ORIGINAL ARTICLE

Computational evaluation of anticipated PE_PGRS39 protein involvement in host–pathogen interplay and its integration into vaccine development

Khyati Patni[1] · Preeti Agarwal[1,2] · Ajit Kumar[1,2] · Laxman S. Meena[1,2]

Abstract

Mycobacterium tuberculosis causes more than 1 million deaths every year, which is higher than any other bacterial pathogen. Its success depends on its interaction with the host and its ability to regulate the host's immune system for its own survival. *Mycobacterium tuberculosis* H_{37}Rv (Mtb) proteome consists of unique PE_PGRS family proteins, which present a significant role in bacterial pathogenesis over the past years. Earlier evidence suggests that some PE_PGRS proteins display fibronectin-binding activity. In this manuscript, computational characterization of the PE_PGRS39 protein has indicated something peculiar about this protein. Investigation showed that PE_PGRS39 is an extracellular protein that, instead of acting as fibronectin-binding protein, might mimic fibronectin which binds to alpha-5 beta-1 ($\alpha5\beta1$) integrin. PE_PGRS39 protein additionally turned into proven pieces of evidence to have motifs such as DXXG and GGXGXD and PXXP that bind with guanosine triphosphate (GTP), calcium, and host Src homology 3 (SH3) domains, respectively, in conjunction with RGD-integrin binding. These interactions designate the direct role of PE_PGRS39 in bacterial pathogenesis via cell adhesion and signaling. Additionally, the analysis showed that PE_PGRS39 is an antigenic protein and epitope prediction provided functional regions of the protein that trigger a cellular immune response facilitated by T or B cells. Further, an experimental analysis could also open up new avenues for developing novel drugs by targeting signaling motifs or novel vaccines using functional epitopes that could evoke an immune response in the host.

Keywords PE_PGRS · Host–pathogen interaction · RGD motif · Integrin · Fibronectin · Src Homology 3

Abbreviations

TB	Tuberculosis
Mtb	*Mycobacterium tuberculosis* H_{37}RV
WHO	World Health Organisation
MDR-TB	Multidrug-resistant TB
XDR-TB	Extremely drug-resistant TB
TDR-TB	Totally drug-resistant TB
RR-TB	Rifampicin-resistant TB
BCG	Bacille Calmette–Gurein
SH3	Src homology 3
GTP	Guanosine triphosphate
$\alpha5\beta1$	Alpha-5 beta-1 integrins

Preeti Agarwal and Ajit Kumar have equally contributed.

✉ Laxman S. Meena
meena@igib.res.in; laxmansm72@yahoo.com

[1] Academy of Scientific and Innovative Research (AcSIR), CSIR-Institute of Genomics and Integrative Biology, Mall Road, Delhi 110007, India

[2] Academy of Scientific and Innovative Research (AcSIR), CSIR-HRDC, Ghaziabad, Uttar Pradesh 201 002, India

Introduction

Tuberculosis (TB) is a highly transmittable disease carried by one of the most pathogenic bacteria known as *Mycobacterium tuberculosis*. Its strain H37Rv is the most studied strain in research laboratories. TB spread by infectious droplets containing *Mycobacterium tuberculosis* H_{37}Rv (Mtb) from one sick person to another, and it remains a significant contributor to global health challenges, prompting comprehensive studies in infectious disease (Rajni and Meena 2010; Fauci 2001). According to the World Health Organisation (WHO) 2020 study, an estimated 10 million people worldwide have fallen ill with TB in 2019, with 5.6 million males, 3.2 million females, and 1.2 million children (WHO Report 2020). Antibiotic-resistant strains, including totally drug-resistant TB (TDR-TB), rifampicin-resistant TB (RR-TB),

multidrug-resistant TB (MDR-TB), and extensively drug-resistant TB (XDR-TB) (Meena and Rajni 2010), are the most imperative factors associated with the development of TB worldwide. In 2019, 206,030 people worldwide were detected with MDR/RR-TB, which comes out to be a 10% rise from 186,883 in 2018 (WHO Report 2020). Data showed that TB rates are rising alarmingly, but Bacille Calmette–Guérin (BCG) is still the solely certified vaccine for preventing TB (Montagnani et al. 2014). Thus, there is a great need for the production of potential therapeutics for TB.

Despite lots of studies on Mtb, there are still many PE/PPE family proteins that have not been studied yet (Via et al. 2015). In this manuscript, we worked on the subgroup of the PE/PPE protein family. These gene families of proline-glutamic (PE) and proline-proline-glutamic acid (PPE) extends to approximately 10% of total Mtb genome coding potential (Singh et al. 2016; Meena 2019). The preserved N-terminus domains of both PE and PPE proteins are about 110 aa and 180 aa, respectively, and are divided into separate subclasses based on their variable C-terminus domain (Sampson 2011; Meena 2015). The PE_PGRS protein, one of the PE/PPE family subgroups, has retained the N-terminus PE domain, which shows similarity to the PE family. Almost all PE_PGRS proteins are known to have consensus sequence GGXGG and demonstrated to contain GTP and calcium-binding motifs (Bachhawat and Singh 2007). Some PE_PGRS proteins are well known for their contribution to the replication and survival of Mtb within macrophages and are likely to propose the function as a source of antigenic variability to evade host immune responses (Mahadevan 1998). Other observations indicate that PE_PGRS proteins may be the components of the cell surface that could facilitate the binding of cell adhesion molecules, affect the bacterial cell structure or interfere with host immune responses by inhibiting antigen processing of antigen (Bachhawat and Singh 2007).

Evidence showed that some PE_PGRS proteins possess fibronectin-binding activity, which plays a potential part in cell adherence and immune invasion (Meena and Meena 2015). Apart from the mycobacterium, it was shown that in *Staphylococcus aureus* (*S. aureus*), fibronectin-binding significantly increases the phagocytic activity of macrophages on binding with alpha-5 and beta-1 chains, which were associated with the cytoskeleton (Kevlani and Meena 2017). Fibronectin consists of several domains classified as FnI to FnIII that show binding specificities for specific host membrane receptors such as integrins. Shreds of evidence have shown that the two significant domains within fibronectin that interact with beta-1 integrins are FnIII-9 and FnIII-10 (Speziale et al. 2019), but the most critical element for binding energy is the aspartate residue of the RGD motif within FnIII-10 (Aota et al. 1994). Integrins represent 24

αβ heterodimeric members of a broad cell adhesion class, and one-third of integrins have been shown to bind to the RGD sequence (Fenn et al. 2020). These integrins not only facilitate the attachment of cells to the extracellular matrix or conduct cell–cell interactions (Zimmermann et al. 2010) but are also equally responsible for migration and signal transduction (Kevlani and Meena 2017).

Understanding the interaction between host and pathogen is necessary to eliminate any disease. Many pathogens mimic human sequences to interact with host components and use host machinery for their survival. Thus, in this manuscript, we have analyzed how previously uncharacterized PE_PGRS39 protein might be responsible for Mtb pathogenesis by mimicry with host or other ways. Through bioinformatics tools, we have structurally and functionally characterized PE_PGRS39. Additionally, we have provided an insight into the integration of this protein for vaccine development.

Materials and methods

Protein sequence retrieval and analysis of its physiochemical properties

For the recovery of the protein sequence of the gene Rv2340c (PE_PGRS39) of Mtb, we used Mycobrowser (Beg et al. 2018). Mycobrowser database provides genomic and proteomic information related to mycobacteria. It provides information about protein sequence, function, drug interaction, and orthologous relations (Lew et al. 2011). It is also possible to recover genes and their protein sequences from the NCBI database, Uniport, and KEGG. The physiochemical properties such as theoretical pI, molecular weight, amino acid composition, atomic composition, aliphatic index, extinction coefficient, instability index, and grand average of hydropathicity (GRAVY) of the PE_PGRS39 were studied using the ProtParam tool, which takes input as a protein sequence (Godfrey et al. 2005). Further, we found functional motifs computationally by motif scan and manually by surfing through the NCBI (National Center for Biotechnology Information) database. Motif scan includes Prosite, Pfam, and HAMAP profiles.

Sub-cellular localization prediction

The subcellular localization is performed using CELLO2GO. CELLO2GO indicates four subcellular localizations in archaea and gram-positive bacteria utilizing multiple integrated machine-learning classifiers. Specified localizations for gram-positive bacteria are the cytoplasm, the extracellular space, the cytoplasmic membrane, and the

cell wall (Yu et al. 2014). To confirm CELLO2GO results, LocTree3 sever was used (Goldberg et al. 2012).

Functional analysis

(a) Functional category prediction: We used the VICMpred server to predict in which functional category PE_PGRS39 is involved, providing input in FASTA format and standard prediction approach as a pattern. To predict functional groups, VICMpred proposes a support vector machine-based method. Broadly, this server offers data for four available groups, i.e., metabolism molecular, virulence factors, the cellular process, and information molecules (Sharma et al. 2019). The precision of VICMpred is 70.75% (Saha and Raghava 2006).

(b) Molecular activity prediction: Molecular activity of PE_PGRS39 was predicted using the COFACTOR online server (Agarwal et al. 2020). COFACTOR takes either sequence or 3D structure of the model and threads the query through the BioLiP protein function database, which searches global and local structure matches to identify molecular activity based on functional sites and homologies. It annotates the biological function of a given protein in terms of gene ontology (GO). Protein–protein interaction inferrals can obtain these GO names from STRING or by sequence and sequence-profile alignments from UniProt-GOA. This server uses the best available homology templates to derive gene ontology (Roy et al. 2012).

Functional disorder regions and phosphorylation site prediction

We used a disordered phosphorylation prediction (DEPP) server to predict phosphorylation sites in PE_PGRS39 protein. This server indicates the presence of threonine, serine, and tyrosine phosphorylation sites in proteins. It is well known that disordered protein regions frequently contain short linear peptide motifs, such as binding sites for the SH3 domain and targeting signals essential for protein function (Linding et al. 2003). Thus, to identify those disorder regions, we used PONDR® (Predictor of Natural Disordered Region). PONDR® takes protein sequence as an input that searches for the size of 9–21 amino acids. It is typically a neural network based on hydropathy, particular amino acid composition, or complexity of a given protein sequence (Delamain et al. 2019). There are eight forms of PONDER, but we used PONDR VL-XT because it is the most reliable one (Agarwal et al. 2020).

Structure prediction and validation

(a) Secondary structure prediction: A user-friendly server named PSIPRED was used to evaluate the secondary structure of PE_PGRS39. PSIPRED is a secondary structure prediction method based on PSI-blast that predicts beta-sheets, alpha helices, and coils from the primary sequence successfully (Monu 2016).

(b) Tertiary structure prediction: To obtain the tertiary structure of PE PGRS39, Raptor-X, I-TASSER, and Phyre2 were used. Lastly, for further experimentation, the best model predicted by I-TASSER was chosen. I-TASSER is an interactive framework designed for template-based predictions of protein structure and function. For multiple threading alignments and iterative structural assembly simulations, I-TASSER first generates full-length atomic structural models using the highest significant template, followed by atomic-level structure refinement starting from the target protein amino acid sequence (Zhang 2008). For multiple threading alignments, I-TASSER uses LOMETS (Roy et al. 2010).

(c) Tertiary structure refinement: To generate structure models, protein structure prediction tools rely on the similarity between the target and available template structures in different databases. Therefore, improving model structures based on templates beyond accuracy has been a primary concern among the bioinformatics community. The GalaxyRefine webserver was used to enhance both global and local structure consistency. This tool showed optimum results when the structure was predicted via servers such as I-TASSER and ROSETTA (Kathwate 2020; Heo et al. 2013).

(d) Structure validation: The best-refined structure of protein PE_PGRS39 was authenticated through SAVES v5.0 (structure analysis and verification server version 5) and the ProSA server. SAVES v5.0 is a complete package of five programs that test the overall consistency of a protein structure. Out of five programs, we used VERIFY-3D to test the 3-D sequence profile for protein models and PROCHECK to validate structure through the Ramachandran plot. PROCHECK checks the stereochemical nature of a protein structure by analyzing residue-by-residue geometry and overall structure geometry (Beg et al. 2018).

Protein interaction and docking

(a) Protein–protein docking: To show protein–protein interactions, we used the suite of docking programs called pyDock (Cheng et al. 2007). pyDock is a protein–protein docking, FFT-based algorithm that uses electrostatics, desolvation energy, and a small contri-

bution of van der Waals contribution to scoring rigid-body docking poses (pyDockSER).

The docking solutions evaluated for rigid-body docking with the pyDockSER module optimized are given by the equation: $BE = EE + DE + W.VE$, where BE is equal to total binding energy, EE is total electrostatics binding energy (lies between 1.0 and -1.0 kcal/mol), DE is desolvation energy upon binding, VE is van der Waals binding energy(set at 1.0 kcal/mol to avoid much noise from the docking of rigid body surfaces), and W refers to weight taken at default equal to 0.1 (Degryse et al. 2008). During simulations, biological restrictions were used, which enables the docking at specific predicted PE_PGRS39 protein regions. Figures were visualized and drawn with PyMol.

(b) Protein–ligand docking: Molecular docking for each ligand was completed using the AutoDock Vina (Meena and Meena 2019) and COACH-D (Wu et al. 2018). We used AutoDock tools, (1) to add Kollman charges and hydrogen atoms (polar) in the protein framework, (2) to adjust the size and the grid box center, and (3) to prepare the final input of the PDBQT file. The grid size was set at $40 \times 36 \times 66$ for x, y, and z points, and the grid center was selected at 73.052, 78.820, and 94.354 for x, y, and z dimensions, respectively. The final configurated file was opened in the notepad to run Auto-Dock Vina. AutoDock Vina scoring function was used to calculate ligand–protein binding affinities, predicted as negative Gibbs free energy (ΔG) scores (kcal/mol). Visualization of the docked structure was done using PyMol and Discovery Studio Biovia, which depicts the sizes and locations of binding sites, electrostatic interactions, H-bond interactions, and interaction radii using bonding distances. Lastly, the result was selected based on the binding energy of the ligand.

Mutation analysis

We used EASE-MM, which stands for evolutionary, amino acid, and structural encodings with multiple models, to find the outcome of the mutation on potential binding sequences. EASE-MM supports the prediction of missense mutation specified as amino acid substitutions. EASE-MM takes protein sequence in a specified format and combines multiple specialized machine learning techniques to predict $\Delta\Delta G$ (Folkman et al. 2016).

Prediction of antigenicity and potential epitopes

(a) Antigenicity: We predicted the protein's antigenic nature using the VaxiJen sever, setting a threshold of 0.4. This server is based on the physicochemical properties of the protein (Kardani et al. 2020).

(b) Prediction of B-cell epitopes: To predict the B-cell binding region within the PE_ PGRS39 protein sequence, we used ABCpred. A threshold of 0.5 and 16mer amino acids default window were selected (Meena and Meena 2016).

(c) Prediction of T-cell epitopes: We used the HLAPred tool to predict MHC class-II binding region in the antigenic protein sequence of PE_PGRS39. Multiple DR-β1 (DRB) alleles like HLA-DRB1*0101, HLA-DRB1*0102, HLA-DRB1*0301, and HLA-DRB1*0305 were selected at default prediction parameters (de la Caridad Addine Ramírez et al. 2013).

Results

Recovery of the PE_PGRS39 protein sequence

We used mycobrowser to obtain the protein sequence of the Rv2340c gene in FASTA format. Results showed that the retrieved protein belongs to the PE_PGRS family protein called PE_PGRS39, a glycine-rich protein containing 22.0% glycine. This protein is 413 amino acids long, with an isoelectric point at 6.5888, and has a molecular mass of 38722.3 Da. Through the ProtParam tool, we found that the grand average of hydropathicity (GRAVY) is 0.086. The protein sequence contains 28 negatively charged residues and 23 positively charged residues (Table 1). ProtParam also determined the instability index's value as 22.72, which categorizes PE_PGRS39 as a stable protein.

Through MotifScan, we found that PE_PGRS39 protein contains an RGD cell attachment motif that consists of the amino acids Arg-Gly-Asp (Speziale et al. 2019). PE_PGRS39 also confirmed the presence of conserved GTP and calcium-binding motifs, which is represented by GXXXXGK/DXXG/NKXD and GGXGXD/NXUX, respectively, where X represents any amino acid, and U is a non-polar or hydrophobic large residue that binds to Ca^{2+}. Along with the above motifs, PE_PGRS39 has a PXXP motif which binds with SH3-containing proteins (Fig. 1). In particular, some PXXP motifs are unique to Mtb (Chandra et al. 2004).

Protein subcellular localization prediction

PE_PGRS39 was predicted to be an extracellular protein with a cumulative score of 3.448, which was more significant than the other expected cell wall, plasma membrane, and cytoplasm divisions (Table 2). The localization probability chart is represented in Fig. 2.

Table 1 Physiochemical properties of protein

Physiochemical properties	Values
Number of amino acids	413
Total number of negatively charged residues (Asp + Glu)	28
Total number of positively charged residues (Arg + Lys)	23
Molecular weight	38722.33
Theoretical PI	6.10
Extinction coefficients	24,980/M cm, at 280 nm measured in water
Instability index	22.72
Aliphatic index	70.48
Total number of atoms	5339
Grand average of hydropathicity (GRAVY)	− 0.086

Fig. 1 Systematic representation of different motifs found and their location in PE_PGRS39 protein: different coloured highlights shows different motifs, calcium-binding (red) GTP binding (yellow), guanine base recognition site (violet), integrin binding (purple), SH3 domain binding (cyan)

>Mycobacterium tuberculosis H37Rv|Rv2340c|PE_PGRS39

MSHVTAAPNVLAASAGELAAIGSTMRAANAAAAAPTAGVLAAGGDDVSAGIAALF
GARAQAYQAISAQAALFHDRFVQILQEGAAAYAMAEAANALPLQKAQGVVSELAQ
DRTGGTGTGQSRGAGGFGGVGQAGGKGWDGGPIGNGQVGEQHGAGQLGSTDGNP
GVAGAAHGSGVSASHGSGATGAAGVADPGGSGAGVGSAAGNGTGAGSADAVGGA
GTGRDIVGSVRGDGGVGMASGDGGLSTGAAGASAEGGLMPGFGGAPWVGGHWGL
GGEGHSGAIGGVGEQVAPAVATAPAVSPATTSAVAAESGSTPATKAQAMHATTNPG
NAAHQGNPADPGNSARRADGGRDEQLLLLPLTSLRGLRHTLKKLSGLRARNGLLTA
SGDNASGSGRPWDRDQLLRALGLRPPGHE

■ GGxGXD/NxUx (Calcium binding motif)

□ GXXXXGK/DXXG/NKXD (GTP binding motif)

■ (T/G)(C/S)A (Buttreness the guanine base recognition site)

■ RGD (integrin binding motif)

■ PXXP (Protein containing SH3 domain binding motif)
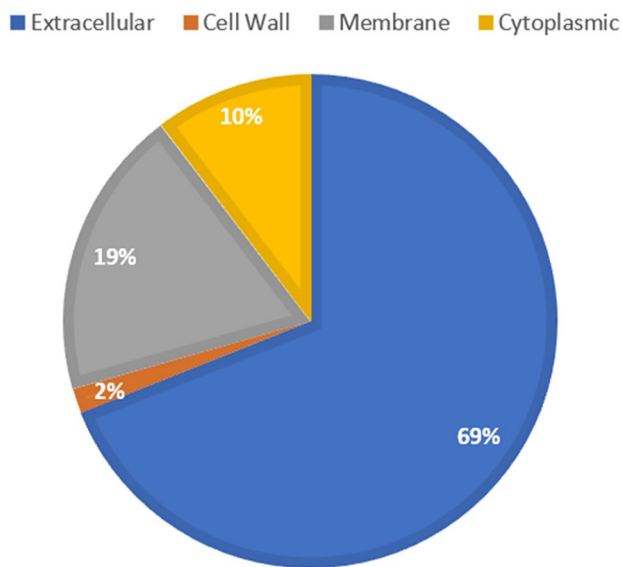
Table 2 Localization prediction of protein PE_PGRS39

Cello predictor for gram + model	
Different functional class	Score
Extracellular	3.448
Cell wall	0.081
Membrane	0.953
Cytoplasmic	0.518

Function prediction

(a) Functional category prediction: The VICMpred server was used to predict which of the functional groups PE_PGRS39 belongs to. VICMpred showed that the PE_PGRS39 involve in the cellular process with a score of − 0.952733511 (Table 3).

(b) Molecular function prediction: We used COFACTOR to predict the probable molecular activity of PE_PGRS39. It is expected that this protein establishes GO terms of structural molecule activity with CscoreGO as 0.31 that contributes to the structural integrity or its assembly within or outside a cell. It also shows functions like receptor binding activity, complex macromolecular binding, heterocyclic compounds binding, etc. (Table 4). Here, CscoreGO is the confidence score for standard GO terms on the COFACTOR server. CscoreGO values are within the range of 0–1, where a higher value infers greater confidence in predicting protein function. Although, CscoreGO value of 0.31 has low confidence level, we have still considered it to have a fair idea regarding protein's possible function.

(c) Phosphorylation site and functional disorder regions prediction: The PONDR server indicated active disorder regions within the PE_PGRS39 protein sequence.

LOCALIZATION PROBABILITY

■ Extracellular ■ Cell Wall ■ Membrane ■ Cytoplasmic

Fig. 2 Pie chart representation of localization probability prediction of protein PE_PGRS39 shows the highest per cent of extracellular represented by blue colour

Table 3 Function category prediction of protein PE_PGRS39

Functional class prediction by VICM	
Different functional class	Score
Cellular process	− 0.95273351
Information molecule	− 1.5166671
Metabolism	− 4.3987188
Virulence factors	− 1.3315684

Table 4 Structure-based molecular function prediction by COFACTOR server

GO term	CscoreGO	Name
GO:0005198	0.31	Structural molecule activity
GO:0005201	0.18	Extracellular matrix structural constituent
GO:1901363	0.16	Heterocyclic compound binding
GO:0097159	0.16	Organic cyclic compound binding
GO:0097367	0.14	Carbohydrate derivative binding
GO:0003824	0.12	Catalytic activity
GO:0044877	0.11	Macromolecular complex binding
GO:0042802	0.11	Identical protein binding
GO:0005102	0.11	Receptor binding

The total number of DRs (Disorder Regions) counted in PE_PGRS39 is 10, in which 109 residues comes in the most prolonged DR. Total number of disordered resi-

dues came out to be 291, and the overall disordered percent came out as 70.46 (Fig. 3). DEPP results showed that 22 out of 33 (73.3333%) serines were phosphorylated, 6 out of 21 (28.5714%) threonines were phosphorylated, and 0 out of 2 (0.000%), i.e., no tyrosines were phosphorylated in our protein (Fig. 4).

Structure prediction and validation

(a) Secondary structure prediction: PRISPERD predicts that the secondary structure of PE_PGRS39 contains 67.12% coil, 32.4% alpha-helix, and 0.48% beta-strand. For the projected secondary structure, the pictorial representation is shown in Fig. 5.

(b) Tertiary structure modeling: I-TASSER predicted the tertiary structure of the protein. Initially, ten models were obtained using threading templates from the PDB structural database according to their Z score values ranging from 1.18 to 4.22, from which five best models were selected with C score values ranging from − 3.15 to − 0.68, where a greater C score means greater confidence. The C score of − 0.68 with a TM-score of 0.63 ± 0.14 and RMSD of 8.4 ± 4.5 Å was chosen for the further experiment (Fig. 6a) TM-score greater than 0.5 suggests a valid topology model, and a TM-score less than 0.17 implies random similarity.

(c) Tertiary structure refinement: We used the GalaxyRefine server for refining I-TASSER predicted model with the highest C score. The initial model was structurally changed in terms of GDT-HA, RMSD, and MolProbity score. One model with the values of the following parameters as MolProbity (2.298), GDT-HA (0.9546), RMSD (0.414), clash score (15.8), low rotamers score (0.4), and the Ramachandran plot score of 88.1 was selected as a final model based on model quality scores for all refined models (Fig. 6b).

(d) Final refined structure validation: The refined I-TASSER structure from the galaxy refine tool was verified using the SAVES v5.0 server. SAVES v5.0 consists of a package of five programs. Out of which, the result from the VERIFY-3D program presented that 80.63% of the residues of PE_PGRS39 protein had a 3D-1D arrangement, which is greater than the threshold value of 0.2. So, this model was successfully passed according to VERIFY-3D. ERRAT score of 84.02 validated and predicted the overall sound quality of the structure. Through PROCHECK Ramachandran plot was evaluated. Ramachandran plot depicted that 82.7% of protein residues were in the most favored region (Fig. 6c). Furthermore, 13.0% and 1.7% residues were found in allowed and generously allowed areas,
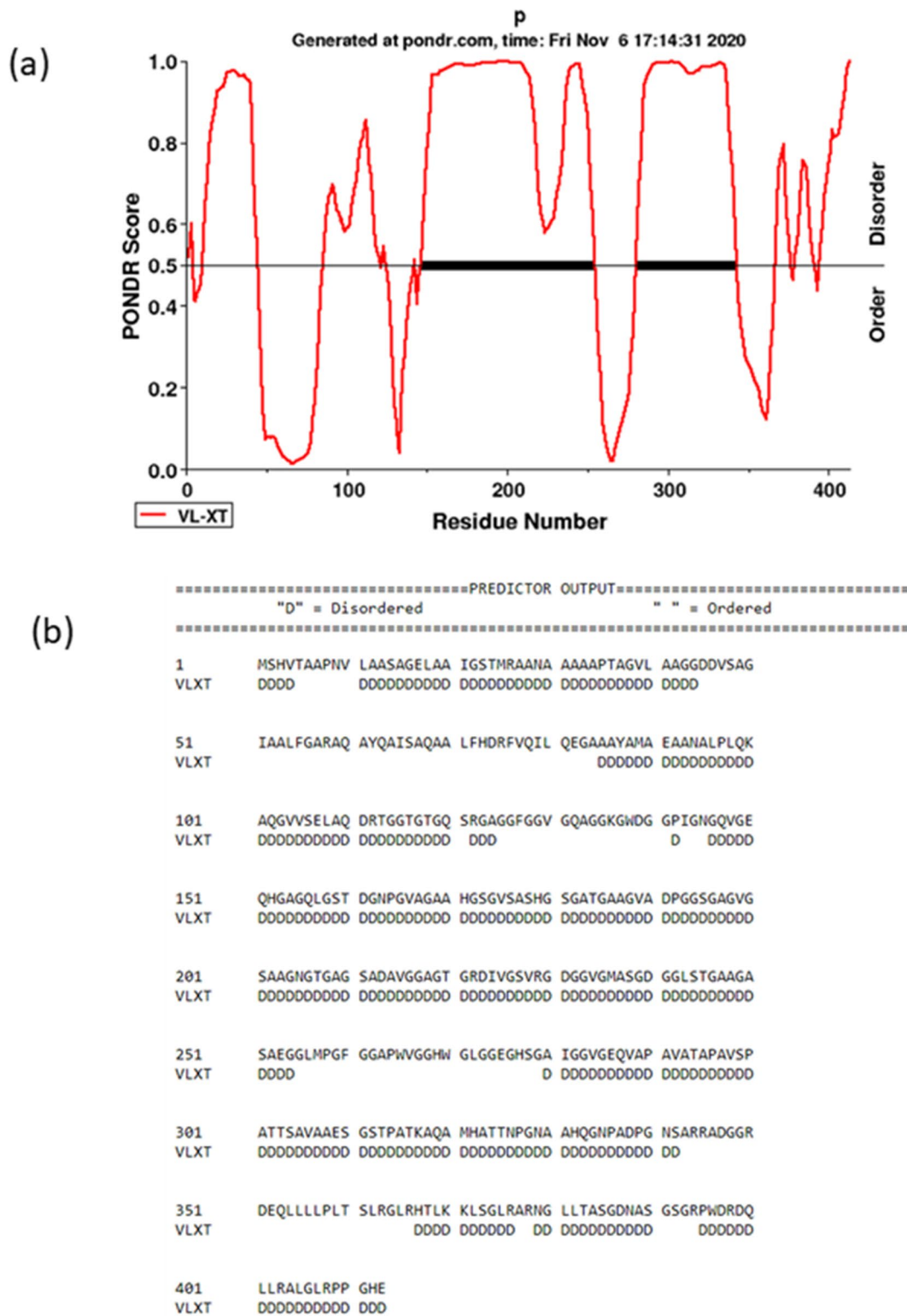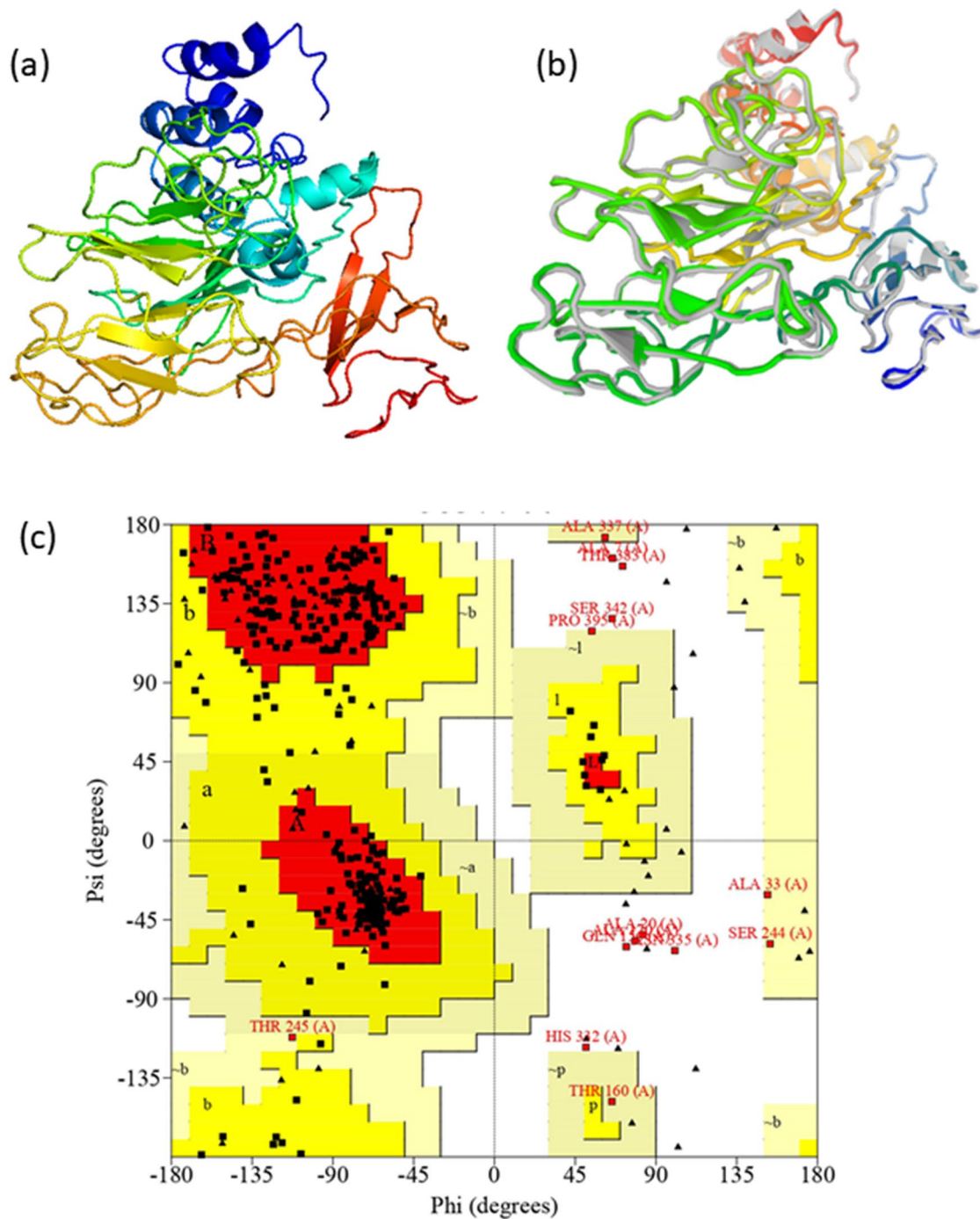
Fig. 3 Disorderness prediction **a** graphical representation of disordered regions. Amino acid residues with disorders are represented with red curves above the 0.5 threshold line. **b** Depicts the disordered areas (D) within the sequence
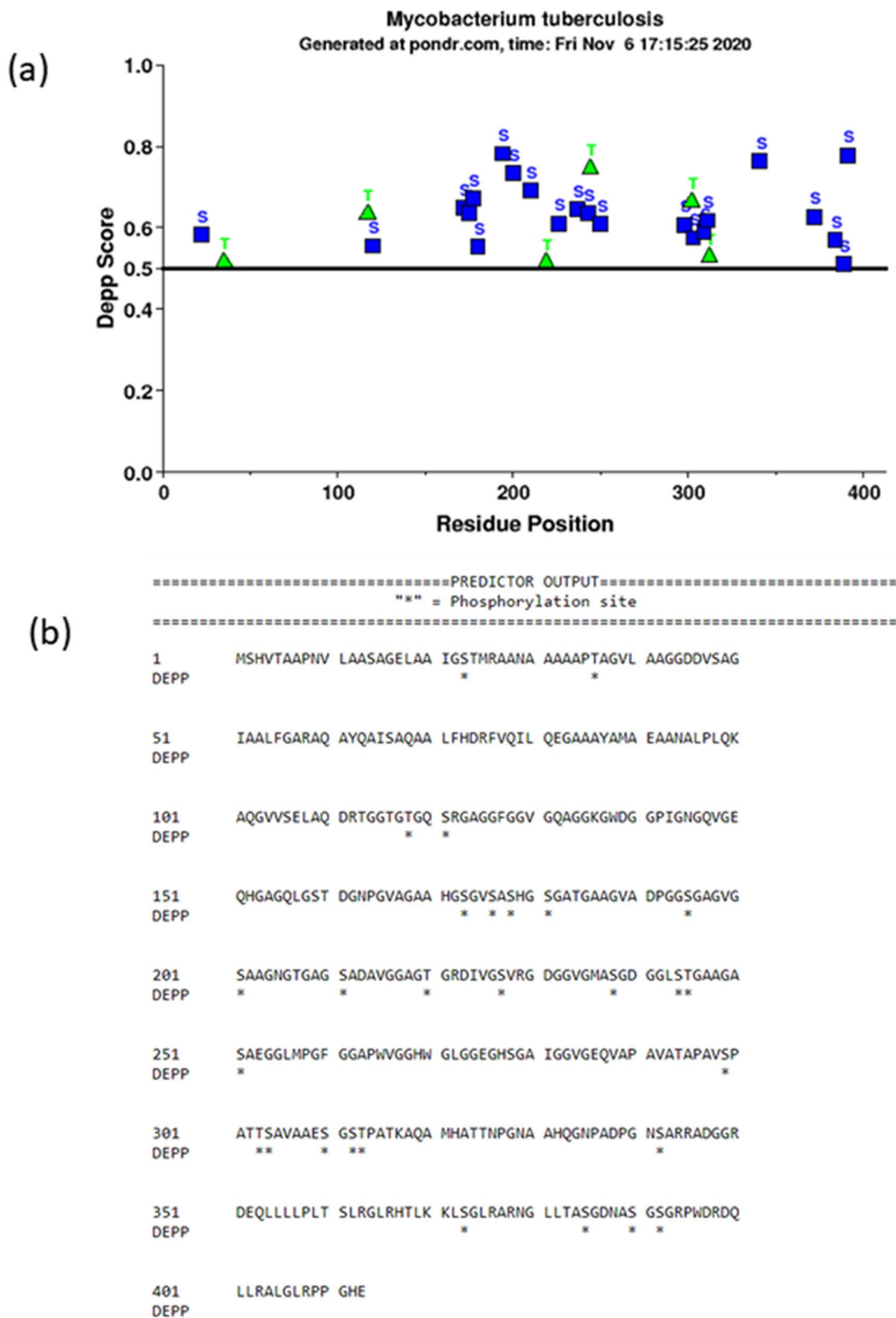
Fig. 4 Phosphorylation sites prediction **a** graphical representation serine (blue), tyrosine (red) and tryptophan (green) residues above the 0.5 threshold value. **b** Showing amino acid residues with phosphorylation denoted by asterisk (*)

Fig. 5 Secondary structure prediction: protein sequence of PE_PGRS39 shows helix (H), strands (E), and coil (C). Four parameters such as secondary structure prediction, confidence of prediction, three-state assignment cartoon, and target sequence are shown here. In the cartoon parameter helix, strands, and coils are denoted by pink, yellow colour, and grey line, respectively

Fig. 6 Predicted tertiary model and its refinement **a** I-TASSER model with highest *C* score of −0.68, **b** refined I-TASSER model by GAL-AXY Webtool, **c** structure validation: Ramachandran plot shows 82% (249 aa) amino acid residues in most favored regions [A, B, L], 13.0% (39 aa) residues in additional allowed regions [a, b, l, p], 1.7% (5 aa) residue in generously allowed regions [~a, ~b, ~l, ~p], and only 0.6% (1aa) residue in disallowed region which infers protein model predicted is good model. Glycine residues (91 aa) are shown in the triangle

respectively, and only 2.7% of the residues were present in the disallowed region. The above data confirmed that the predicted model could be used for further experiments. Additionally, when we analyzed the model with the ProSA server, *Z* score provided the negative value of −3.09, which confirms that the overall model quality is good.

Protein interaction analysis

(a) Protein–protein docking: To confirm the above results and to know the binding affinity of each interacting host protein, molecular docking was performed, the interaction of PE_PGRS39 with integrins and SH3 domain was modeled by pyDock. We performed docking with three integrins with PDB code; namely, 4m76 (β_2 integrin; score: -129.39), 4o02 (β_3 integrin; score: -122.25) and 3vi3 ($\alpha5\beta1$), out of that $\alpha5\beta1$ showed maximum affinity (total binding energy) of -138.678. This $\alpha5\beta1$ integrins act as a triggering molecule for the recruitment of various protein assemblies and strengthen cell adhesion (Speziale et al. 2019). Further, we used PDB code, namely, 1W1F, which represents the human SH3 domain, and it was found that this SH3 domain binds explicitly to the PXXP motif of Mtb with a total binding energy of -91.782. SH3 domain-containing proteins involve in cytoplasmic signaling to perform various essential functions. Molecular visualization was carried out using PyMol software in the cartoon and surface types (Fig. 7).

(b) Protein–ligand docking: Docking simulations were performed with purine nucleotides. Out of which docking simulations of GTP (Fig. 8a) with PE_PGRS39 protein using Autodock Vina showed a maximum affinity with ΔG score of -7.8 kcal/mol (Fig. 8b). This result indicates that GTP has an excellent binding affinity for PE_PGRS39. For further analysis, we used the Discovery Studio to show GTP interaction with amino acids. It showed that GTP formed four carbon–hydrogen bonds with protein, involving the amino acid residues HIS152, GLY188, GLY141, GLY129, and two conventional hydrogen bonds GLY141 and GLU150. GTP also showed an electrostatic interaction with ASP139 and ARG122 (Fig. 8c). Additionally, protein interaction with calcium ligand was confirmed by COACH-D results (Fig. 8d)
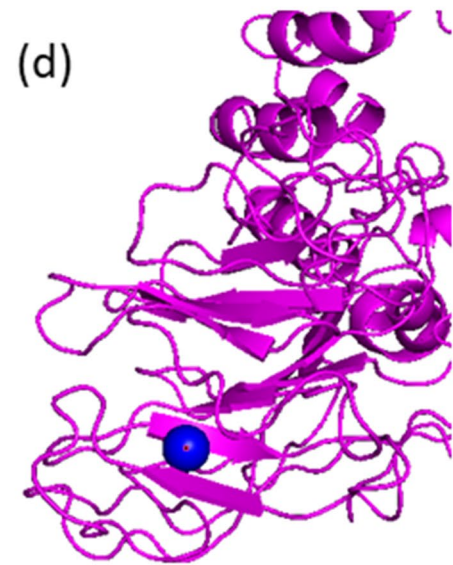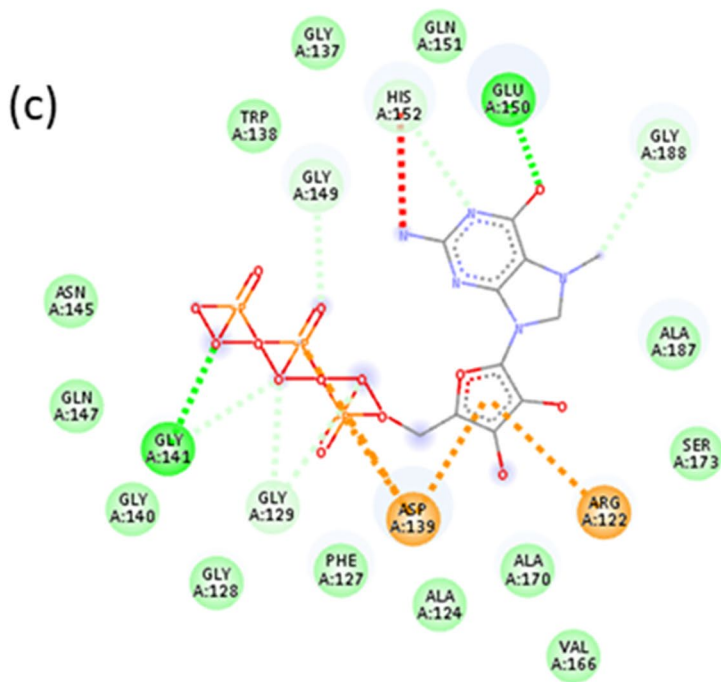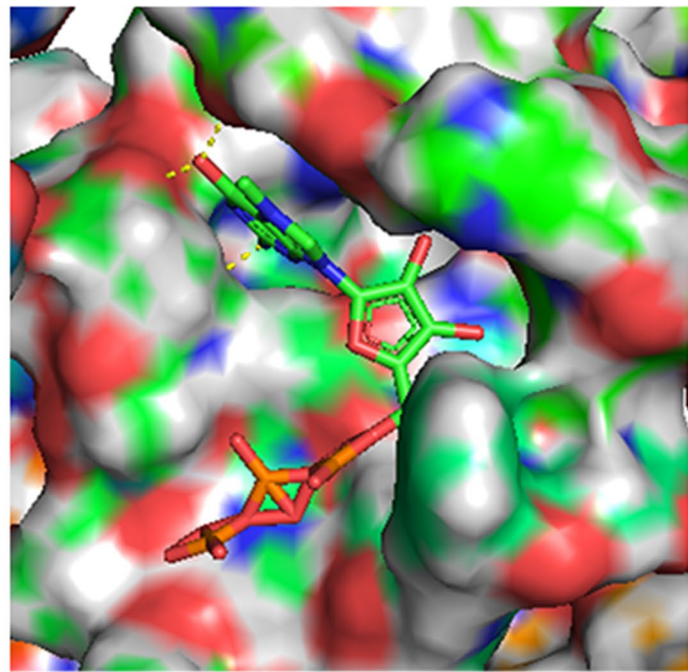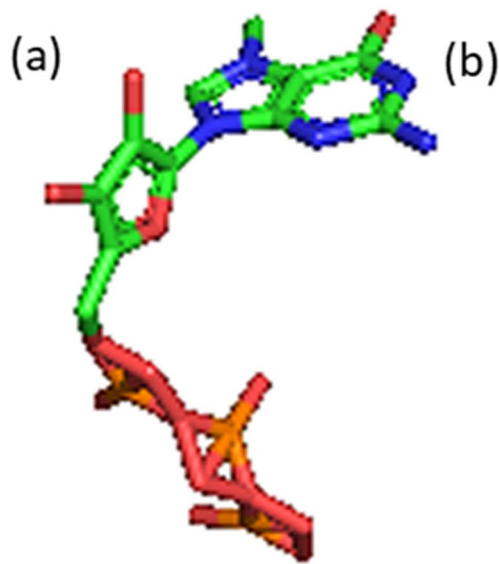
Mutational studies

Missense mutation causes enormous variation in protein binding, thus to forecast the effect of mutations at potential binding sites, an EASE-MM webserver was used. We predict that mutation at RGD, PXXP, calcium, and GTP binding sites can affect binding in one way or another. The extreme value of $\Delta\Delta G$ equal to -1.6171 kcal/mol was detected when we replaced glycine with tryptophan at position 299 in the RGD motif, suggesting a substantial decrease in protein stability. For the SH3 domain binding site, mutation at the P336 position showed the highest destabilization of -1.4727. Mutation at GTP and calcium-binding site showed destabilization of -1.5209 and -1.7457 at G114 and G135 places, respectively, on mutating with tryptophan.

Fig. 7 Pymol representation of protein–protein docking by pyDock **a** cartoon representation of integrin (cyan) binding with RGD (yellow) sequence of PE_PGRS39 (magenta), **b** surface representation of integrin (cyan) binding with RGD (yellow) sequence of PE_PGRS39 (magenta), **c** cartoon representation of SH3 Domain (green) binding with PXXP (orange) sequence of PE_PGRS39 (magenta), **d** surface representation SH3 Domain (green) binding with PXXP (orange) sequence of PE_PGRS39 (magenta)

◀**Fig. 8** Binding conformation of the GTP in the protein binding site **a** representation of GTP ligand used for docking, **b** PyMol representation of GTP-protein binding with a hydrogen bond between amino acid residue, **c** different types of interaction between ligand GTP and PE_PGRS39 receptor, **d** calcium docking with COACH-D

Antigenicity and epitope prediction

(a) Protein antigenicity prediction: PE_PGRS39 was analyzed for its immunogenicity using an independent alignment server called Vaxijen—v2. The threshold for evaluation was kept at 0.4 to forecast whether PE_PGRS39 protein is antigenic or non-antigenic. The protein was predicted as antigenic with a score of overall 1.0072 (probable ANTIGEN). To determine the accuracy of the result, we have also taken an experimentally proved immunogenic Mtb protein (Schierloh et al. 2014), i.e., aceE (contributes to acetyl-CoA production). Using the same Vaxijen—v2 server at 0.4 threshold, aceE predicted to be probable antigen with the score of 0.4770, which ultimately showed that PE_PGRS39 is highly antigenic as its value is more significant than already proven immunogenic protein.

(b) B-cell epitope prediction: We used ABCpred to find several B-cell epitopes in PE_PGRS39. For analysis, default window of 16 amino acids was chosen. Outcomes showed the highest score of 0.91 for the region "GRDIVGSV<u>RGD</u>GGVGM" that starts at 221st residue (Table 5).

(c) T-cell epitope prediction: Using HLAPred, we carried out a prediction of the T-cell epitope. MHC class II binders: we found the "IVGSV<u>RGD</u>G" epitope containing the RGD sequence that starts at 224th residue binds to MHC class II. Sequence search against the human genome showed no identity with humans (A eukaryotic organism). This indicates that this predicted binder can be a vaccine candidate.

Discussion

TB is a dangerous bacterial disease that causes numerous deaths every year. In 2019, it was estimated that a total of 1.4 million persons perished from TB (counting 208,000 individuals with HIV) (WHO Report 2020). Despite vigorous research on Mtb, several proteins have not been studied, and there is still no protective and corrective approach to remove TB, except for BCG. Studies show that BCG can be protective against TB, but it also fails to resolve drug-resistant TB. Thus, there is a high need to consider how different Mtb proteins interact with the host to improve therapeutics.

Several bacteria have evolved to interact with host cell components through protein short linear motifs that mimic host short linear motifs. This interface similarity permits pathogen to smartly alter the host signaling, thus facilitating their internalization and pathogenesis (Guven-Maiorov et al. 2017). In this manuscript, we have tried to analyze an uncharacterized PE_PGRS39 protein functionally and structurally to provide its role in pathogenesis. Further, observing the importance of PE_PGRS39 in host–pathogen interaction and molecular mimicry, we looked forward to the clues for vaccine development (Fig. 9).

PE_PGRS39, a glycine-rich protein made up of 413 amino acids and has a molecular mass of 38722.3 Da (Table 1). Through subcellular localization tools, i.e., cello2GO, we have found that PE_PGRS39 is an extracellular protein that implies that this protein might directly play a role in host–pathogen interaction (Fig. 2, Table 2). Further functional analysis was performed using VICMpred and COFACTOR, which indicated that PE_PGS39 is involve in the cellular process (Table 3), and serves a molecular function as structural molecule activity (Table 4). DEPP server showed that 73.3333% serines, 28.5714% threonines, and no tyrosines were phosphorylated in PE_PGRS39 (Fig. 4). PONDER server predicted functional disordered region (Fig. 3). It is a well-known study that disordered protein regions are functionally active regions of the protein that frequently contain short linear peptide motifs, such as binding sites for the SH3 domain and targeting signals essential for protein function (Linding et al. 2003).

Further, in favor of functional analysis, several motifs like RGD (cell-attachment motif), PXXP, DXXG, and GGXGXD were recognized in PE_PGRS39 (Fig. 1), which binds with integrin receptor, host SH3 domain-containing protein, GTP, and calcium, respectively. To precisely confirm the interaction between PE_PGRS39 protein and host, structural and interaction analyses were also performed. For this, we generated secondary (Fig. 5) and tertiary structure of PE_PGRS39 protein with refinement and validation values, which showed that the predicted protein model is a good quality model (Fig. 6). This good quality model was further used for interaction analysis to validate the binding of PE_PGRS39 with host proteins (integrin and SH3 domain) and ligands (calcium and GTP) which successfully showed a positive result (high affinity) for α5β1 integrin, GTP, and calcium (Fig. 8). However, SH3 domain-PXXP docking results showed low affinity but high specificity (Fig. 7), which is preferably ideal for transient protein signaling in the cellular context (Chandra et al. 2004). Also, from previous studies, it is depicted that several bacteria use the PXXP motif to bind the host SH3 domain and manipulate its cell signaling pathways. In this regard, a cell wall-associated/secretory Rv1917c antigen, characterized by PXXP-SH3 binding can be taken as a positive control which has shown to induce selective

Table 5 Epitope prediction in protein PE_PGRS39

Rank	Sequence	Start position	Score
1	STPATKAQAMHATTNP	312	0.93
2	GRDIVGSVRGDGGVGM	221	0.92
3	GGHWGLGGEGHSGAIG	267	0.91
3	TGTGQSRGAGGFGGVG	116	0.91
4	AISAQAALFHDRFVQI	64	0.90
4	PGGSGAGVGSAAGNGT	192	0.90

maturation of human dendritic cells by controlling the signaling of PI3K-MAPK-NF-B3 and facilitating a change in the subsequent immunity to the Th2 phenotype that helps mycobacterial immune evasion (Bansal et al. 2010).

Pieces of evidence showed that calcium (Ca^{2+}) is a universal signaling ion, and GTP-binding proteins (G-proteins) have also proved to be highly conserved signaling molecules (Meena et al. 2008). Concerning calcium-binding proteins, studies have shown that various bacterial proteins like SdrC and SdrD in *S. aureus* and BapA in *Paracoccus denitrificans*, ClfA and ClfB in *S. aureus* and *Vibrio cholerae*, etc. bind and respond to the host Ca^{2+} by its surface motifs, triggering the molecular mechanisms of adhesion, biofilm formation and processes enabling the development of persistent infections (Elíes et al. 2020).

In this in-silico experiment, after carrying out docking with various β1, β2, and β3 integrins. α5β1, also known as the fibronectin receptor, was shown to have the highest affinity. This result indicates that PE_PGRS39 might be similar to host fibronectin that contains an integrin-binding RGD motif on its extracellular surface. A similar phenomenon has been experimentally verified for the CagL protein of the *Helicobacter pylori* that binds with α5β1 integrin. It mimics fibronectin in triggering cell spreading, focal adhesion formation, and activation of several tyrosine kinases in an RGD-dependent manner, which gives an experimental point into our prediction. They showed that CagL activated tyrosine kinases focal adhesion kinase and Src, etc. Interestingly, host fibronectin also activates a similar range of tyrosine kinases (Tegtmeyer et al. 2010).

Next, we inquired whether these motifs are essential for PE_PGRS39-induced signaling. We found that mutating PXXP, calcium, and GTP binding motif by the EASE-MM tool showed a high decrease in critical stability of the protein. Point mutation of glycine to tryptophan in RGD domain also showed a decrease in stability with a determined score of $\Delta\Delta G$ (change in Gibbs free energy) equal to -1.6171 kcal/mol, signifying that the RGD motif of PE_PGRS39 could be vital for provoking the host cell responses, which are also stimulated by human fibronectin.

Researchers showed that MHC-II peptides flexibly bind with the antigens of extracellular regions (Villadangos 2001). Thus, after considering the importance of PE_PGRS39, we took our experiment towards vaccine development. By B-cell and MHC-II epitope prediction tool, we found that PE_PGRS39 is highly antigenic, and both B-cell and T-cell epitopes contain the RGD sequence (Table 5). Also, comparing the 16mer epitope sequence with the human host genome shows no match, which means that these epitopes are unique to mycobacterium and can also be used in vaccine development to evoke immune response against the pathogen. Studies showed that calcium-dependent signaling mechanisms are potential targets for inhibiting macrophage activation by Mtb (Watanabe et al. 1996). Design of peptidomimics can be initiated by screening-specific distinctive PXXP-containing motifs in the Mtb proteome (Chandra et al. 2004). Therefore, targeting this PE-PGRS39 protein to disrupt its functional features could also prove a successful step towards creating an effective tuberculosis drug. Further experimental analysis on this protein will open a door for host–pathogen interaction studies.

Conclusion

In this study, various computational studies like functional, structural, interaction, and mutational analysis were performed to highlight the role of an unexplored PE_PGRS39 protein. Investigation showed that PE_PGRS39 is a functionally important protein that can interact with host cell components through its short linear motifs. PE_PGRS39 contains motifs like RGD (cell-attachment motif), PXXP, DXXG, and GGXGXD, which binds alpha-5 beta-1 (α5β1) integrin receptor, host SH3 domain-containing protein, GTP, and calcium, respectively with good affinity. Previous studies in different pathogens showed that these motifs mimic the host motifs. Their interaction with the host could alter cellular processes via adhesion and cell signaling, directly playing a potential role in pathogenesis. Further, PE_PGRS39 has the potential to evoke an immune response which gives an insight into vaccine development.
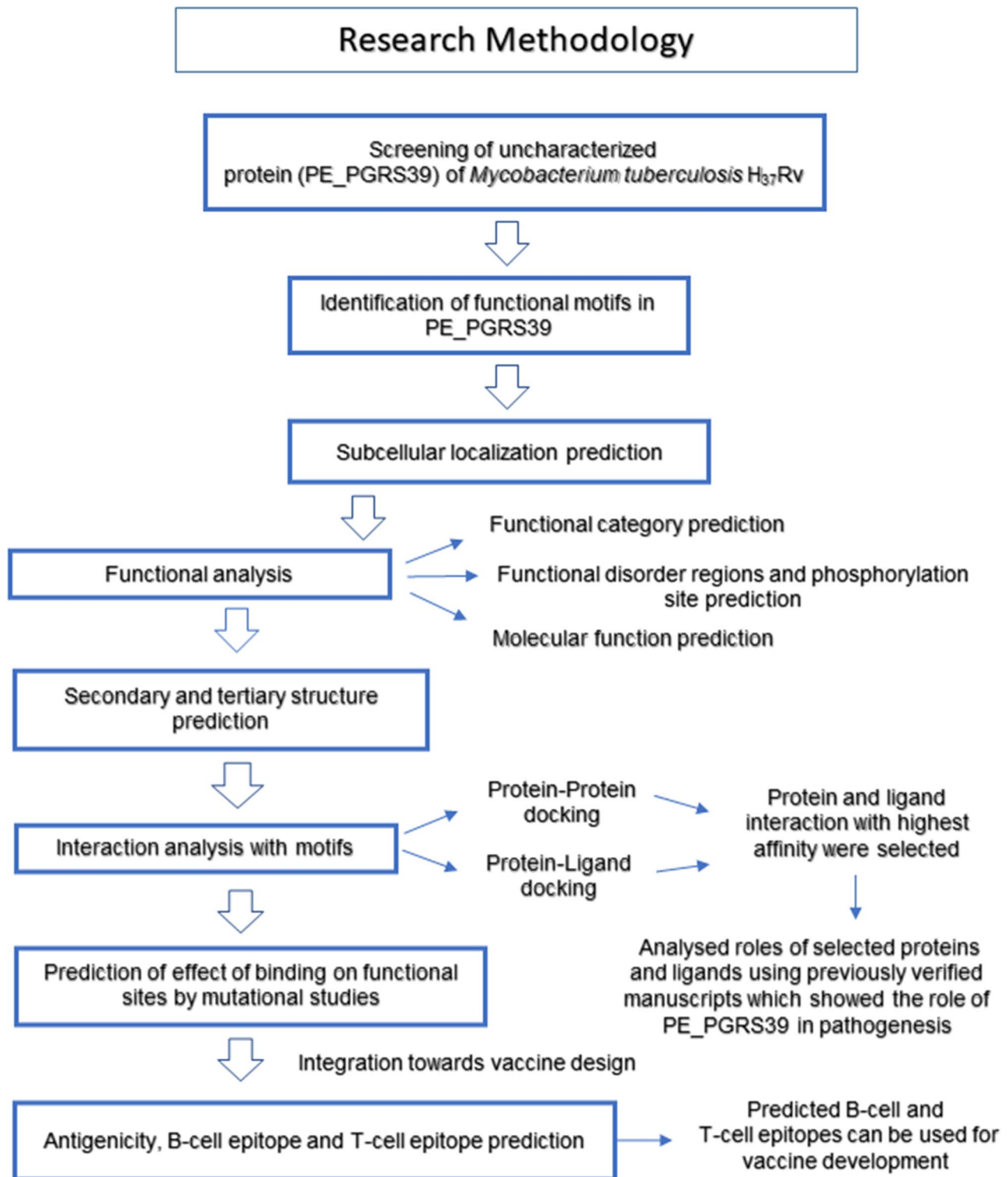
Fig. 9 Flow chart for research methodology

Data availability Yes, the authors agreed for the availability of data to ensure data transparency norms.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate All authors participant agree to publish this work in your esteem Journal: 3 Biotech.

Consent to publish Yes, all authors have given their consent.

References

Agarwal P, Meena S, Meena LS (2020) Comprehensive analysis of GTP cyclohydrolase I activity in *Mycobacterium tuberculosis* H$_{37}$Rv via in silico studies. Biotechnol Appl Biochem. https://doi.org/10.1002/bab.1988

Aota SI, Nomizu M, Yamada KM (1994) The short amino acid sequence Pro-His-Ser-Arg-Asn in human fibronectin enhances cell-adhesive function. J Biol Chem 269:24756–24761

Bachhawat N, Singh B (2007) Mycobacterial PE_PGRS proteins contain calcium-binding motifs with parallel β-roll folds. GPB 5:236–241

Bansal K, Sinha AY, Ghorpade DS et al (2010) Src homology 3-interacting domain of Rv1917c of *Mycobacterium tuberculosis* induces selective maturation of human dendritic cells by regulating PI3K-MAPK-NF-κB signaling and drives Th2 immune responses. J Biol Chem 285:36511–36522. https://doi.org/10.1074/jbc.M110.158055

Beg MA, Shivangi TSC, Meena LS (2018) Structural prediction and mutational analysis of Rv3906c gene of *Mycobacterium tuberculosis* H$_{37}$Rv to determine its essentiality in survival. Adv Bioinform. https://doi.org/10.1155/2018/6152014

Chandra BR, Gowthaman R, Akhouri RR et al (2004) Distribution of proline-rich (PxxP) motifs in distinct proteomes: functional and therapeutic implications for malaria and tuberculosis. Protein Eng Des Sel 17:175–182. https://doi.org/10.1093/protein/gzh024

Cheng TM, Blundell TL, Fernandez-Recio J (2007) pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. Proteins: Struct Funct Bioinf 68:503–515. https://doi.org/10.1002/prot.21419

de la Caridad Addine Ramírez B, Marrón R, Calero R et al (2013) In silico identification of common epitopes from pathogenic mycobacteria. BMC Immunol 14(Suppl 1):S6. https://doi.org/10.1186/1471-2172-14-S1-S6

Degryse B, Fernandez-Recio J, Citro V et al (2008) In silico docking of urokinase plasminogen activator and integrins. BMC Bioinform 9:1–9. https://doi.org/10.1186/1471-2105-9-S2-S8

Delamain C, Spring J, Delamain C, Spring J (2019) Predicting. Speak List Underst. https://doi.org/10.4324/9781351131117-11

Mahadevan R (1998) Reconciling the spectrum of Sagittarius A* with a two-temperature plasma model. Nature 394:651–653. https://doi.org/10.1038/29241

Elíes J, Yáñez M, Pereira TMC et al (2020) An update to calcium binding proteins. Adv Exp Med 1131:183–213

Fauci AS (2001) Infectious diseases: considerations for the 21st century. Clin Infect Dis 32:675–685. https://doi.org/10.1086/319235

Fenn K, Wong CT, Darbari VC (2020) *Mycobacterium tuberculosis* uses Mce proteins to interfere with host cell signaling. Front Mol Biosci 6:1–6. https://doi.org/10.3389/fmolb.2019.00149

Folkman L, Stantic B, Sattar A, Zhou Y (2016) EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. J Mol Biol 428:1394–1405. https://doi.org/10.1016/j.jmb.2016.01.012

Godfrey WL, Rudd CJ, Iyer S, Recktenwald D (2005) Purification of cellular and organelle populations by fluorescence-activated cell sorting for proteome analysis. In: Walker JM (ed) The proteomics protocols handbook. Springer protocols handbooks. Humana Press. https://doi.org/10.1385/1-59259-890-0:067

Goldberg T, Hamp T, Rost B (2012) LocTree2 predicts localization for all domains of life. Bioinformatics 28:458–465. https://doi.org/10.1093/bioinformatics/bts390

Guven-Maiorov E, Tsai CJ, Ma B, Nussinov R (2017) Prediction of host-pathogen interactions for helicobacter pylori by interface mimicry and implications to gastric cancer. J Mol Biol 429:3925–3941. https://doi.org/10.1016/j.jmb.2017.10.023

Heo L, Park H, Seok C (2013) GalaxyRefine: protein structure refinement driven by side-chain repacking. Nucleic Acids Res 41:384–388. https://doi.org/10.1093/nar/gkt458

Kardani K, Bolhassani A, Namvar A (2020) An overview of in silico vaccine design against different pathogens and cancer. Expert Rev Vaccines 19:699–726. https://doi.org/10.1080/14760584.2020.1794832

Kathwate GH (2020) In silico design and characterization of multi-epitopes vaccine for SARS-CoV2 from its Spike proteins. bioRxiv 56:1115–1135. https://doi.org/10.31857/s0016675820100094

Kevlani N, Meena L (2017) Prominent role of FnBPs of *Mycobacterium tuberculosis* in cell adhesion, immune invasion and pathogenesis. Open J Bacteriol 1:007–012. https://doi.org/10.17352/ojb.000002

Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList - 10 years after. Tuberculosis 91:1–7. https://doi.org/10.1016/j.tube.2010.09.008

Linding R, Jensen LJ, Diella F et al (2003) Protein disorder prediction: implications for structural proteomics. Structure 11:1453–1459. https://doi.org/10.1016/j.str.2003.10.002

Meena LS, Rajni (2010) Survival mechanisms of pathogenic *Mycobacterium tuberculosis* H$_{37}$Rv. FEBS J 277:2416–2427. https://doi.org/10.1111/j.1742-4658.2010.07666.x

Meena LS (2015) An overview to understand the role of PE_PGRS family proteins in *Mycobacterium tuberculosis* H$_{37}$Rv and their potential as new drug targets. Biotechnol Appl Biochem 62:145–153. https://doi.org/10.1002/bab.1266

Meena LS (2019) Interrelation of Ca 2+ and PE_PGRS proteins during *Mycobacterium tuberculosis* pathogenesis. J Biosci 44:1–7. https://doi.org/10.1007/s12038-018-9828-4

Meena PR, Meena LS (2015) Fibronectin binding protein and Ca^{2+} play an access key role to mediate pathogenesis in *Mycobacterium*

tuberculosis; an overview. Biotechnol Appl Biochem. https://doi.org/10.1002/bab.1434.This

Meena LS, Meena J (2016) Cloning and characterization of a novel PE_PGRS60 protein (Rv3652) of *Mycobacterium tuberculosis* H_{37}Rv exhibit fibronectin-binding property. Biotechnol Appl Biochem 63:525–531. https://doi.org/10.1002/bab.1411

Meena LS, Meena S (2019) To elucidate the association of Rv0526 gene with the pathogenic potential of *Mycobacterium tuberculosis* H_{37}Rv. Tuberculosis 1:1–6

Meena LS, Chopra P, Bedwal RS, Singh Y (2008) Cloning and characterization of GTP-binding proteins of *Mycobacterium tuberculosis* H_{37}Rv. Enzyme Microb Technol 42:138–144. https://doi.org/10.1016/j.enzmictec.2007.08.008

Montagnani C, Chiappini E, Galli L, Martino De M (2014) Vaccine against tuberculosis: what's new? BMC Infect Dis 14:1–9. https://doi.org/10.1186/1471-2334-14-S1-S2

Monu LSM (2016) Biochemical characterization of PE_PGRS61 family protein of *Mycobacterium tuberculosis* H_{37}Rv reveals the binding ability to fibronectin. Iran J Basic Med Sci 19:1105–1113. https://doi.org/10.22038/ijbms.2016.7736

Rajni, Meena LS (2010) Guanosine triphosphatases as novel therapeutic targets in tuberculosis. Int J Infect Dis 14:e682–e687. https://doi.org/10.1016/j.ijid.2009.11.016

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738. https://doi.org/10.1038/nprot.2010.5

Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40:471–477. https://doi.org/10.1093/nar/gks372

Saha S, Raghava GPS (2006) VICMpred: an SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition. Genomics Proteomics Bioinform 4:42–47. https://doi.org/10.1016/S1672-0229(06)60015-6

Sampson SL (2011) Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin Dev Immunol. https://doi.org/10.1155/2011/497203

Schierloh P, Klepp L, Vazquez C et al (2014) Differential expression of immunogenic proteins on virulent *Mycobacterium tuberculosis* clinical isolates. Biomed Res Int. https://doi.org/10.1155/2014/741309

Sharma H, Meena S, Meena LS (2019) In silico screening of protein Rv3228 to have a vision towards survival and pathogenesis of *Mycobacterium tuberculosis* H_{37}Rv. J Microbiol Biotechnol 8:21–33

Singh N, Upadhyay S, Jaiswar A, Mishra N (2016) Central bringing excellence in open access in silico analysis of protein. Anal Protein J Bioinform 1:1007

Speziale P, Arciola CR, Pietrocola G (2019) Fibronectin and its role in human infective diseases. Cells. https://doi.org/10.3390/cells8121516

Tegtmeyer N, Hartig R, Delahay RM et al (2010) A small fibronectin-mimicking protein from bacteria induces cell spreading and focal adhesion formation. J Biol Chem 285:23515–23526. https://doi.org/10.1074/jbc.M109.096214

Via A, Uyar B, Brun C, Zanzoni A (2015) How pathogens use linear motifs to perturb host cell networks. Trends Biochem Sci 40:36–48. https://doi.org/10.1016/j.tibs.2014.11.001

Villadangos JA (2001) Presentation of antigens by MHC class II molecules: getting the most out of them. Mol Immunol 38:329–346

Watanabe N, Suzuki J, Kobayashi Y (1996) Role of calcium in tumor necrosis factor-α production by activated macrophages. J Biochem 120:1190–1195. https://doi.org/10.1093/oxfordjournals.jbchem.a021540

WHO Report (2020) https://www.who.int/publications/i/item/9789240013131. Accessed 14 Oct 2020

Wu Q, Peng Z, Zhang Y, Yang J (2018) COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Res 46:W438–W442. https://doi.org/10.1093/nar/gky439

Yu CS, Cheng CW, Su WC et al (2014) CELLO2GO: a web server for protein subCELlular lOcalization prediction with functional gene ontology annotation. PLoS ONE 9:1–9. https://doi.org/10.1371/journal.pone.0099368

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinform 9:1–8. https://doi.org/10.1186/1471-2105-9-40

Zimmermann L, Peterhans E, Frey J (2010) RGD motif of lipoprotein T, involved in adhesion of *Mycoplasma conjunctivae* to lamb synovial tissue cells. J Bacteriol 192:3773–3779. https://doi.org/10.1128/JB.00253-10