Volume 13 • Number 4 • December 2022 • em00805

JOURNAL OF CLINICAL AND EXPERIMENTAL INVESTIGATIONS

RESEARCH ARTICLE

In silico subtractive genomics approach characterizes a hypothetical protein (MG_476) from *microplasma genitalium* G37

Naznin Jahan [1,2], *Tanvir Ahamed* [1], *Arun Das* [2], *Md. Arif Khan* [2], *Sharif Hossain* [1], *Satya Ranjan Sarker* [1], *Mohammad Mahfuz Ali Khan Shawan* [3*]

[1] Department of Biotechnology & Genetic Engineering, Jahangirnagar University, Dhaka, Bangladesh

[2] Bio-Bio-1 Bioinformatics Research Foundation, Dhaka, Bangladesh

[3] Department of Biochemistry and Molecular Biology, Jahangirnagar University, Savar, Bangladesh

ABSTRACT

Background: *Mycoplasma genitalium* is a gram negative, parasitic pathogenic bacterium, usually transmitted sexually into human and frequently causing urethritis in men and women as well as cervicitis and pelvic inflammation in women. This is an extremely small self-replicating entity whose genome has been sequenced. This genome sequencing is advantageous in understanding pathogenesis and identifying therapeutic targets. In this study different bioinformatics tools and databases were adopted to analyze the functions of different hypothetical proteins from *M. genitalium* G37.

Methodology: A total of 75 hypothetical proteins (HPs) were retrieved from KEGG database, while CDD-BLAST, Pfam, and InterProScan servers were used for conserved domains analysis. After that, those HPs were broadly analyzed for physicochemical properties, subcellular localization, GO annotation, and virulence factors.

Results: Based on best score, hypothetical protein MG_476 was selected for homology modelling which produced a fairly good quality 3D model. The active site within MG_476 was predicted using CASTp server that helps to explore the surface features of the protein. Other approaches include the use of NetCTL, IEDB, Bcepred, and ABCpred servers to predict the location of B and T cell epitopes. Among the CD8+ T cell epitopes tested, ILQIIMFIL scored highest (0.23718) in terms of immunogenicity.

Conclusion: Moreover, this analysis recommended MG_476 as a non-homologous protein and the data generated in this study may facilitate the experimental designing of novel drug and vaccines against *M. genitalium*.

Keywords: *M. genitalium* G37, sexually transmitted disease, urethritis, genital tract, HPs, CDD-BLAST, CASTp server and immunogenicity

Correspondence:

Mohammad Mahfuz Ali Khan Shawan

Address: Department of Biochemistry and Molecular Biology, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

Email: mahfuz_026shawan@juniv.edu

Received: 04.04.2022,
Accepted: 11.08.2022
https://doi.org/10.29333/jcei/12377

INTRODUCTION

With the availability of the human genome sequence and the sequence of many microbial genomes, novel approaches to understand host-pathogen interactions have been developed. Using bioinformatics and comparative analysis of the genome of a pathogenic microbe, one can identify essential genes necessary for the survival of that pathogen. Essential genes encode proteins not found in the host or not homologous to the host, which can be used as drug targets [1].

Mycoplasma genitalium is a pathogenic gram-positive bacterium that causes disease in humans. The cultural process of *M. genitalium* is fastidious and really challenging and even when successful, it takes several weeks or even months for each isolate to grow. This microorganism is involved in genitourinary infections and faces oxidative stress during the colonization of mucosal epithelium [2]. *M. genitalium* is the main culprit for sexually transmitted disease in women and is responsible for reproductive diseases particularly urethritis, infertility, and pelvic inflammation [3]. However, 475 protein-coding genes are

located in the genome of *M. genitalium* G37, and among them, 75 hypothetical proteins have been recognized with no known function. The hypothetical proteins are proteins whose presence has been predicted, but functions are not readily assigned [4].

The genomes of many organisms are still incompletely annotated and contain genes and proteins with unknown functions and structures. The bioinformatics approach could be an excellent alternative to laboratory-based methods to estimate functional and structural annotation of hypothetical proteins. 3D structures are more evolutionary conserved than sequences, and they are a great source of information [5]. The hypothetical complete protein sequences of *M. genitalium* G37 were downloaded from the KEGG database (http://www.genome.jp/kegg/) and saved as FASTA files for further analysis. Each sequence was also assigned an NCBI GI accession number and a Uniprot ID [6]. Three web-tools such as CDDBLAST, INTERPRO and Pfam were used in the present study to search the presence of conserved domain in 75 hypothetical proteins (HPs) [7]. The proteins that contained at least one domain by any of the servers were considered. Among the remaining 75 proteins, 41 HPs contained at least one domain. One of these HPs was MG_476 [8].

After that the study is primarily directed at structural characteristics such as physicochemical properties, subcellular localization, and functional annotation of the *M. genitalium* hypothetical protein MG_476 [9, 10]. In addition, homology modeling to generate a three-dimensional (3D) model, primary and secondary sequence structure analysis, active sites, protein-protein interaction and T cell, and B cell prediction were performed [8, 11].

METHODOLOGY

Structural Characterization

Retrieval of the sequence of MG_476 and prediction of its conserved domain

Protein MG_476: Hypothetical protein sequence was retrieved from the KEGG (Kyoto Encyclopaedia of Genes and Genomes) database. Primary NCBI accession numbersABC59632, Uniprot entry number P58061, and entry name SecG MYCGE.A multi-tool bioinformatics program including CDD-BLAST, Pfam, and InterPro was utilized to reveal conserved domains within close orthologous family members of *M. genitalium* hypothetical proteins. The CDD-BLAST is a simple technique that searches multiple databases containing domain models to identify conserved domains in protein queries [12]. The protein family database Pfam contains a massive collection of protein families. Each family is represented by multiple sequence alignments and Hidden Markov Models (HMMs) [13]. A tool that combines various protein signature recognition methods into one resource is called InterPro [14].

Physicochemical features and subcellular localization of MG_476

With ExPASy's ProtParam program, the MG_476 hypothetical protein was evaluated for its physicochemical properties, including atomic composition, amino acid composition, molecular weight, theoretical pI, grand average of hydropathicity (GRAVY) molecular weight, and stability index [15]. Protein subcellular localization can help explain why a protein is a therapeutic candidate or vaccine candidate. In order to determine the subcellular localization of MG_476, PSORTb v.3.0 was used. Additionally, PSORTb results were verified using the PSLpred server [16].

Functional annotation of MG_476 HP via gene ontology

The annotation of hypothetical protein MG_476 was performed by a specialized server Argot2.5 (annotation retrieval of gene oncology terms) [17]. First, the sequence in FASTA formatwas processed using BLAST and HMMER searches vs. UniProKB and Pfam databases, respectively. These sequences were then annotated with gene ontology (GO) terms (biological method, molecular function, and cellular component) retrieved from the UniProtKB-GOA info [9].

Prediction of Structure

Homology modeling of MG_476

An alternate method to build the 3D model of a protein based on sequence homology uses sequence alignment to find the most similar 3D structure, which considers the conserved portions, loops, and side chains in the database. For example, using potential template methods, Phyre2 was used to predict the 3D structure of MG_476, a protein [18].

Generated 3D structure model validation

The generated 3D model of MG_476 validation was performed by the PROCHECK program, which creates the Ramachandran plot [19]. Models are further validated by QMEAN (qualitative model energy analysis), which calculates a global score of the whole model reflecting the predicted model reliability ranging from mean 0 to standard deviation 1 [20]. In addition, the model of MG_476 was validated by Verify3D, ERRAT, ProSA [21, 22]. Finally, protein structure visualization tools PyMOLand JMOL were used to facilitate simple visualization of all structures, binding sites, docking epitopes, structural similarities, and alterations.

Secondary structure determination of MG_476

Secondary structure from its amino acid sequence was predicted by using SOPMA (self-optimized prediction method with alignment) [23] and PSIpred. It is based on the frequency of residues that initiate a helix, sheets, and turns [24].

Active sites prediction

Active sites of proteins and DNAs are often associated with structural pockets and cavities. In this study, the CASTp (computed atlas of surface topography of proteins) tool was used to identify all pockets and measure voids a protein's PDB structure. In addition, JMOL plugins facilitate the visualization of annotated residues and pockets [25].

Prediction of Protein-Protein Interaction

In order to evaluate and discover information related to protein-protein interactions, the STRING database has been developed. Indirect (functional) and direct (physical) protein-protein interactions are added to the list from four sources: genomic context, high throughput experiments, conserved/co-expressed, and previous studies [26]. The protein interacting partner searches were performed using STRING 9.05 to determine what proteins interact with phyre2-modeled proteins [27].

T Cell and B Cell Prediction

Prediction of CD8+ T-cell and CD4+T-cell epitopes

NetCTL 1.2, an application based on neuronal network architecture, was used to identify the CD8+ T cell epitope candidates derived from the in vivo processing of peptides [28]. In addition, artificial neural networks are used to predict the binding of MHC class I to 12 MHC supertypes, and the cleavage of MHC class I. As a second prediction, IEDB (immune epitope database) was utilized in order to identify alleles associated with these epitopes (MHC 1-binders) and to determine IpMHC immunogenicity scores [29]. Finally, using the HLApred and IEDB (MHC II binding prediction tool), the CD4+T-cell epitopes were predicted. For this prediction, seven abundant HLA class II alleles DRB1*01:01, DRB1*03:01, DRB1*04:01, DRB1*07:01, DRB1*11:01, DRB1*13:01, and DRB1*15:01 were selected from the selection panel of both servers [30].

Continuous (linear) and discontinuous epitope identification

A combination of Bcepred and ABCpred was used to predict Linear B cells [31]. In the context of predicting discontinuous (configurational) epitopes for B cells, FLLIPRO integrated a protein's 3D structure model [32].

RESULTS AND DISCUSSION

Conserved Domain Assessment of MG_476

Using three web-based tools such as CDDBLAST, INTERPRO, and Pfam, the present study examined whether the MG_476 hypothetical protein contains conserved domains. The protein contained at least one domain by any of the servers was considered. CDD-BLAST conserved domain searching program suggested that MG_476 contained SecGsuper family domain which is currently classified as protein of unknown function. There is also a consensus regarding Pfam and InterProScan's predictions MG_476 does not depict any domain. Therefore, the confidence level of MG_476 HP of *M. genitalium* was 33.3%.

Physiochemical Characteristics of MG_476

The physicochemical characteristics of protein were analyzed by ExPASy's ProtParam server including the amino acid sequence of the hypothetical protein MG_476. The hypothetical protein MG_476 contains 77 amino acids, with a molecular weight of 8543.56 Daltons. In addition, the isoelectric point (PI) of MG_476 was 9.78, which means that, it is a positively charged protein since above seven indicates a positively charged protein and an instability index of 33.79 denotes a stable protein. The positive GRAVY range 1.253 points out the possibility of being a hydrophobic protein rather than hydrophilic. Lysine (13) and isoleucine (13 each), the most abundant amino acid residue was found in MG_476 HP. The lowest amino acid residues were cysteine (1) and glutamic acid (1).

Determination of Subcellular Localization of MG_476

Knowing where unknown proteins are located in the cell could provide insight into their cellular function. This information helps select proteins for further study, and helps develop new drugs, and understand disease mechanisms [33]. For example, the subcellular localization of the hypothetical protein MG_476 was predicted to be a cytoplasmic-membrane protein with a localization score of 9.55 out of 10 by PSORTbv3.0. On the other hand, it was predicted to be an inner membrane protein by PSLPRED, and the expected accuracy was 90.20% [34].

Determination of Probable Function Based on GO Annotation

Ontology for gene expression specifies each gene product's biological process, molecular function, and cellular component categories [35]. Various functional classes contain various enzymes, different types of binding protein, catalytic proteins, regulatory proteins, carriers, and transporters. Annotation of the MG_476 protein indicates that transport across a membrane is driven by the hydrolysis of diphosphate bonds of inorganic pyrophosphate, ATP, or another nucleoside triphosphate. In contrast, the substrate does not get phosphorylated, even if the transport protein gets transiently phosphorylated.

Homology Modeling (3D Structure Modeling)

According to Phyre2, there are 20 possible models for MG_476 based on alignment with different templates. We obtained the best model with the highest scoring template (PDB id: c3hhcB) that states interferon-lambda is functionally interferon, but structurally related to il-10 [36]. This sequence has 53% identity with its template, whereas a minimum requirement is 30-40% sequence identity (represented in **Figure 1**). In terms of confidence and coverage, the prediction was 43% accurate. As a result, the MG_476 provided a very reliable structure by satisfying all the validation criteria.

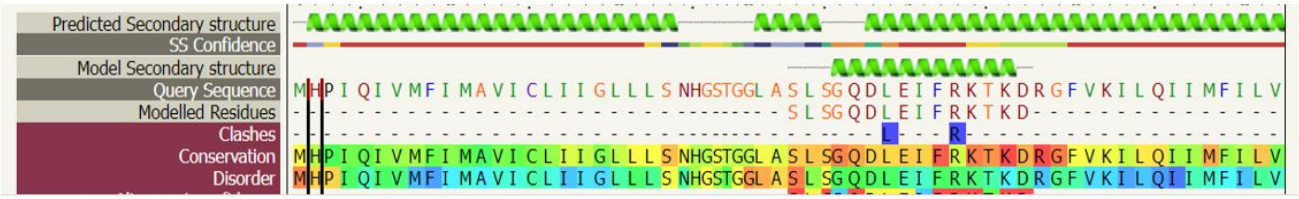

Figure 1. Phyre2 aligned secondary structure between MG_476 and c3hhcB

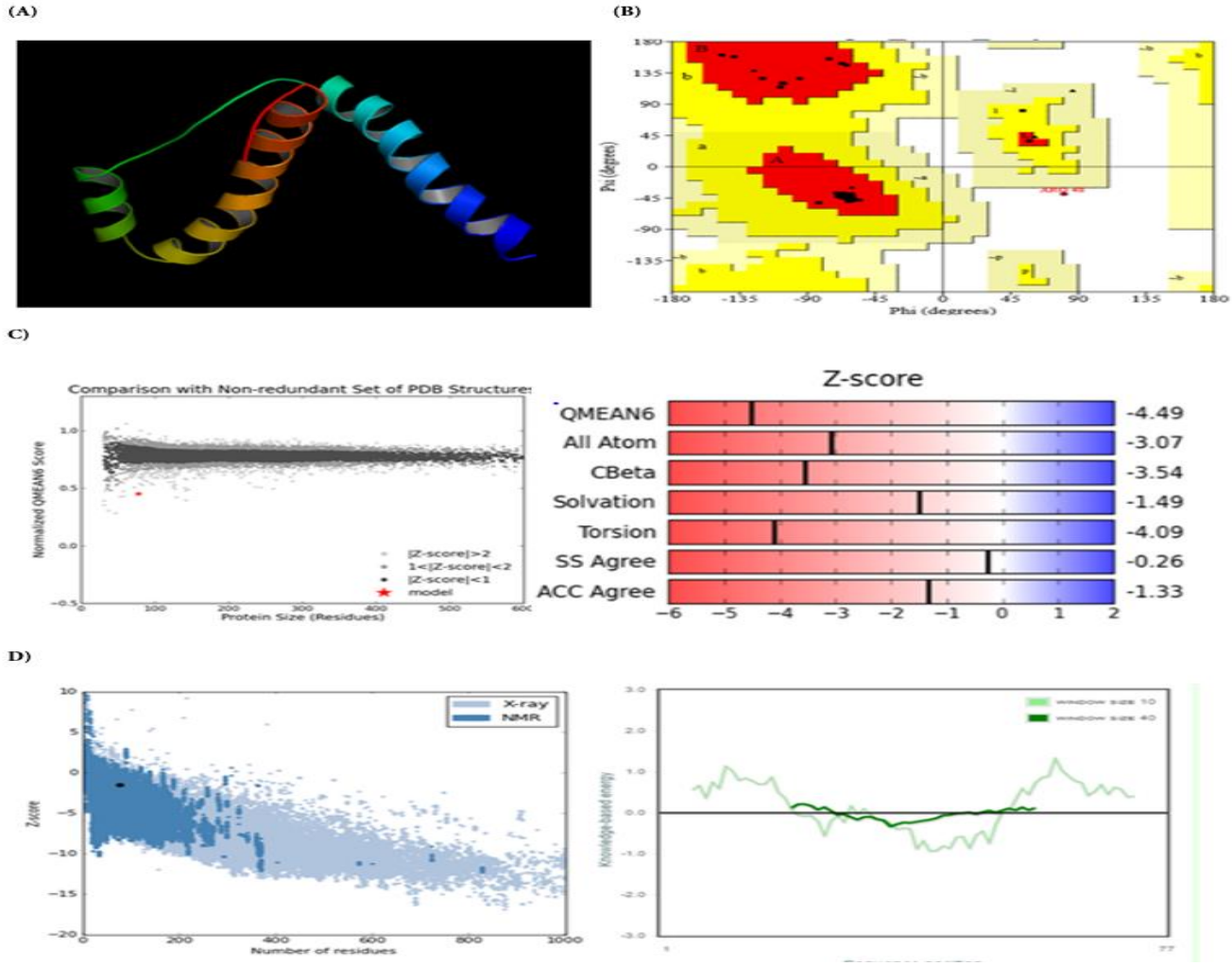

Figure 2. MG_476's refined 3D structure shown in PyMOL (A). Ramachandran plot of MG_476 (B). Embedded QMEAN value & z-score in the dark region indicates the protein of interest (C). ProSA server determination of energy plot and z-score determined by X-ray crystallography or NMR spectroscopy (D)

MG_476 Modification, Validation, and Energy Minimization

Using two steps, ModRefiner refined the protein structure successfully. A main-chain model based on a hydrogen-bonding network on a backbone topology is constructed. A composite force field composed of physics and knowledge is used in the second step to add side chains to the backbone conformation. PYMOL's view of the refined model is depicted in figure 2A. Next, PROCHECK checks the stereochemical quality and accuracy of the predicted protein model by using the Ramachandran plot [37]. The predicted model had 97.0%, 1.5%, 0.0%, and 1.5% residues in the most favorable regions, the additional allowed regions, generously allowed regions, and the disallowed regions, respectively. Ramachandran plots show a good correlation between protein residues and their position in the predicted model based on the percentage distribution. Hence, a reliable prediction model should have a G-factor score beyond -0.50 [39]. Using the present model, dihedral bond G-factor is 0.16, covalent bond G-factor is -2.99, and the overall G-factor is -0.17. The distribution of the main chain bond lengths and bond angles was 61.5% and 69.3%, respectively which was also within limits. The predicted 3D model was evaluated and validated using QMEAN analysis (**Figure 2**), and the Q value for MG_476 was -5.41 (**Figure 2C**).

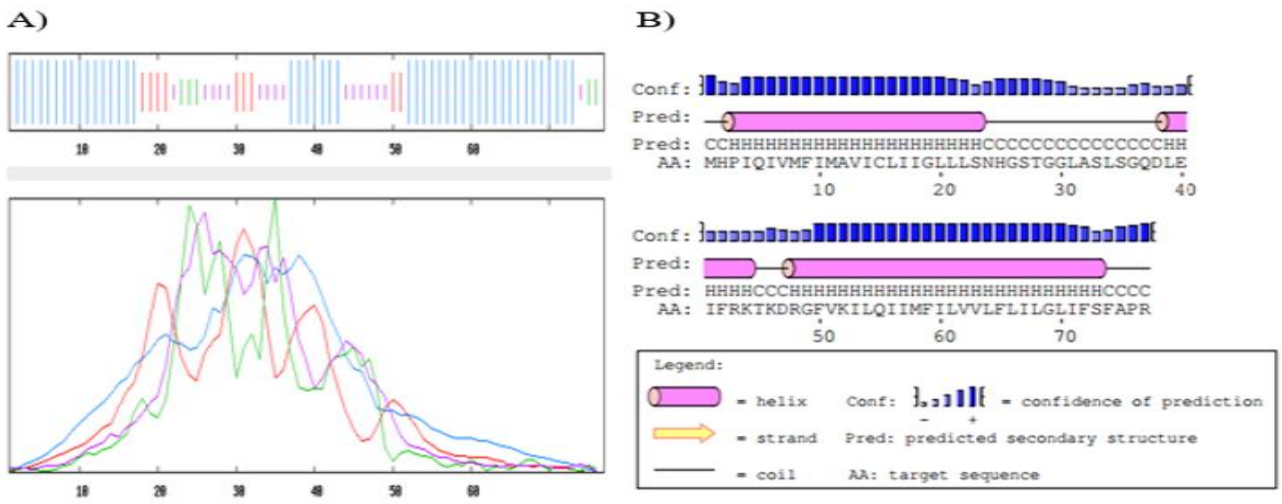

Figure 3. Secondary structure prediction of MG_476 by SOPMA (A). Secondary structure prediction of MG_476 by PSIpred server (B)

Figure 4. CASTp server determined 3D structure of active site with area and volume

When normalized QMEAN score and protein size were plotted, the difference in z-scores for different parameters including C-beta interactions, interactions between all atoms, solvation, torsion, SSE agreement and ACC agreement could be seen. The estimated absolute model quality graph, protein, model quality was in the dark region with an excellent global score (**Figure 2C**). As part of the quality assessment, Verify3D was also used. The analysis result of MG_476 revealed 6.49% of the residues which had an averaged 3D-1D score above 0.2.

From the ERRAT analysis, the overall quality factor of the predicted model was found 59.091 which was below the limit of error value between 95% and 99%. So, the predicted model was fairly a good quality model. A z-score and a plot with residue energies were shown on the web page for the input structure [22]. A model scoring value of -1.55 was obtained, within the acceptable range of -10 to 10, located within the area of the proteins subject to NMR (**Figure 2D**) [38]. It measures the difference between a structure's total energy and an energy distribution derived from random conformations. In the energy plot, knowledge-based energies were plotted as a function of amino acid sequence position to measure the local model quality (**Figure 2D**).

Secondary Structure Determination

At the 8.0 threshold from SOPMA the secondary structure was disclosed with the presence of 45 (58.44%) α-helix, 9 (11.69%) extended sheet, 5 (6.49%) beta-turn, and 18 (23.38%) random coils (**Figure 3A**). Alpha helices were found to be most frequent, followed by a random coil and extended strand. The dominance of the alpha helices and coiled regions showed the high level of conservation and stability of the protein structures [36]. Determination of secondary structure from PSIpred also disclosed that helices were more noticeable, but no beta-sheets have been shown in this structure. The prediction confidence is indicated by blue bars (**Figure 3B**), and most of the helices and coils are relatively associated with better confidence.

Analysis of Active Site

The green region filling space indicates the active site as predicted by the CASTp server in **Figure 4**. Total 12 binding pockets were found in MG_476 protein and the largest pocket is usually considered as standard. So, the largest pocket 12 has an area of 182.3 and a volume 315.4Ao. The residues occurring in this pocket were ALA32, SER33, LEU34, LEU39, ARG43, and LEU54.

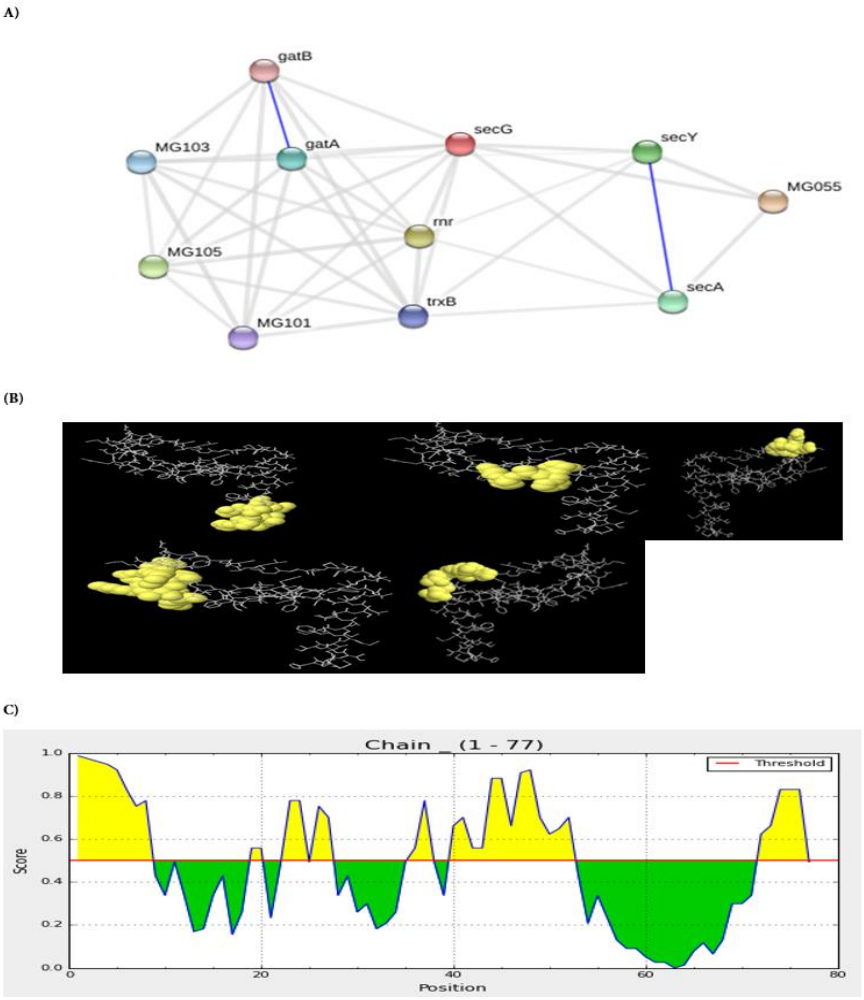

Figure 5. Protein-protein interaction network of MG_476 and predicted functional partners of the protein MG_476 (A). Ellipro discontinuous epitope prediction of MG_476 visualized in Jmol (B). The X and Y axes denote the number of residues and scores, yellow regions indicate potential epitopes above the threshold of 0.5 (C)

Analysis of Interacting Network

STRING 9.05 was used to identify the functional partners of MG_476 from its interacting network presented. STRING forecasts a confidence score and 3D structures of protein and protein domains. Confidence scores were generated based on different parameters, like the neighborhood, co-occurrence, co-expression, and homology [39]. The protein-protein interaction network demonstrated that MG_476 interacts with 10 other proteins of the same strain (**Figure 5A**). The highest confidence was 0.873 and observes with preprotein translocate subunit SecE (MG055).

Other interacting partners were ribonuclease R, 3'-5' exoribonuclease that participates in an essential cell function, preprotein translocase subunit SecY, involved in protein export and interacts with secA and secE, preprotein translocase subunit SecA, part of the Sec protein translocase, aspartyl/glutamyl-tRNA amidotransferase subunit A, three hypothetical protein, thioredoxin-disulfide reductase, andaspartyl/glutamyl-tRNA amidotransferase subunit B with their confidence score.

Determination of Epitopes

T cell (CD8+ and CD4+) epitopes

According to all MHC (A1-B62) supertypes, the NetCTL prediction tool predicted 69 different epitopes from the MG_476 protein sequence. A maximum of four highly promising epitopes were selected based on their high combinatorial scores [40]. A search was conducted to identify which alleles were associated with each of these 4 epitopes within the MHC-I molecules in the IEDB panel that bind peptides. A list of the four best CD8+ T cell epitopes, IMFILIVLVLF epitope interacted with most MHC-1 alleles, particularly HLA-B*15:03(0.3), HLA-B*40:13(1.1), HLA-A*32:01(0.2), HLA-A*32:15(0.6), HLA-B*15:01(0.3), and HLA-A*29:02(0.8). IpMHC immunogenicity score of IMFILVVLF was 0.22834 (**Table 1**). HLApred and MHC class II binding prediction tools have identified five common epitopes on MG_476 protein that bind strongly to HLA-DRB1*01:01, HLADRB1*07:01. Hence, FLILGLIFS, MFILVVLFL, MFIMAVICL, IMAVICLII, and VVLFLILGL can act as potential CD4+ T-cell epitopes and can stimulate an immune response.

Table 1. Predicted Total Scores for CD8+ T-Cell Epitopes with Interacting MHC-1 Alleles

Epitope	NetCTL combined score	Interacting MHC-1 allele with an affinity of IC50<200 (Total score for proteasome, TAP, MHC processing)	IpMHC immunogenicity prediction score
IMFILVVLF	1.24; A24 0.99; B58 1.31; B62	HLA-B*15:03(0.3) HLA-B*40:13(1.1) HLA-A*32:01(0.2) HLA-A*32:15(0.6) HLA-B*15:01(0.3) HLA-A*29:02(0.8)	0.22834
LILGLIFSF	1.03; B58 1.16; B62	HLA-A*02:06(1.7) HLA-A*32:01(0.8) HLA-B*15:01(1.1) HLA-B*15:03(2.8)	0.14712
FRKTKDRGF	1.51; B8 1.22; B27	HLA-C*07:02(0.5) HLA-C*06:02(1.1) HLA-B*27:20(1.3) HLA-C*07:01(2.6)	-0.15658
ILQIIMFIL	1.03; A2 1.21; B8	HLA-A*02:02(0.9) HLA-A*02:12(1.3) HLA-A*02:02(0.9)	0.23718

Table 2. Estimated B-Cell Epitope by ABCpred

Sequence	Start position	Score
RGFVKILQIIMFILWLFLI	48	0.85
EIFRKTKDRGFVKILQIIMF	40	0.71
GSTGGLASLSGQDLEIFRKT	26	0.63
IQIVMFIMAVICLIIGLLLS	4	0.61
MAVICLIIGLLSNHGSTGG	11	0.52

Table 3. Discontinuous Epitopes Prediction of MG_476 Protein by Ellipro

No	Residues	NoR	S
1	_:M1,_:H2,_:P3,_:I4,_:Q5,_:I6,_:V7,_:M8	8	0.894
2	_:R43,_:T45,_:K46,_:D47,_;R48,_:G49,_:F50,_:V51,_:K52	9	0.734
3	_:S73,_:F74,_:A75_:P76,_:R77	5	0.73
4	_:S35,_:G36,_:Q37,_:E40,_:I41,_:F42	6	0.626
5	_;G19,_;L20,_;S23,_;N24,_:H25,_;G26,_:S27,_:T28	8	0.26

Note. NoR: Number of residues & S. Score

B cell continuous and discontinuous epitope determination

Bceppred can predict linear B-cell epitopes with 58.7% accuracy using flexibility, hydrophobicity, polarity, surface properties combined at a threshold of 2.38, and ten epitopes were found by combining seven physicochemical properties. Five epitopes were found by using the ABCpred server (represented in **Table 2**) and predicted B cell epitopes were ranked according to their score obtained by a trained recurrent neural network. Higher scores indicate a greater likelihood that peptide will act as an epitope.

Five discontinuous epitopes were predicted by Ellipro with residual specification and corresponding scores summarized in **Table 3**. **Figure 5B** shows the ball-and-stick models of the predicted epitopes. The epitope residues are yellow while the antibody chains are white in color. The 2D score chart denotes in **Figure 5C**.

MG_476 is not known to have any known functions. An integrated bioinformatics workflow was used to perform functional annotation on MG_476 utilizing several tools and databases. Our search for Mg_476 conserved domains and possible functions utilized three web tools. The SecGsuper family domains of MG_476 are classified as protein with unknown function which have been predicted by Pfam, NCBI-CDD, and InterProScan [41]. MG_476 has been further characterized using comparative genomics after functional annotation had been done. A proteome search, the estimation of essentiality, and involvement with metabolic pathways were involved. In order to determine whether MG_476 has any human homologs, a BLASTP search against the human proteome was conducted first. MG_476 is a distinct protein of *M. genitalium* and does not show any homology with human proteins. Since these proteins contain no homology to human proteins, they can be used effectively as drug targets [42].

Active sites of this protein were determined by the CASTp program, and it recognized 12 binding pockets within the protein [25]. As the protein is a hypothetical one, no known ligands have been found for this study. T and B cell epitopes were designed for an integrative in-silico vaccine/drug design. In antiviral immunity, CD8+ and CD4+ T cells are crucial to its effectiveness, and they are also important in antigen-mediated clonal expansion of B cells. Based on the results of this study, some possible epitopes for molecules in class I and class II MHC were identified. Among the four potential CD8+ T cell epitopes tested, ILQIIMFIL scored (0.23718) the highest in terms of immunogenicity. The discontinuous epitope with the highest score in this study was MHPIOIVM (1-8) with eight residues in length.

CONCLUSION

3D structure of MG_476 was predicted using a homology-based model. Following validation by different structural assessment methods, a fairly good quality model emerged. Next, we predicted B cell and T cell epitopes that were not identical to those in humans. An effective drug or vaccine might be developed based on these protein epitopes. Although not exclusively applicable to drug discovery, such a method might be useful for defining chemical targets for other clinically important pathogens. Further studies are in progress to validate experimentally the data found from this study.

Author contributions: All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: This study was supported by the Department of Biotechnology and Genetic Engineering and Department of Biochemistry and Molecular Biology, Jahangirnagar University, Savar, Bangladesh.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

1. Vetrivel U, Subramanian G, Dorairaj S. A novel in silico approach to identify potential therapeutic targets in human bacterial pathogens. Hugo J. 2011;(1-4):25-34. doi:10.1007/s11568-011-9152-7 PMid:23205162 PMCid:PMC3238024
2. Saikolappan S, Sasindran SJ, Yu HD, Baseman JB, Dhandayuthapani S. The mycoplasma genitalium MG_454 gene product resists killing by organic hydroperoxides. J. Bacteriol. Res. 2009;191:6675-82. doi:10.1128/JB.01066-08 PMid:19717589 PMCid: PMC2795314
3. Butt AM, Batool M, Tong Y. Homology modeling, comparative genomics and functional annotation of Mycoplasma genitalium hypothetical protein MG_237. Bioinformation. 2011;7:299. doi:10.6026/007/97320630007299 PMid:22355225 PMCid:PMC3280499
4. Paul S, Saha M, Bhoumik NC, Talukdar SN. In silico structural and functional annotation of mycoplasma genitalium hypothetical protein MG_377. Int J Bioautomation. 2015;19.
5. Jensen JS. Mycoplasma genitalium: The aetiological agent of urethritis and other sexually transmitted diseases. J Eur Acad Dermatol Venereol. 2004;18:1-1. doi:10.1111/j.1468-3083.2004.00923.x PMid:14678525
6. Ijaq J, Chandrasekharan M, Poddar R, Bethi N, Sundararajan VS. Annotation and curation of uncharacterized proteins-challenges. Front Genet. 2015;6:119. doi:10.3389/fgene.2015.00119 PMid: 25873935 PMCid:PMC4379932
7. Sanmukh SG, Paunikar WN, Ghosh TK, Chakrabarti T. Structure and function predictions of hypothetical proteins in vibrio phages. IJBB. 2010;4:161-75.
8. Gazi MA, Kibria MG, Mahfuz M, et al. Functional, structural and epitopic prediction of hypothetical proteins of Mycobacterium tuberculosis H37Rv: An in silico approach for prioritizing the targets. Gene. 2016;591:442-55. doi:10.1016/j.gene.2016.06.057 PMid: 27374154
9. Falda M, Toppo S, Pescarolo A, et al. Argot2: A large scale function prediction tool relying on semantic similarity of weighted gene ontology terms. BMC Bioinform. 2012;13:1-9. doi:10.1186/1471-2105-13-S4-S14 PMid:22536960 PMCid:PMC3314586
10. Lubec G, Afjehi-Sadat L, Yang JW, John JP. Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol. 2005;77:90-127. doi:10.1016/j.pneurobio.2005.10.001 PMid:16271823
11. Rambabu R, Peri S, Allam A. Computational analysis and function prediction of a hypothetical protein 1RW0. Int J Comp Bioinform In Silico. 2012:58-62.
12. Marchler-Bauer A, Anderson JB, Derbyshire MK, et al. CDD: A conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007;35:237-40. doi:10.1093/nar/gkl951 PMid:17135202 PMCid: PMC1751546
13. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. Nucleic Acids Res. 2010;38:211-22. doi:10.1093/nar/gkp985 PMid:19920124 PMCid: PMC2808889
14. Priya VS, Muddapur UM, Mehta M. Function and structure prediction of Rv2004c, a hypothetical protein from M. tuberculosis. IJRSEIT. 2013;2(9):4467-77.
15. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. Proteomics Protoc Handbook. 2005;571-607. doi:10.1385/1-59259-890-0:571
16. Bhasin M, Garg A, Raghava GP. PSLpred: Prediction of subcellular localization of bacterial proteins. Bioinform. 2005;21:2522-4. doi:10.1093/bioinformatics/bti309 PMid:15699023
17. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. 2006;15:1550-6. doi:10.1110/ps.062153506 PMid: 16672240 PMCid:PMC2242549
18. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10:845-58. doi:10.1038/nprot.2015.053 PMid:25950237 PMCid: PMC5298202

19. Gupta CL, Akhtar S, Bajpaib P, Kandpal KN, Desai GS, Tiwari AK. Computational modeling and validation studies of 3-D structure of neuraminidase protein of H1N1 influenza A virus and subsequent in silico elucidation of piceid analogues as its potent inhibitors. EXCLI J. 2013;12:215.
20. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009;37:510-4. doi:10.1093/nar/gkp322 PMid:19429685 PMCid:PMC2703985
21. Colovos C, Yeates TO. Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Sci.1993;2:1511-9. doi:10.1002/pro.5560020916 PMid: 8401235 PMCid:PMC2142462
22. Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007;35:407-10. doi:10.1093/nar/gkm290 PMid: 17517781 PMCid:PMC1933241
23. Geourjon C, Deleage G. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinform. 1995;11:681-4. doi:10.1093/bioinformatics/11.6.681 PMid:8808585
24. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. JMB. 1999;292:195-202. doi:10.1006/jmbi.1999.3091 PMid: 10493868
25. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res. 2006;34:116-8. doi:10.1093/nar/gkl282 PMid:16844972 PMCid:PMC1538779
26. Von Mering C, Jensen LJ, Snel B, et al. STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids res. 2005;33:433-7. doi:10.1093/nar/gki005 PMid:15608232 PMCid:PMC539959
27. Zhao XM, Chen L, Aihara K. Protein function prediction with high-throughput data. Amino Acids. 2008;35:517-30. doi:10.1007/s00726-008-0077-y PMid:18427717
28. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinform. 2007;8:1-2. doi:10.1186/1471-2105-8-424 PMid:17973982 PMCid:PMC2194739
29. Buus S, Lauemøller SL, Worning P, et al. Sensitive quantitative predictions of peptide-MHC binding by a 'query by committee' artificial neural network approach. Tissue Antigens. 2003;62:378-84. doi:10.1034/j.1399-0039.2003.00112.x PMid:14617044
30. Kobayashi H, Wood M, Song Y, Appella E, Celis E. Defining promiscuous MHC class II helper T-cell epitopes for the HER2/neu tumor antigen. Cancer Res. 2000;60:5228-36.
31. Saha S, Raghava GP. Predicting virulence factors of immunological interest. Immunoinform. 2007:407-15. doi:10.1007/978-1-60327-118-9_31 PMid:18450019
32. Ponomarenko J, Bui HH, Li W, et al. ElliPro: A new structure-based tool for the prediction of antibody epitopes. BMC Bioinforms. 2008;9:1-8. doi:10.1186/1471-2105-9-514 PMid:19055730 PMCid: PMC2607291
33. Zanotti G, Cendron L. Structural and functional aspects of the Helicobacter pylori secretum. WJG. 2014;20:1402. doi:10.3748/wjg.v20.i6.1402 PMid:24587618 PMCid: PMC3925851
34. Zhang R, Ou HY, Zhang CT. DEG: A database of essential genes. Nucleic Acids Res. 2004;32:271-2. doi:10.1093/nar/gkh024 PMid:14681410 PMCid: PMC308758
35. Silva PFF, Novaes E, Pereira M, Soares CMA, Borges CL, Salem-Isacc SM. In silico characterization of hypothetical proteins from paracoccidioides lutzii. Genet Mol Res. 2015;14(4):17416-25. doi:10.4238/2015.December.21.11 PMid:26782383
36. Pilley HH. In-silico prediction of structural and functional aspects of a hypothetical protein of capnocytophaga canimorsus Cc5. J Adv Bioinfo Appl Res. 2002;2:206-10.
37. Sharon FB, Daniel RR. Homology modeling of nitrogenase iron protein of nitrogen fixing Actinomycete Arthrobacter sp. IJCA. 2013;61. doi:10.5120/9891-4457
38. Chhabra G, Sharma P, Anant A, et al. Identification and modeling of a drug target for clostridium perfringens SM101. Bioinformation. 2010;4:278. doi:10.6026/97320630004278 PMid:20978600 PMCid:PMC2957761
39. Hasan A, Mazumder HH, Khan A, Hossain MU, Chowdhury HK. Molecular characterization of legionellosis drug target candidate enzyme phosphoglucosamine mutase from Legionella pneumophila (strain Paris): An in silico approach. Genomics Inform. 2014;12:268. doi:10.5808/GI.2014.12.4.268 PMid:25705169 PMCid:PMC4330265
40. Shawan MM, AlMahmud H, Hasan MM, Parvin A, Rahman MN, Rahman SB. In silico modeling and immunoinformatics probing disclose the epitope based peptidevaccine against zika virus envelope glycoprotein. IJPBR. 2014;2:44. doi:10.30750/ijpbr.2.4.10
41. Butt AM, Batool M, Tong Y. Homology modeling, comparative genomics and functional annotation of Mycoplasma genitalium hypothetical protein MG_237. Bioinformation. 2011;7:299. doi:10.6026/007/97320630007299 PMid:22355225 PMCid:PMC3280499

42. Chhabra G, Sharma P, Anant A, et al. Identification and modeling of a drug target for clostridium perfringens SM101. Bioinformation. 2010;4:278. doi:10.6026/97320630004278 PMid:20978600 PMCid:PMC2957761

