematic in China. Wei Sheng Wu Xue Bao. 2013;53(6):521–30. PMID: 24028053.

[2] Hajishengallis G, Wang M, Liang S, Triantafilou M, Triantafilou K. Pathogen induction of CXCR4/TLR2 cross-talk impairs host defense function. Proc Natl Acad Sci U S A. 2008;105(36):13532–7. doi: 10.1073/pnas.0803852105.

[3] Datta HK, Ng WF, Walker JA, Tuck SP, Varanasi SS. The cell biology of bone metabolism. J Clin Pathol. 2008;61(5):577–87. doi: 10.1136/jcp.2007.048868.

[4] Lundberg K, Wegner N, Yucel-Lindberg T, Venables PJ. Periodontitis in RA-the citrullinated enolase connection. Nat Rev Rheumatol. 2010;6(12):727–30. doi: 10.1038/nrrheum.2010.139.

[5] Olsen I, Progulske-Fox A. Invasion of Porphyromonas gingivalis strains into vascular cells and tissue. J Oral Microbiol. 2015;7:28788. doi: 10.3402/jom.v7.28788.

[6] Fukasawa A, Kurita-Ochiai T, Hashizume T, Kobayashi R, Akimoto Y, Yamamoto M. Porphyromonas gingivalis accelerates atherosclerosis in C57BL/6 mice fed a high-fat diet. Immunopharmacol Immunotoxicol. 2012;34(3):470–6. doi: 10.3109/08923973.2011.627866.

[7] Hayashi C, Viereck J, Hua N, Phinikaridou A, Madrigal AG, Gibson FC 3rd, et al. Porphyromonas gingivalis accelerates inflammatory atherosclerosis in the innominate artery of ApoE deficient mice. Atherosclerosis. 2011;215(1):52–9. doi: 10.1016/j.atherosclerosis.2010.12.009.

[8] Li L, Messas E, Batista EL Jr, Levine RA, Amar S. Porphyromonas gingivalis infection accelerates the progression of atherosclerosis in a heterozygous apolipoprotein E-deficient murine model. Circulation. 2002;105(7):861–7. doi: 10.1161/hc0702.104178.

[9] Madan M, Amar S. Toll-like receptor-2 mediates diet and/or pathogen associated atherosclerosis: proteomic findings. PLoS One. 2008;3(9):e3204. doi: 10.1371/journal.pone.0003204.

[10] Maekawa T, Takahashi N, Tabeta K, Aoki Y, Miyashita H, Miyauchi S, et al. Chronic oral infection with Porphyromonas gingivalis accelerates atheroma formation by shifting the lipid profile. PLoS One. 2011;6(5):e20240. doi: 10.1371/journal.pone.0020240.

[11] Brodala N, Merricks EP, Bellinger DA, Damrongsri D, Offenbacher S, Beck J, et al. Porphyromonas gingivalis bacteremia induces coronary and aortic atherosclerosis in normocholesterolemic and hypercholesterolemic pigs. Arterioscler Thromb Vasc Biol. 2005;25(7):1446–51. doi: 10.1161/01.ATV.0000167525.69400.9c.

[12] Rams TE, Sautter JD, van Winkelhoff AJ. Emergence of antibiotic-resistant Porphyromonas gingivalis in United States periodontitis patients. Antibiotics. 2023;12(11):1584. doi: 10.3390/antibiotics1211158.

[13] Pihlstrom BL, Michalowicz BS, Johnson NW. Periodontal diseases. Lancet. 2005;366(9499):1809–20. doi: 10.1016/s0140-6736(05)67728-8.

[14] Tribble GD, Lamont GJ, Progulske-Fox A, Lamont RJ. Conjugal transfer of chromosomal DNA contributes to genetic variation in the oral pathogen Porphyromonas gingivalis. J Bacteriol. 2007;189(17):6382–8. doi: 10.1128/jb.00460-07.

[15] Kuleš J, Horvatić A, Guillemin N, Galan A, Mrljak V, Bhide M. New approaches and omics tools for mining of vaccine candidates against vector-borne diseases. Mol Biosyst. 2016;12(9):2680–94. doi: 10.1039/c6mb00268d.

[16] Kumar Jaiswal A, Tiwari S, Jamal SB, Barh D, Azevedo V, Soares SC. An in silico identification of common putative vaccine candidates against Treponema pallidum: A reverse vaccinology and subtractive genomics based approach. Int J Mol Sci. 2017;18(2):402. doi: 10.3390/ijms18020402.

[17] Rizwan M, Naz A, Ahmad J, Naz K, Obaid A, Parveen T, et al. VacSol: A high throughput in silico pipeline to predict potential therapeutic

targets in prokaryotic pathogens using subtractive reverse vaccinology. BMC Bioinforma. 2017;18(1):106. doi: 10.1186/s12859-017-1540-0.

[18] Chen T, Siddiqui H, Olsen I. In silico Comparison of 19 Porphyromonas gingivalis strains in genomics, phylogenetics, phylogenomics and functional genomics. Front Cell Infect Microbiol. 2017;7:28. doi: 10.3389/fcimb.2017.00028.

[19] Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021;49(D1):D10–d7. doi: 10.1093/nar/gkaa892.

[20] Magrane M. UniProt Knowledgebase: A hub of integrated protein data. Database (Oxford). 2011;2011:bar009. doi: 10.1093/database/bar009.

[21] Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, et al. The Gene Ontology Annotation (GOA) database: Sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 2004;32(Database issue):D262–6. doi: 10.1093/nar/gkh021.

[22] Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9. doi: 10.1093/bioinformatics/btl158.

[23] Naveed M, Tehreem S, Mubeen S, Nadeem F, Zafar F, Irshad M. In-silico analysis of non-synonymous-SNPs of STEAP2: To provoke the progression of prostate cancer. 2016;11(1):402–16. doi: 10.1515/biol-2016-0054.

[24] Zhang R, Ou HY, Zhang CT. DEG: A database of essential genes. Nucleic Acids Res. 2004;32(Database issue):D271–2. doi: 10.1093/nar/gkh024.

[25] Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. 2004;13(5):1402–6. doi: 10.1110/ps.03479604.

[26] Bakheet TM, Doig AJ. Properties and identification of antibiotic drug targets. BMC Bioinforma. 2010;11:195. doi: 10.1186/1471-2105-11-195.

[27] Garmory HS, Titball RW. ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies. Infect Immun. 2004;72(12):6757–63. doi: 10.1128/iai.72.12.6757-6763.2004.

[28] Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 2003;31(13):3692–7. doi: 10.1093/nar/gkg600.

[29] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277–80. doi: 10.1093/nar/gkh063.

[30] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server issue):W182–5. doi: 10.1093/nar/gkm321.

[31] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The protein data bank. Acta Crystallogr D Biol Crystallogr. 2002;58(Pt 6 No 1):899–907. doi: 10.1107/s0907444902003451.

[32] Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): The single global macromolecular structure archive. Methods Mol Biol. 2017;1607:627–41. doi: 10.1007/978-1-4939-7000-1_26.

[33] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296–w303. doi: 10.1093/nar/gky427.

[34] Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: A program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993;26(2):283–91. doi: 10.1107/S0021889892009944.

[35] Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: Assessment of protein models with three-dimensional profiles. Methods Enzymol. 1997;277:396–404. doi: 10.1016/s0076-6879(97)77022-8.

[36] Colovos C, Yeates TO. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Sci. 1993;2(9):1511–9. doi: 10.1002/pro.5560020916.

[37] Goodacre NF, Gerloff DL, Uetz P. Protein domains of unknown function are essential in bacteria. mBio. 2013;5(1):e00744–13. doi: 10.1128/mBio.00744-13.

[38] Desler C, Suravajhala P, Sanderhoff M, Rasmussen M, Rasmussen LJ. In Silico screening for functional candidates amongst hypothetical proteins. BMC Bioinforma. 2009;10:289. doi: 10.1186/1471-2105-10-289.

[39] Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A. 1998;95(11):5857–64. doi: 10.1073/pnas.95.11.5857.

[40] de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006;34(Web Server issue):W362–5. doi: 10.1093/nar/gkl124.

[41] Hogue CWV, Ohkawa H, Bryant SH. A dynamic look at structures: WWW-entrez and the molecular modeling database. Trends Biochem Sci. 1996;21(6):226–9. doi: 10.1016/S0968-0004(96)80021-1.

[42] Barh D, Tiwari S, Jain N, Ali A, Santos AR, Misra AN, et al. In silico subtractive genomics for target identification in human bacterial pathogens. Drug Dev Res. 2011;72(2):162–77. doi: 10.1002/ddr.20413.

[43] Chhabra G, Sharma P, Anant A, Deshmukh S, Kaushik H, Gopal K, et al. Identification and modeling of a drug target for Clostridium perfringens SM101. Bioinformation. 2010;4(7):278–89. doi: 10.6026/97320630004278.

[44] Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D. In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. Silico Biol. 2006;6(1–2):43–7.

[45] Sarangi AN, Aggarwal R, Rahman Q, Trivedi N. Subtractive genomics approach for in silico identification and characterization of novel drug targets in Neisseria Meningitides Serogroup B. J Computer Sci Syst Biol. 2009;2(5):255–8. doi: 10.4172/jcsb.1000038.

[46] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2. doi: 10.1093/bioinformatics/bts565.

[47] Gao F, Luo H, Zhang CT, Zhang R. Gene essentiality analysis based on DEG 10, an updated database of essential genes. Methods Mol Biol. 2015;1279:219–33. doi: 10.1007/978-1-4939-2398-4_14.

[48] Zhou Y, Yang W, Kirberger M, Lee H-W, Ayalasomayajula G, Yang JJ. Prediction of EF-hand calcium-binding proteins and analysis of bacterial EF-hand proteins. Proteins: Struct Funct Bioinforma. 2006;65(3):643–55. doi: 10.1002/prot.21139.

[49] Das S, Sillitoe I, Lee D, Lees JG, Dawson NL, Ward J, et al. CATH FunFHMMer web server: Protein functional annotations using functional family assignments. Nucleic Acids Res. 2015;43(W1):W148–53. doi: 10.1093/nar/gkv488.

[50] Uddin R, Siraj B, Rashid M, Khan A, Ahsan Halim S, Al-Harrasi A. Genome subtraction and comparison for the identification of novel drug targets against Mycobacterium avium subsp. hominissuis. Pathogens. 2020;9(5). doi: 10.3390/pathogens9050368.

[51] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–d61. doi: 10.1093/nar/gkw1092.

[52] Uddin R, Siddiqui QN, Azam SS, Saima B, Wadood A. Identification and characterization of potential druggable targets among hypothetical proteins of extensively drug resistant Mycobacterium tuberculosis (XDR KZN 605) through subtractive genomics approach. Eur J Pharm Sci. 2018;114:13–23. doi: 10.1016/j.ejps.2017.11.014.

[53] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The Pfam protein families database. Nucleic Acids Res. 2002;30(1):276–80. doi: 10.1093/nar/30.1.276.