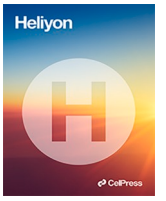
Heliyon 10 (2024) e31713

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Unveiling the mysteries: Functional insights into hypothetical proteins from *Bacteroides fragilis* 638R

Thomas Jebastin [a,1,*], M.H. Syed Abuthakir [b,c,1], Ilangovan Santhoshi [a], Muniraj Gnanaraj [d], Mansour K. Gatasheh [e], Anis Ahamed [f], Velusamy Sharmila [g]

[a] *Computer Aided Drug Designing Lab, Department of Bioinformatics, Bishop Heber College (Autonomous), Tiruchirappalli, 620017, Tamil Nadu, India*
[b] *Department of Bioinformatics, Bharathiar University, Coimbatore, 641046, Tamil Nadu, India*
[c] *Institute of Systems Biology, Universiti Kebangsaan Malaysia, 43600, UKM Bangi, Selangor, Malaysia*
[d] *Department of Biotechnology, School of Life Sciences, St Joseph's University, 36 Lalbagh Road, Bengaluru, 560027, Karnataka, India*
[e] *Department of Biochemistry, College of Science, King Saud University, P.O. Box 2455, Riyadh, 11451, Saudi Arabia*
[f] *Department of Botany and Microbiology, College of Science, King Saud University, Saudi Arabia*
[g] *Department of Biotechnology, Nehru Arts and Science College (NASC), Thirumalayampalayam, Coimbatore, 641 105, Tamil Nadu, India*

A R T I C L E I N F O

Keywords:
Bacteroides fragilis
Hypothetical proteins
Functional annotation
Gut microbiota
Essential genes
Gene ontology

A B S T R A C T

Humans benefit from a vast community of microorganisms in their gastrointestinal tract, known as the gut microbiota, numbering in the tens of trillions. An imbalance in the gut microbiota known as dysbiosis, can lead to changes in the metabolite profile, elevating the levels of toxins like *Bacteroides fragilis* toxin (BFT), colibactin, and cytolethal distending toxin. These toxins are implicated in the process of oncogenesis. However, a significant portion of the *Bacteroides fragilis* genome consists of functionally uncharacterized and hypothetical proteins. This study delves into the functional characterization of hypothetical proteins (HPs) encoded by the *Bacteroides fragilis* genome, employing a systematic *in silico* approach. A total of 379 HPs were subjected to a BlastP homology search against the NCBI non-redundant protein sequence database, resulting in 162 HPs devoid of identity to known proteins. CDD-Blast identified 106 HPs with functional domains, which were then annotated using Pfam, InterPro, SUPERFAMILY, SCANPROSITE, SMART, and CATH. Physicochemical properties, such as molecular weight, isoelectric point, and stability indices, were assessed for 60 HPs whose functional domains were identified by at least three of the aforementioned bioinformatic tools. Subsequently, subcellular localization analysis was examined and the gene ontology analysis revealed diverse biological processes, cellular components, and molecular functions. Remarkably, E1WPR3 was identified as a virulent and essential gene among the HPs. This study presents a comprehensive exploration of *B. fragilis* HPs, shedding light on their potential roles and contributing to a deeper understanding of this organism's functional landscape.

* Corresponding author.
E-mail address: stephajebastin@gmail.com (T. Jebastin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e31713
Received 21 February 2024; Received in revised form 21 May 2024; Accepted 21 May 2024
Available online 22 May 2024

1. Introduction

The gut microbiome represents a vast population of generally symbiotic bacteria, numbering between 10 and 100 trillion [1]. However, if certain species escape their natural habitats, they can become highly pathogenic, causing severe infections. This escape often occurs in instances of compromised gut health, such as in cases of ulcers, cancer, trauma, surgery, or other factors. *Bacteroides fragilis* (*B. fragilis*) is a beneficial commensal that provides benefits such as the digestion of complex polysaccharides and the development of immunity [2–6]. Among the species that can transition into an opportunistic pathogen, *B. fragilis* stands out. Gram-negative anaerobe *B. fragilis* normally occurs in the human gut as a commensal, constituting a small portion of gut bacteria. The Bacteroides stand out as the most frequent Bacteroides isolated from human infections despite its relatively low abundance [6]. The presence of a metalloprotease toxin, known as *B. fragilis* toxin (BFT), emphasizes the pathogenic potential of *B. fragilis*. This toxin can induce intestinal inflammation, thereby playing a role in the development of colorectal cancer [7,8].

Deciphering the key genes essential for *B. fragilis* survival provides novel insights that can be utilized in developing innovative treatments for bacterial infections and enhances understanding of the organism's resistance. The identification of these crucial genes opens avenues for targeted treatment approaches, advancing the management of *B. fragilis* infections [6]. However, a comprehensive understanding of the factors influencing *B. fragilis*'s unique pathogenicity is still lacking [9].

Hypothetical proteins (HPs) are proteins whose functions are not yet known. These proteins are often identified through genomic sequencing, but their roles in biological processes, virulence, and pathogenesis remains unclear [9]. Computational tools and techniques are used to analyze the physicochemical characteristics, protein-protein interactions, sub-cellular localization, functional classification, and antigenicity of these proteins, with a focus on their potential as drug targets [10,11]. Structural analysis of hypothetical proteins can provide insights into their functions and interactions with other molecules [12]. The identification and characterization of hypothetical proteins are important for understanding the complexity of the organism and its potential implications in normal and pathological conditions [13]. Therefore, in this investigation, we elucidated the molecular function of HPs derived from the *B. fragilis* strain 638R. Our approach involved employing an annotation-based workflow incorporating multiple *in silico* databases, tools and software. This methodology aimed to uncover the functions of HPs, potentially unveiling novel pharmacological targets for screening, drug development, and the design of treatments for *B. fragilis* infections.

2. Materials and methods

2.1. Sequence retrieval and similarity identification

The proteome of *B. fragilis* strain 638R (UP000008560) sourced from the Universal Protein Resource (UniProt) and consisting of 4284 protein sequences were employed in this study (The UniProt Consortium, 2023). A total of 379 HPs sequences were obtained in FASTA format. Subsequently, a homology search was conducted for the FASTA sequences of the 379 HPs using Blastp [14]. The search utilized the NCBI protein non-redundant (nr) database, considering hits with an identity range of 30–100 %, query coverage between 70 and 100 %, and an expected threshold value of 0.001. Sequences lacking identity with hypothetical proteins were selected for further analysis.

2.2. Assessment of operational biological characteristics

Protein domains are distinct molecular evolutionary units and are commonly associated with specific aspects of the molecular and cellular functions of a given protein sequence. The Conserved Domain Database (CDD-Blast) was utilized to search for conserved domains in cytoplasmic and periplasmic HPs using default parameters [15,16].

2.3. Functional annotation of HPs

The functional roles of HPs based on their similarity are elucidated with Pfam [17], InterPro [18], SUPERFAMILY [19], and SCANPROSITE [20]. Furthermore, SMART (Simple Modular Architecture Research Tool) [21] and CATH [22] were employed to investigate the functions of HPs, focusing on domain architecture and categorizing domains within the structural hierarchy, respectively. All these analyses were conducted using default parameters. HPs were selectively chosen for further analysis based on the presence of functional domains or motifs identified by at least three bioinformatic tools.

2.4. Exploring subcellular localization

Subcellular localization of the selected proteins was identified using CELLO [23], BUSCA [24], and PSORTb [25]. BUSCA and PSORTb integrate both experimental and computational datasets, while CELLO utilizes a two-level support vector machine (SVM)-based system.

2.5. Prediction of physiochemical parameters

Consequently, the physiochemical features of the selected HPs were estimated using the Protparam (Protein Parameters) interface on the ExPASy (Expert Protein Analysis System) server. Physical and chemical properties, including molecular weight (Mw),

theoretical isoelectric point (pI), number of negatively charged residues (Asp + Glu), number of positively charged residues (Arg + Lys), instability index, aliphatic index, and grand average of hydropathicity (GRAVY), were determined [26].

2.6. Prediction of gene ontology

The gene ontology of selected HPs was predicted based on the confidence scores using Argot$^{2.5}$ (Annotation Retrieval of Genel Ontology Terms) [27].

2.7. Virulent and essential genes prediction

The virulent proteins are identified based on a bi-layer cascade Support Vector Machine (SVM) using VirulentPred 2.0 [28]. Subsequently, the Database of Essential Genes (DEG) was employed to identify essential genes within the screened HPs [29]. The search was conducted against the genomes of *Bacteroides* using default parameters.

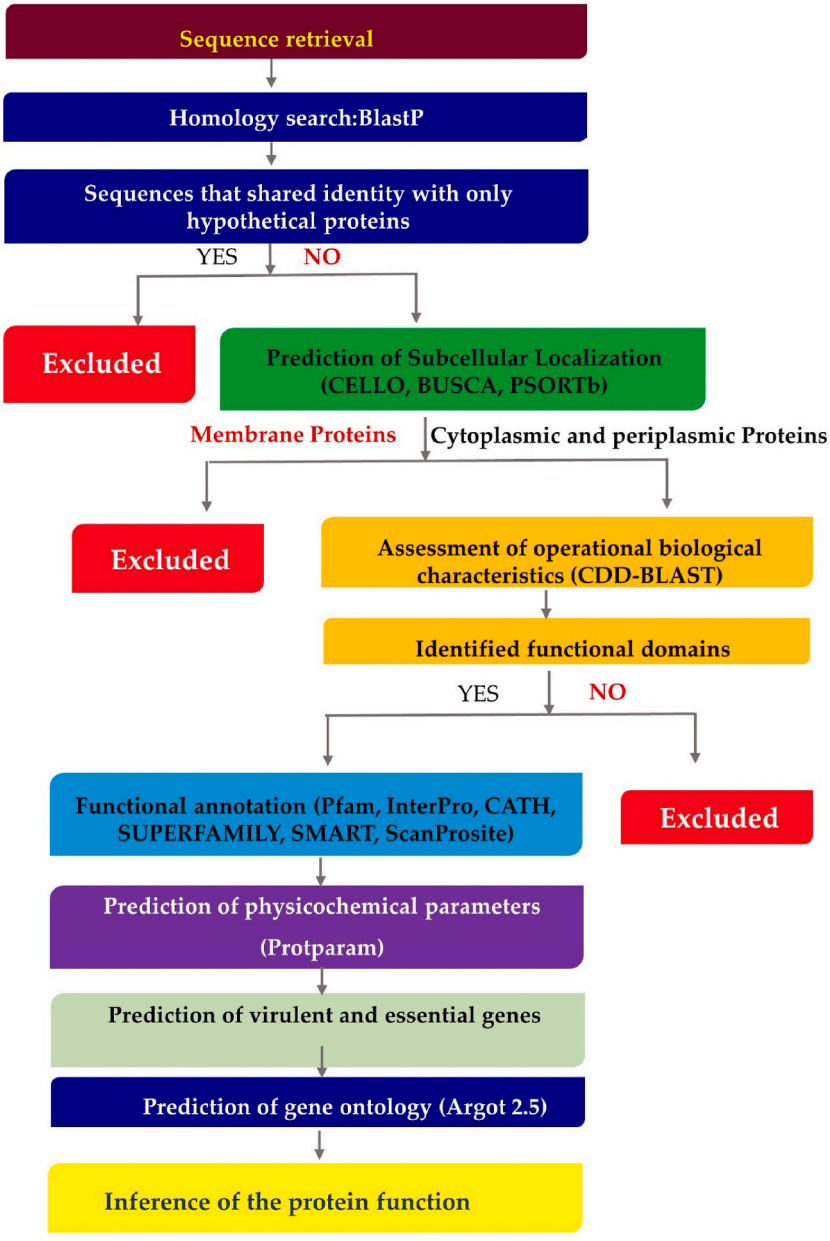


Fig. 1. Methodology that illustrates the overarching concept of the investigation.

3. Results

In this study, the HPs encoded by the *B. fragilis* genome underwent analysis through an *in silico* approach to identify proteins for categorization as either enterotoxins or non-enterotoxins. The workflow employed in this study is depicted in Fig. 1.

3.1. Sequence retrieval and similarity identification

A total of 379 Hypothetical Proteins (HPs) from the proteome of *B. fragilis* strain 638R were retrieved (Table S1). The NCBI protein non-redundant database was accessed using BLASTp to find homologous matches between the retrieved hypothetical proteins and known proteins. We excluded sequences that were similar to other hypothetical proteins due to the unknown functional roles of hypothetical proteins, it is critical to concentrate on HPs that are comparable to known proteins for future research. The remaining 162 HPs that show identity with known proteins were selected for biological characteristics assessment. (Table S2).

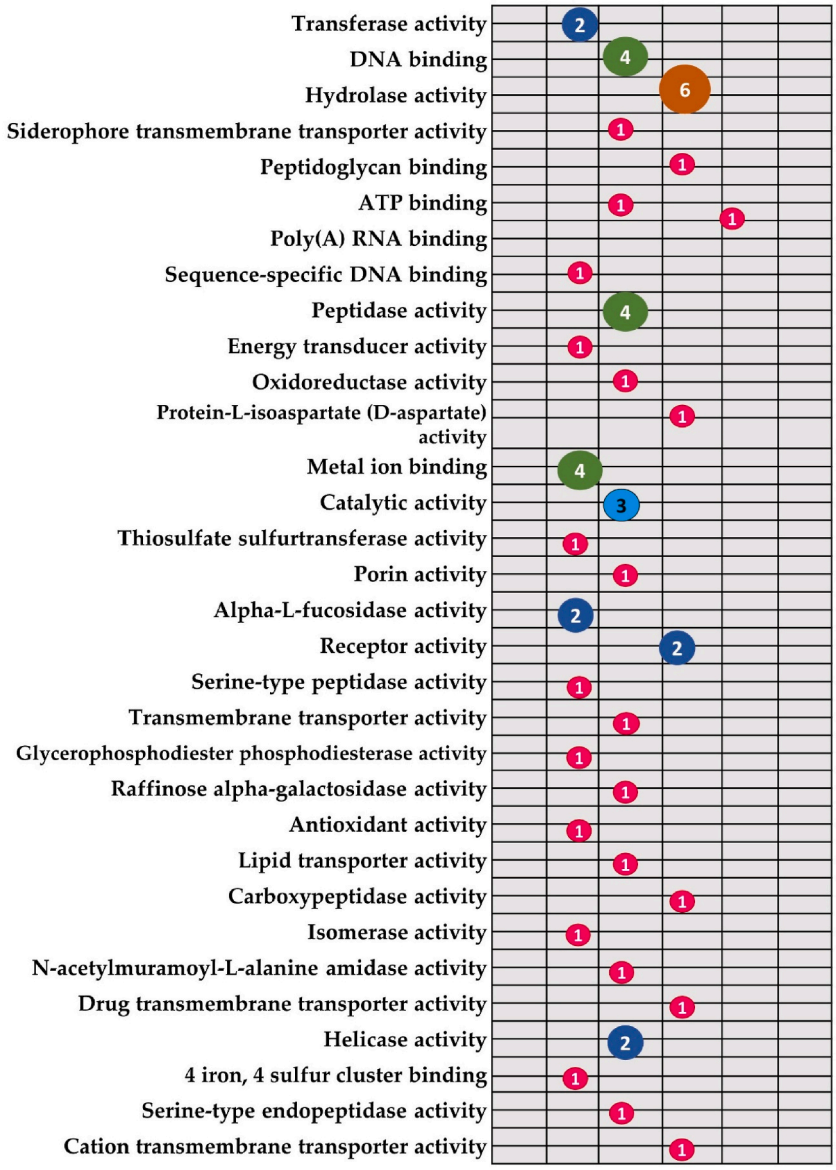


Fig. 2. Graph of molecular functions (In this graph, the distribution of GO terms is depicted on the Y-axis, and the size of the bubbles corresponds to the relative number of proteins found in each category).

3.2. Prediction of operational biological characteristics

CDD-Blast was used to search for conserved domains in cytoplasmic and periplasmic HPs. 106 HPs with identified functional domains were selected for functional annotation (Table S3).

3.3. Functional analysis of HPs

Pfam, InterPro, SUPERFAMILY, SCANPROSITE, SMART, and CATH were utilized to identify the functional roles of HPs (Table S4). Based on the presence of functional domains identified by at least three of the aforementioned bioinformatic tools, only 60 HPs were selected for examining physiological properties.

3.4. Subcellular localization

Predicting protein functions relies on a clear understanding of the subcellular localization of proteins. The effective transport of a protein to its designated location is crucial for its function, as proteins have evolved to operate optimally in specific subcellular localizations [30]. Subcellular localization of the retrieved proteins was predicted using CELLO, BUSCA, and PSORTb, categorizing HPs into membrane proteins, cytoplasmic proteins, and periplasmic proteins. Detailed results of subcellular localization are listed in Table S5.

3.5. Physiochemical properties

The physiochemical properties of proteins reflect their structural and functional characteristics. Physicochemical properties, including molecular weight, theoretical isoelectric point, charged residues, instability index, aliphatic index, and GRAVY, were predicted using Protparam. In the group of HPs, E1WSB3 (969AA) exhibits an extended protein length, while E1WTW5 (91AA) has the

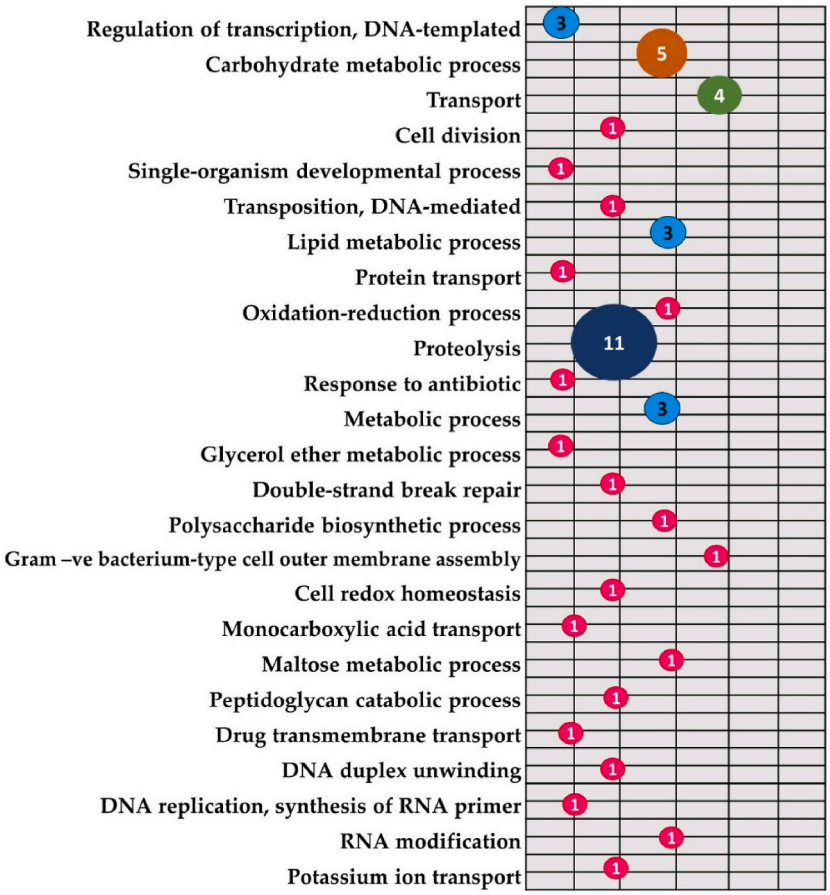


Fig. 3. Graph of biological processes (In this graph, the distribution of GO terms is depicted on the Y-axis, and the size of the bubbles corresponds to the relative number of proteins found in each category).

shortest length. The theoretical pI of HPs ranges from 4.52 to 9.54, and their determined molecular weight varies between 10.6 KDa and 108.4 KDa. The pH value at which a protein carries no net electrical charge is termed the isoelectric point (pI). Among 60 HPs 26 proteins exhibited an acidic nature (pI $<$ 7), while others were found to be basic. The aliphatic index (AI), used to assess protein thermostability, ranged from 66 to 115.9. The instability index (II) provided insights into *in vitro* protein stability, revealing that eighteen HPs were considered stable, while 42 were unstable. A cut-off value of $>$40 and $<$ 40 was employed to categorize proteins as stable and unstable, respectively. The GRAVY index, indicating a protein's interaction with water, showed negative values for 55 HPs, signifying their hydrophilic nature while the other 5 HPs are hydrophobic nature. Further detailed physiochemical data is presented in Table S6.

3.6. Prediction of gene ontology

Further, the gene ontology of HPs was analyzed in Argot$^{2.5}$ which provides results based on the confidence scores. Table S7 illustrates the distribution of hypothetical proteins (HPs) across three distinct Gene Ontology (GO) categories, encompassing the fundamental aspects of biological systems: cellular component, biological process, and molecular function. Specifically, there are 6 hypothetical proteins in the cellular component category, 25 in the biological process category, and 32 in the molecular function category. This result provides insight into diverse roles and locations in the cellular systems of these hypothetical proteins, shedding light on their potential functional significance. Among the three categories, molecular functions were observed to have the most extensive cluster, followed by Cellular components and biological processes. Thirty-two GO terminologies were identified in the molecular function category, primarily indicating hydrolase activity, with other terms related to DNA binding, metal ion binding and peptidase activity (Fig. 2). There were 25 different GO terminologies identified for biological processes, most of which (11 HPs) were associated with proteolysis whereas five HPs are involved in the carbohydrate metabolic process (Fig. 3). A total of six distinct GO terminologies were found in the cellular component category, of which 27 were specifically associated to the integral component of membrane (Fig. 4).

3.7. Detection of virulent essential genes

Among the HPs, 18 HPs are identified as virulent in VirulentPred 2.0 (Table 1). Then HPs were examined to identify essential genes in the DEG database against the genomes of Bacteroides. From the result, we find one essential protein (E1WPR3) among the targeted 60 HPs.

4. Discussion

In the present study, we utilized various effective computational tools and databases for annotating HPs. Initially, FASTA sequences of 379 HPs were retrieved from Uniprot. The UniProt proteome database is significant because it provides a comprehensive and freely accessible collection of protein sequences annotated with functional information [31,32]. A homology search using BLASTp was conducted for the FASTA sequences of the HPs to identify similarities and potential relationships with known proteins in the non-redundant protein databases. This approach allows for the inference of protein structure, function, and evolution [33]. The goal of the homology search was to analyze the protein sequences and detect similarities that may not be evident from sequence alone, but can be inferred from protein spatial structures [34]. Sequences lacking identity with hypothetical proteins were selected for further analysis. The purpose of selecting sequences lacking identity with only hypothetical proteins for further analysis was to validate the existence of predicted or hypothetical proteins at the protein level.

CDD-Blast was used in this study to search for conserved domains in cytoplasmic and periplasmic HPs. CDD offers live search capabilities and an archive of pre-computed domain annotations for sequences tracked by the NCBI's Entrez protein database [35]. 60 HPs with identified functional domains were selected for functional annotation. We aimed to annotate the locations of these identified functional domains, along with the inferred functional sites and motifs. This allows for a more comprehensive understanding of the functional characteristics of the sequences being analyzed [36]. We conducted a thorough analysis of the 60 HPs utilizing Pfam, InterPro, SUPERFAMILY, SCANPROSITE, SMART, and CATH. The analysis aimed to assign functions to HPs and determine if they

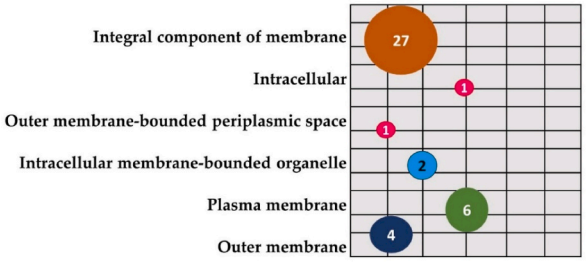


Fig. 4. Graph of cellular components (In this graph, the distribution of GO terms is depicted on the Y-axis, and the size of the bubbles corresponds to the relative number of proteins found in each category).

Table 1
Virulent gene prediction using VirulentPred 2.0

S.No.	Protein Name	Prediction results	Predicted Scores
1	E1WJS8	Virulent	0.2570
2	E1WJT5	Virulent	1.0074
3	E1WKT7	Virulent	0.8974
4	E1WLC6	Non-Virulent	−1.037
5	E1WMP7	Non-Virulent	−0.091
6	E1WNA1	Virulent	0.9642
7	E1WNL2	Non-Virulent	−0.983
8	E1WS18	Non-Virulent	−1.057
9	E1WS20	Non-Virulent	−1.024
10	E1WU44	Non-Virulent	−0.239
11	E1WU50	Non-Virulent	−1.016
12	E1WW97	Non-Virulent	−1.068
13	E1WWA1	Non-Virulent	−1.01
14	E1WQN4	Non-Virulent	−1.028
15	E1WJQ1	Virulent	0.9859
16	E1WK44	Virulent	0.9478
17	E1WKF5	Virulent	0.6376
18	E1WKH1	Non-Virulent	−0.865
19	E1WKH8	Virulent	0.9914
20	E1WKN0	Virulent	1.0163
21	E1WKT4	Non-Virulent	−0.864
22	E1WKZ5	Non-Virulent	−0.507
23	E1WLC7	Non-Virulent	−1.094
24	E1WLE3	Non-Virulent	−0.905
25	E1WLT5	Non-Virulent	−0.861
26	E1WLV2	Non-Virulent	−0.993
27	E1WN93	Virulent	1.0092
28	E1WNB4	Non-Virulent	−1.039
29	E1WNC6	Non-Virulent	−1.031
30	E1WNH4	Virulent	0.1975
31	E1WNN9	Virulent	0.0415
32	E1WPR3	Virulent	0.9633
33	E1WPR8	Virulent	0.9583
34	E1WQD7	Non-Virulent	−0.838
35	E1WRH7	Virulent	0.1103
36	E1WRQ8	Non-Virulent	−0.831
37	E1WTI3	Non-Virulent	−0.974
38	E1WTV0	Virulent	0.9814
39	E1WU07	Non-Virulent	−1.073
40	E1WUZ3	Virulent	0.1787
41	E1WVC0	Virulent	1.0455
42	E1WVD8	Virulent	1.0771
43	E1WVI3	Non-Virulent	−0.915
44	E1WWC4	Virulent	1.0163
45	E1WWP2	Non-Virulent	−0.67
46	E1WL48	Virulent	1.0926
47	E1WQN3	Non-Virulent	−1.01
48	E1WRI1	Non-Virulent	−1.019
49	E1WSB3	Virulent	1.1372
50	E1WTS9	Virulent	0.9848
51	E1WRA3	Virulent	0.6762
52	E1WN82	Non-Virulent	−0.918
53	E1WNR9	Non-Virulent	−0.069
54	E1WT37	Virulent	1.0425
55	E1WVP3	Non-Virulent	−0.954
56	E1WWX5	Non-Virulent	−1.089
57	E1WLZ0	Virulent	0.854
58	E1WS21	Non-Virulent	−0.903
59	E1WTB4	Non-Virulent	−0.913
60	E1WTW5	Virulent	0.95

exhibited similar functions according to three or more of the aforementioned programs [37,38]. The results showed that 19 HPs had similar functions based on the analysis of multiple programs. Computational approaches to identify the structural and functional qualities of a protein rely heavily on physiochemical properties, such as molecular weight, theoretical isoelectric point, charged residues, instability index, aliphatic index, and GRAVY, partition coefficient, aqueous solubility, pKa, melting point, etc. [39,40]. Physicochemical properties of proteins play a crucial role in their stability and activity of proteins [41]. The co-receptor binding affinities crucial for viral or bacterial tropism are governed by both the physicochemical and structural traits of proteins, particularly in

specific regions [42].

The subcellular localization of proteins is vital for their biological functions. Proteins perform their roles only when they arrive at their target location in the cell [43]. protein subcellular localization is important for determining protein function, revealing molecular interaction mechanisms, and developing drug targets [44].

Additionally, the gene ontology of selected HP was examined using Argot$^{2.5}$, yielding results based on the confidence scores. The identified functions were then classified into three GO categories: Cellular components, molecular functions, and biological processes. Numerous biological processes involving transcriptional regulation, DNA repair, and DNA modification depend on interactions between DNA and proteins, both sequence-specific and sequence-nonspecific. Aside from their few enzymatic and essential functions, some proteins could be involved in maintaining the gut microbiome [45].

There is still uncertainty regarding Bacteroides spp's ecological distribution, composition, and impact on health, despite the prevalence of Bacteroides spp in gut microbiota [46]. Gut microbiota includes *B. fragilis* 638R, which is commonly found in human guts. *B. fragilis* usually coexist harmlessly, but certain strains may pose infection risks, especially to those with weakened immune systems. The influence of it on health or disease depends on a variety of factors [47]. According to experimental animal models, bacteria like *B. fragilis* and *Bacteroides vulgatus* disrupt the intestinal epithelial barrier, which could lead to inflammation, as well as in patients with inflammatory bowel disease (IBD) [48–51]. In *B. fragilis,* specific proteins serve distinct functions.

The identification of potential drug targets is a crucial step in the development of novel therapeutic interventions, particularly in combating virulent bacterial strains. Identifying essential genes in the DEG database is crucial as it provides essential information about genes necessary for the survival of organisms [29]. This knowledge aids in understanding fundamental cellular processes that sustain life and can have implications for various research fields, including biology, genomics, and drug discovery [52]. Hence, in this study, we ultimately identified a candidate protein, E1WPR3, with implications in virulence as well as essential gene.

Since the Bam complex is known to play a crucial role in the biogenesis of proteins found on the surface of bacteria's outer membranes, our findings that E1WPR3 is a member of the BamD protein family are noteworthy. To ensure the appropriate folding and insertion of outer membrane proteins (OMPs), the Bam complex, which includes BamD and BamA are crucial [53,54]. This complex affects the functioning and integrity of the bacterial outer membrane. It has been shown that breaking down the Bam complex increases bacterial sensitivity to drugs and reduces their pathogenicity in several different species.

The possibility of E1WPR3's involvement in *B. fragilis* virulence pathways is indicated by its relationship with the BamD protein family. Bacterial pathogenicity-inducing mechanisms may be amenable to disruption by targeting proteins like BamD, which are involved in outer membrane protein synthesis [53,54]. Because of this, E1WPR3 could be a promising novel target to create new treatments to fight *B. fragilis* infections.

The use of other bioinformatics tools, such as SUPERFAMILY, SCOP, CDD-Blast, CATH, and InterPro, further supports our results by confirming that E1WPR3 is a member of the BamD protein family. The importance of E1WPR3 as a possible therapeutic target is highlighted by this agreement across many bioinformatics platforms, which increases the certainty of our findings.

5. Conclusion

In conclusion, our study presents ample proof indicating E1WPR3, identified as a member of the BamD protein family, as a possible therapeutic target for treating aggressive strains of *B. fragilis*. The development of BamD-targeting inhibitors and further investigation into the specific function of E1WPR3 in bacterial pathogenicity offers significant potential for the progress of antimicrobial treatment against illnesses caused by *B. fragilis*.

Human and animal rights

This study does not include human or animal subjects.

Data availability statement

All data are included in the manuscript as supplementary files:

CRediT authorship contribution statement

Jebastin Thomas: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **M.H. Syed Abuthakir:** Validation, Methodology, Data curation, Conceptualization. **Ilangovan Santhoshi:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muniraj Gnanaraj:** Writing – review & editing, Visualization. **Mansour K. Gatasheh:** Writing – review & editing, Validation, Funding acquisition. **Anis Ahamed:** Writing – review & editing. **Velusamy Sharmila:** Writing – review & editing.

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.

Acknowledgment

The work was supported by the Researchers Supporting Project number (RSP2024R393), King Saud University, Riyadh, Saudi Arabia.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e31713.

References

[1] P.J. Turnbaugh, R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight, J.I. Gordon, The human microbiome project, Nature 449 (7164) (2007) 804–810.
[2] D.W. Wareham, M. Wilks, D. Ahmed, J.S. Brazier, M. Millar, Anaerobic sepsis due to multidrug-resistant Bacteroides fragilis: microbiological cure and clinical response with linezolid therapy, Clin. Infect. Dis. 40 (7) (2005) e67–e68.
[3] H.M. Wexler, Bacteroides: the good, the bad, and the nitty-gritty, Clin. Microbiol. Rev. 20 (4) (2007) 593–621.
[4] I. Brook, The role of anaerobic bacteria in bacteremia, Anaerobe 16 (3) (2010) 183–189.
[5] E.B. Troy, D.L. Kasper, Beneficial effects of Bacteroides fragilis polysaccharides on the immune system, Front. Biosci.: J. Vis. Literacy 15 (2010) 25.
[6] Y. Veeranagouda, F. Husain, E.L. Tenorio, H.M. Wexler, Identification of genes required for the survival of B. fragilis using massive parallel sequencing of a saturated transposon mutant library, BMC Genom. 15 (2014) 1–11.
[7] E. Valguarnera, J.B. Wardenburg, Good gone bad: one toxin away from disease for Bacteroides fragilis, J. Mol. Biol. 432 (4) (2020) 765–785.
[8] C.E. DeStefano Shields, J.R. White, L. Chung, A. Wenzel, J.L. Hicks, A.J. Tam, C.L. Sears, Bacterial-driven inflammation and mutant BRAF expression combine to promote murine colon tumorigenesis that is sensitive to immune checkpoint therapy, Cancer Discov. 11 (7) (2021) 1792–1807.
[9] M. Naveed, S.I. Makhdoom, G. Abbas, M. Safdari, A. Farhadi, S. Habtemariam, S. Tehreem, The virulent hypothetical proteins: the potential drug target involved in bacterial pathogenesis, Mini Rev. Med. Chem. 22 (20) (2022) 2608–2623.
[10] P.B.S. Varma, Y.B. Adimulam, S. Kodukula, In silico functional annotation of a hypothetical protein from Staphylococcus aureus, Journal of infection and public health 8 (6) (2015) 526–532.
[11] A. Mukherjee, P. Dandapat, M.Z. Haque, S. Mandal, P.S. Jana, S. Samanta, C. Guha, Computational analysis of hypothetical proteins from Mycobacterium orygis identifies proteins with therapeutic and diagnostic potentials, Anim. Genet. (2023) 200154.
[12] I. Yuvaraj, S.K. Chaudhary, J. Jeyakanthan, K. Sekar, Structure of the hypothetical protein TTHA1873 from Thermus thermophilus, Acta Crystallogr. F: Structural Biology Communications 78 (9) (2022) 338–346.
[13] Y.N. Chirgadze, E.A. Boshkova, A.M. Kargatov, N.Y. Chirgadze, Functional identification of 'hypothetical protein'structures with unknown function, J. Biomol. Struct. Dyn. 41 (11) (2023) 5362–5366.
[14] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, T.L. Madden, NCBI BLAST: a better web interface, Nucleic Acids Res. 36 (suppl_2) (2008) W5–W9.
[15] A. Marchler-Bauer, M.K. Derbyshire, N.R. Gonzales, S. Lu, F. Chitsaz, L.Y. Geer, S.H. Bryant, CDD: NCBI's conserved domain database, Nucleic Acids Res. 43 (D1) (2015) D222–D226.
[16] S. Lu, J. Wang, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, A. Marchler-Bauer, CDD/SPARCLE: the conserved domain database in 2020, Nucleic Acids Res. 48 (D1) (2020) D265–D268.
[17] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L. Sonnhammer, A. Bateman, Pfam: the protein families database in 2021, Nucleic Acids Res. 49 (D1) (2021) D412–D419.
[18] M. Blum, H.Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, R.D. Finn, The InterPro protein families and domains database: 20 years on, Nucleic Acids Res. 49 (D1) (2021) D344–D354.
[19] J. Gough, K. Karplus, R. Hughey, C. Chothia, Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, J. Mol. Biol. 313 (4) (2001) 903–919.
[20] E. De Castro, C.J. Sigrist, A. Gattiker, V. Bulliard, P.S. Langendijk-Genevaux, E. Gasteiger, N. Hulo, ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, Nucleic Acids Res. 34 (suppl_2) (2006) W362–W365.
[21] I. Sillitoe, T.E. Lewis, A. Cuff, S. Das, P. Ashford, N.L. Dawson, C.A. Orengo, CATH: comprehensive structural and functional annotations for genome sequences, Nucleic Acids Res. 43 (D1) (2015) D376–D381.
[22] I. Letunic, S. Khedkar, P. Bork, SMART: recent updates, new developments and status in 2020, Nucleic Acids Res. 49 (D1) (2021) D458–D460.
[23] C.S. Yu, C.J. Lin, J.K. Hwang, Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions, Protein Sci. 13 (5) (2004) 1402–1406.
[24] C. Savojardo, P.L. Martelli, P. Fariselli, G. Profiti, R. Casadio, BUSCA: an integrative web server to predict subcellular localization of proteins, Nucleic Acids Res. 46 (W1) (2018) W459–W466.
[25] N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, F.S. Brinkman, PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26 (13) (2010) 1608–1615.
[26] V.K. Garg, H. Avashthi, A. Tiwari, P.A. Jain, P.W. Ramkete, A.M. Kayastha, V.K. Singh, MFPPI–multi FASTA ProtParam interface, Bioinformation 12 (2) (2016) 74.
[27] E. Lavezzo, M. Falda, P. Fontana, L. Bianco, S. Toppo, Enhancing protein function prediction with taxonomic constraints–The Argot2. 5 web server, Methods 93 (2016) 15–23.
[28] A. Sharma, A. Garg, J. Ramana, D. Gupta, VirulentPred 2.0: an improved method for prediction of virulent proteins in bacterial pathogens, Protein Sci. 32 (12) (2023) e4808.
[29] H. Luo, Y. Lin, T. Liu, F.L. Lai, C.T. Zhang, F. Gao, R. Zhang, DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools, Nucleic Acids Res. 49 (D1) (2021) D677–D686.
[30] C.S. Yu, Y.C. Chen, C.H. Lu, J.K. Hwang, Prediction of protein subcellular localization, Proteins: Struct., Funct., Bioinf. 64 (3) (2006) 643–651.
[31] Y.C. Lussi, M. Magrane, M.J. Martin, S. Orchard, Consortium UniProt, Searching and navigating UniProt databases, Current Protocols 3 (3) (2023) e700.
[32] The UniProt Consortium, UniProt: the universal protein knowledgebase in 2023, Nucleic Acids Res. 51 (D1) (2023) D523–D531, 2023.
[33] H. Ge, L. Sun, J. Yu, Fast batch searching for protein homology based on compression and clustering, BMC Bioinf. 18 (2017) 1–12.
[34] D.J. Nasko, K.E. Wommack, B.D. Ferrell, S.W. Polson, Fast and sensitive protein sequence homology searches using hierarchical cluster BLAST, bioRxiv (2018) 426098.
[35] A. Marchler-Bauer, J.B. Anderson, P.F. Cherukuri, C. DeWeese-Scott, L.Y. Geer, M. Gwadz, S.H. Bryant, CDD: a conserved domain database for protein classification, Nucleic Acids Res. 33 (suppl_1) (2005) D192–D196.
[36] M. Yang, M.K. Derbyshire, R.A. Yamashita, A. Marchler-Bauer, NCBI's conserved domain database and tools for protein domain analysis, Current protocols in bioinformatics 69 (1) (2020) e90.

[37] S. Guha, S. Das, S. Ganguli, A comparative genomics pipeline for in silico characterization and functional annotation of short hypothetical proteins, Journal of Tropical Life Science 10 (2) (2020).
[38] M.A. Rahman, U.H. Heme, M.A.K. Parvez, In silico functional annotation of hypothetical proteins from the Bacillus paralicheniformis strain Bac84 reveals proteins with biotechnological potentials and adaptational functions to extreme environments, PLoS One 17 (10) (2022) e0276085.
[39] F. Desai, R.K. Kamat, An android java application for evaluating physico-chemical properties of a protein sequence, Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146 3 (2017).
[40] M.I. Hossain, A.T. Asha, M.A. Hossain, S. Mahmud, K. Chowdhury, R.B. Mohiuddin, A.K.M. Mohiuddin, Investigating the role of hypothetical protein (AAB33144. 1) in HIV-1 virus pathogenicity: a comparative study with FDA-Approved inhibitor compounds through in silico analysis and molecular docking, Heliyon 10 (1) (2024) e23123.
[41] T.J. Kamerzell, C.R. Middaugh, The complex inter-relationships between protein flexibility and stability, J. Pharmaceut. Sci. 97 (9) (2008) 3494–3517.
[42] K. Bozek, T. Lengauer, S. Sierra, R. Kaiser, F.S. Domingues, Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage, PLoS Comput. Biol. 9 (3) (2013) e1002977.
[43] G. Dubourg-Felonneau, A. Abbasi, E. Akiva, L. Lee, Improving protein subcellular localization prediction with structural prediction & graph neural networks, bioRxiv 11 (2022), 2022.
[44] S. Wang, K. Zou, Z. Wang, S. Zhu, F. Yang, A novel multi-label human protein subcellular localization model based on gene ontology and functional domain, in: Proceedings of the 2023 3rd International Conference on Bioinformatics and Intelligent Computing, 2023, February, pp. 376–380.
[45] X.Y. Zeng, M. Li, Looking into key bacterial proteins involved in gut dysbiosis, World J. Methodol. 11 (4) (2021) 130;
[a] E. Lavezzo, M. Falda, P. Fontana, L. Bianco, S. Toppo, Enhancing protein function prediction with taxonomic constraints–The Argot2. 5 web server, Methods 93 (2016) 15–23.
[46] H.M. Wexler, Bacteroides: the good, the bad, and the nitty-gritty, Clin. Microbiol. Rev. 20 (4) (2007) 593–621.
[47] G.A. Botta, A. Arzese, R. Minisini, G. Trani, Role of structural and extracellular virulence factors in gram-negative anaerobic bacteria, Clin. Infect. Dis. 18 (Supplement_4) (1994) S260–S264.
[48] T. Shiba, Y. Aiba, H. Ishikawa, A. Ushiyama, A. Takagi, T. Mine, Y. Koga, The suppressive effect of bifidobacteria on Bacteroides vulgatus, a putative pathogenic microbe in inflammatory bowel disease, Microbiol. Immunol. 47 (6) (2003) 371–378.
[49] B.C. Sydora, S.M. MacFarlane, J.W. Walker, A.L. Dmytrash, T.A. Churchill, J. Doyle, R.N. Fedorak, Epithelial barrier disruption allows nondisease-causing bacteria to initiate and sustain IBD in the IL-10 gene-deficient mouse, Inflamm. Bowel Dis. 13 (8) (2007) 947–954.
[50] J. Dicksved, J. Halfvarson, M. Rosenquist, G. Järnerot, C. Tysk, J. Apajalahti, J.K. Jansson, Molecular analysis of the gut microbiota of identical twins with Crohn's disease, ISME J. 2 (7) (2008) 716–727.
[51] E. Sanchez, J.M. Laparra, Y. Sanz, Discerning the role of Bacteroides fragilis in celiac disease pathogenesis, Appl. Environ. Microbiol. 78 (18) (2012) 6507–6515.
[52] B. Gautam, K. Goswami, S. Singh, G. Wadhwa, Genome-wide essential gene identification in pathogens, Current trends in Bioinformatics: Insight (2018) 227–244.
[53] C.M. Sandoval, S.L. Baker, K. Jansen, S.I. Metzner, M.C. Sousa, Crystal structure of BamD: an essential component of the β-barrel assembly machinery of Gram-negative bacteria, J. Mol. Biol. 409 (3) (2011) 348–357.
[54] K.H. Kim, H.S. Kang, M. Okon, E. Escobar-Cabrera, L.P. McIntosh, M. Paetzel, Structural characterization of Escherichia coli BamE, a lipoprotein component of the β-barrel assembly machinery complex, Biochemistry 50 (6) (2011) 1081–1090.