Kootery and Sarojini
Journal of Genetic Engineering and Biotechnology (2022) 20:55
https://doi.org/10.1186/s43141-022-00340-5

Journal of Genetic Engineering and Biotechnology

RESEARCH Open Access

Structural and functional characterization of a hypothetical protein in the RD7 region in clinical isolates of *Mycobacterium tuberculosis* — an in silico approach to candidate vaccines

Kaviya Parambath Kootery and Suma Sarojini*

Abstract

Background: *Mycobacterium tuberculosis* has been ravaging humans by inflicting respiratory tuberculosis since centuries. Bacillus Calmette Guerine (BCG) is the only vaccine available for tuberculosis, and it is known to be poorly effective against adult tuberculosis. Proteins belonging to the ESAT-6 family and PE/PPE family show immune responses and are included in different vaccine trials. Herein, we study the functional and structural characterization of a 248 amino acid long putative protein novel hypothetical protein 1 (NHP1) present in the RD7 region of *Mycobacterium tuberculosis* (identified first by subtractive hybridization in the clinical isolate RGTB123) using bioinformatics tools.

Results: Physicochemical properties were studied using Expasy ProtParam and SMS software. We predicted different B-cell and T-cell epitopes by using the immune epitope database (IEDB) and also tested antigenicity, immunogenicity, and allergenicity. Secondary structure of the protein predicted 30% alpha helices, 20% beta strands, and 48% random coils. Tertiary structure of the protein was predicted using the Robetta server using the *Mycobacterium smegmatis* protein as the putative protein with homology. Structural evaluations were done with Ramachandran plot analysis, ProSA-web, and VERIFY3D, and with GalaxyWEB server, a more stable structure was validated with good stereo chemical properties.

Conclusion: The present study of a subtracted genomic locus using various bioinformatics tools indicated good immunological properties of the putative mycobacterial protein, NHP1. Evidence obtained from the analyses of NHP1 using structure prediction tools strongly point to the fact that NHP1 is an ancient protein having flavodoxin folding structure with ATP binding sites. Positive scores were obtained for antigenicity, immunogenicity, and virulence too, implying the possibility of NHP1 to be a potential vaccine candidate. Such computational studies might give clues for developing newer vaccines for tuberculosis, which is the need of the hour.

Keywords: *Mycobacterium tuberculosis*, *M. canetti*, RD7, PSPIRED, Robetta, Immunogenic, Ramachandran plot, Flavodoxin-like fold, Membrane protein, Vaccine

Background

The World Tuberculosis Report, 2020, estimated ten million newly infected tuberculosis (TB) cases in 2019 of which 1.2 million people were HIV-negative and 2,08,000 people were HIV positive [1]. The only vaccine developed against TB is BCG, which is not very effective against adult tuberculosis [2]. Though many vaccines are in trial

*Correspondence: suma@christuniversity.in
Department of Lifesciences, CHRIST (Deemed to be University), Bengaluru, Karnataka 560029, India

stages, significant efficacy has not yet been reported. Even MVA85A, one of the most promising among candidate vaccines, failed to give significant protection against the disease in a trial conducted on HIV patients [3]. Furthermore, *Mycobacterium tuberculosis* shows resistance to commonly used drugs, which makes the disease more difficult to treat [4]. The first whole genome sequencing of *M. tuberculosis* was done in 1998, which showed that more than half of the genome encodes hypothetical proteins [5]. Members belonging to the *M. tuberculosis* complex share 99% of genome sequence similarity but exhibit different pathogenic outcomes due to the coevolution of the species with animals and humans [6]. A large number of virulent mycobacterial genes studied were membrane-anchored genes associated with lipid metabolism [7]. Biological functions of hypothetical proteins with high GC content are predicted to play a major role in the pathology and virulence of the bacteria [8]. Comparative study of common genes in *M. tuberculosis* and *M. leprae* showed 219 genes unique for *Mycobacterium* species of which a few codes for highly conserved ESAT-6 family protein [9]. Proteins with low molecular weight encoding the ESAT-6 gene are widely used as diagnostic tools and in vaccine studies as they possess multiple epitopes for both B and T cells [10]. RD1, RD7, RD9, and RD11 regions of *M. tuberculosis* are considered as immunodominant regions. Genes specific to human isolates were identified in the RD7 region of H37Rv strain, for example, the *mce*3 gene (mammalian cell entry) located on the cell surface which plays an important role in the survival of the pathogen in the host [11]. RD7 regions were shown to possess eight mce-associated membrane proteins (Rv1966-Rv1973), two integral proteins (Rv1964, Rv1965), and one membrane protein (Rv1974). Although the function of mce genes is not completely studied, computational studies of mce2 genes have shown T-cell epitopes inducing property revealing the immunogenic nature of mce genes [12].

Computer-aided drug discovery studies in the field of tuberculosis research are a good model to find new drug targets against *M. tuberculosis* [13]. SMRT sequencing of the MTBC genome studies has shown that the genomes are highly conserved with 99% identity. Fifty-one secretory proteins of *M. tuberculosis* were identified using computational tools of which only seven secretory proteins were previously reported [14]. Hypothetical proteins belonging to different classes like enzymes, transporters, receptors, and structural proteins were studied using bioinformatics tools. Structure prediction, mutational analysis, and functional studies of different *M. tuberculosis* proteins were done using in silico methods of which few were predicted to have vaccine potential, ribosome binding regions, GTP binding domains, etc. [15]. Reverse vaccinology studies were done to develop highly immunogenic novel multiepitope sub-unit vaccines against tuberculosis [16].

Our present focus, the NHP1 locus, was obtained by subtractive hybridization in an earlier study using *M. tuberculosis* H37Rv and Indian clinical isolates. Using the initial 384 bp subtracted product as a probe for RFLP with different restriction enzymes followed by sequencing and assembly of the subcloned fragments, a 4.5 kb subtractive fragment was obtained by genome walking. Three potential ORFs were found in the 4.5 loci, NHP1 being one among them [17]. NCBI BLAST studies showed 99% sequence similarity with *Mycobacterium canetti* and predicted 5 potential ORFs. The present study is an in silico approach of the functional and structural studies of one of the novel hypothetical proteins (NHP1), present in the N4.5 locus of clinical isolates. Being a primitive protein and also being located in the RD7 region make NHP1 an ideal vaccine candidate to be explored. In order to check the probable use of NHP1 protein as a potential vaccine candidate against tuberculosis, various computational tools including Expasy ProtParam, SMS suite, NCBI CDD blast, PFAM, SMART, TB Pred, TMHMM, InterPro, VaxiJen, AllerTop, AllergenFP, VirulentPred, GOR4, SOPMA, PSIPRED, Robetta, PROCHECK, and GalaxyWEB server were used. Three positive and negative controls each were used for the comparison of results of the in silico study. The 3D structural studies of NHP1 were done using homology protein modeling. Proteins with high sequence similarity were selected as templates using the MODELLER server. Along with the homology template model, a position-specific scoring matrix (PSSM) and hidden Markov model (HMM)-based server Robetta were considered to develop the 3D structure of the query protein. The developed structures were refined and validated using different quality assessment tools including PROCKECK. The refinement of the developed 3D structure of the protein was also done using the GalaxyWEB server which is based on molecular dynamic simulation. The present study aimed at exploring the possibility of NHP1 protein as a potential vaccine candidate against tuberculosis, a disease for which there are no many vaccines in the market.

Methods

Sequence retrieval

The nucleotide sequence of N4.5 deposited earlier by our research group was retrieved from the National Center for Biotechnology Information (NCBI) GenBank (accession number — GU994138.2). NHP1 locus had 747 bp encoding a protein of 248 amino acids. The nucleotides encoding esxA (Rv3875, gene ID: 886209), esxB (Rv3874, gene ID: 886194), and fbpB (Rv1886c, gene ID: 885785)

used as positive controls were retrieved from NCBI GenBank. whiB2 (Rv3260c, ID: 887598), tuf elongation factor (Rv0685, gene ID: 888262), and cyp144 (Rv1777, gene ID: 885839) used as negative control were also retrieved from NCBI GenBank. The physical and chemical characters, immunological properties, and the secondary and tertiary structures of the protein were studied using different bioinformatics tools.

Physicochemical characterization

Expasy's ProtParam [18] https://web.expasy.org/protparam/ server was used for physical and chemical characterization of the proteins. The prediction included molecular weight, isoelectric point, extinction coefficient, and instability index. SMS (Sequence Manipulation Suite) v2.0 [19] https://www.bioinformatics.org/sms2/ was also used to predict the physical properties using the sequence of 248 amino acids along with the positive and negative controls as input.

Conserved domain search

Bioinformatics tools including CDD-BLAST [20] https://www.ncbi.nlm.nih.gov/Structure/cdd/ wrpsb.cgi, PFAM [21] http://pfam.xfam.org/ and SMART [22] http://smart.embl-heidelberg.de/smart/set_mode.cgi were used to search the conserved domains in the target protein and control samples.

Subcellular localization

Subcellular localization prediction tools like TBpred [23] (https://webs.iiitd.edu.in/tbpred/submission.html), CELLO (sub**CEL**lular **LO**calization prediction) [24] (http://cello.life.nctu.edu.tw/), TMHMM [25] (http://www.cbs.dtu.dk/services/TMHMM/), SignalP (https://services.healthtech.dtu.dk/service.php?SignalP-5.0) and SOSUI [26] (https://harrier.nagahama-i-bio.ac.jp/sosui/) were used to study whether the protein is localized in the cytoplasm or on the cell surface. InterPro [27] (https://www.ebi.ac.uk/interpro/) tool which predicts functional analysis of proteins by comparing the referral sequences with predicted databases provided by different softwares was used to study the protein. Proteins are further classified into different families after domain prediction and site analysis. Furthermore, PFP-FunDSeqE (predicting protein fold pattern with functional domain and sequential evolution information) [28] (http://www.csbio.sjtu.edu.cn/bioinf/PFP-FunDSeqE/) server was used to find protein folding patterns based on functional domain and its evolutionary information. NHP1 protein along with positive and negative controls was studied using these servers.

Prediction of antigenicity, allergenicity, and toxicity

VaxiJen v2.0 [29] (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) server was used to predict the antigenicity of the query proteins. Antigenicity of the protein was studied using the server ANTIGENpro (http://scratch.proteomics.ics.uci.edu/explanation.html). Allergenicity is very crucial in the development of vaccines. AllerTOP v.2.0 [30] (https://www.ddg-pharmfac.net/AllerTOP/) and AllergenFP [31] (http://ddg-pharmfac.net/AllergenFP/data.html) servers were used to predict the allergenicity of the protein. ToxinPred [32] (https://webs.iiitd.edu.in/raghava/toxinpred/algo.php) is a free server which helps in predicting the toxic regions of a protein sequence if any. This server can also be used to identify a new toxic peptide present in organisms.

Virulence factor analysis

VirulentPred [33] (http://203.92.44.117/virulent/) was used to predict the virulence factors. VirulentPred is based on the SVM method with 81.8% accuracy.

T-cell epitope prediction

HLAPred (http://crdd.osdd.net/raghava/hlapred/) and NetCTL [34] (http://www.cbs.dtu.dk/services/NetCTL/) servers were used to predict the T-cell epitopes of the proteins. The NetCTL server was used to predict cytotoxic T-lymphocyte (CTL) epitopes of the query proteins. NetCTL prediction depends on MHCI binding affinity, C-terminal cleavage function, and transporter function associated with antigen processing. IEDB MHCII [35] (http://tools.iedb.org/mhcii/) server was used for identifying helper T-lymphocyte (HTL) epitopes. Human/HLA-DR were chosen as the species/locus with 7 allele human leukocyte antigen (HLA), and 15-m-long epitopes were retrieved. The percentile rank was compared with the Swiss-Prot database showing high MHCII affinity. T-cell epitopes immunogenicity prediction tool was used to study if the predicted T-cell epitopes elicit an immune response. Immunogenicity of class-I peptide MHC complex (pMHC) (http://tools.iedb.org/immunogenicity/) is based on the positions and the properties of the amino acids.

B-cell epitope prediction

BcePred [36] (https://webs.iiitd.edu.in/raghava/bcepred/bcepred_submission.html) and ABCpred [37] (https://webs.iiitd.edu.in/raghava/abcpred/ABC_method.html) were used for the prediction of linear B-cell epitopes. BepiPred by IEDB [38] (http://tools.iedb.org/bcell/) resources was also used to predict

linear B epitopes from the protein structure. In this server, amino acid characteristic scales and hidden Markov Model were used to predict linear B-cell epitopes.

Secondary structure prediction

Online PSI-blast-based secondary structure prediction (PSIPRED) [39] (http://bioinf.cs.ucl.ac.uk/psipred/) and self-optimized prediction method with alignment (SOPMA) [40] (https://npsa-prabi.ibcp.fr/NPSA/npsa_sopma.html) and GOR IV [41] (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html) servers were used to predict the secondary structure of NHP1 protein.

Tertiary structure model

The protein tertiary structure was modelled using MODELLER [42] (https://salilab.org/modeller/) and Robetta [43] (https://robetta.bakerlab.org/) servers. The model protein structures obtained were validated using PROCHECK [44] (http://www.csb.yale.edu/Excite/AT-CSBquery.html) and VERIFY3D [45] (https://saves.mbi.ucla.edu/). PROCHECK was used to check the stereochemical parameters along with the bond angles and bond lengths of the protein structure. ProSA-web [46] (https://prosa.services.came.sbg.ac.at/prosa.php) server was used for Z-score validation. Galaxy Refine — GalaxyWEB server [47] (http://galaxy.seoklab.org/cgi-bin/help.cgi?key=METHOD&type= REFINE) was used to refine the predicted homology model.

Results

Physicochemical characterization

In silico studies of the putative 747 bp long gene in the N4.5 region of *M. tuberculosis* has revealed a 248 amino acid long protein — NHP1 — with a molecular weight of 26.34 kDa, pH 4.90 (highly acidic), with a theoretical pI 5.09, and estimated half-life of 30 h. The hypothetical protein had an average GRAVY of 0.035, making the protein hydrophobic. The aliphatic index of the protein is 86.65, as the volume occupied by alanine, valine, leucine, and isoleucine is defined. Various physicochemical parameters of NHP1 are depicted in Table 1. Physical and chemical characterizations of the proteins used as positive and negative control were studied and are depicted in Tables S1 and S2.

Conserved domain search

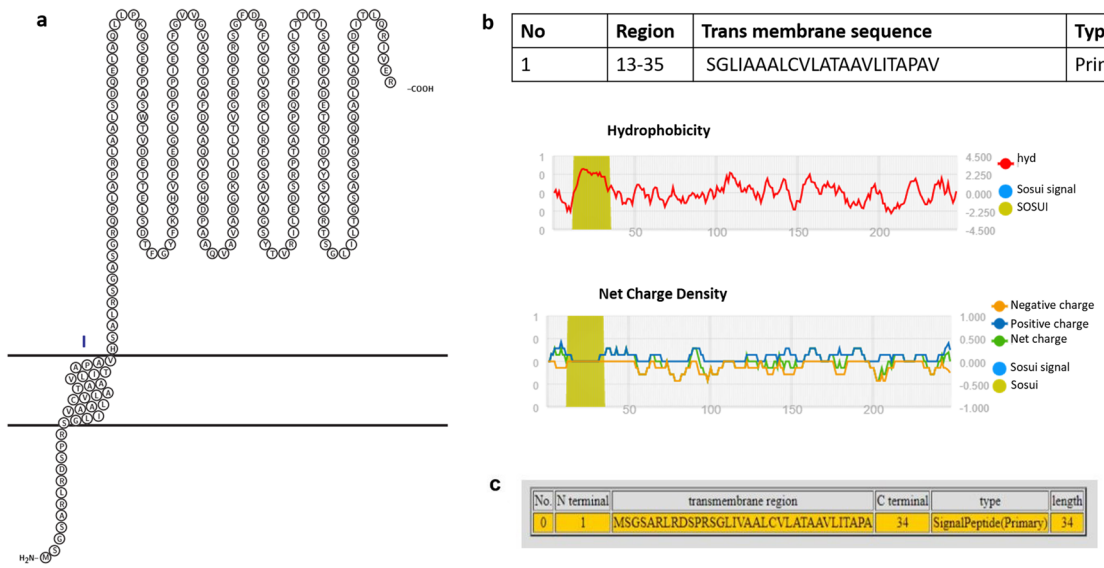
Conserved domains are sequences or structural units reoccurring in an evolutionary perspective. NHP1 nucleotide sequence subjected to NCBI BLAST showed 100% identity with many isolates of *M. tuberculosis* and 99% identity with many isolates of *M. canetti*. CDD-BLAST searches conserved domains by scanning the position-specific scoring matrices (PSSM) with all the conserved domains in the database with a protein query. CDD-BLAST, PFAM, and SMART did not show the presence of any conserved domain in the query protein.

Subcellular localization

The 248 amino acid long sequence was used as the input in SOSUI, TMHMM, and CELLO to study the subcellular localization of the protein. These servers categorize the protein into transmembrane or cytoplasmic ones (Figs. 1 and 2). The results indicated the presence of signal peptide at the N-terminus and a transmembrane domain. Proteins used as positive and negative controls were studied using TMHMM server and did not show any transmembrane helices (Figs. S1 and S2). SignalP 5.0 and LipoP also predicted the presence of signal peptide

Table 1 Physical and chemical characteristics of NHP1 protein predicted using EXPASY ProtParam server

Physicochemical parameters	Values
Number of amino acids	248
Molecular weight	26331.53 Da
pH	4.90
Theoretical isoelectric point (pI)	5.09
Formula	$C_{1167}H_{1826}N_{324}O_{363}S_4$
Total number of atoms	3684
Aliphatic index	86.65
Instability index	44.06
Extinction coefficients (all pairs of Cys residues form cystines)	16055
Extinction coefficients (all Cys residues are reduced)	15930
Total number of negatively charged residues (Asp + Glu)	28
Total number of positively charged residues (Arg + Lys)	21
Grand average of hydropathicity (GRAVY)	0.035



No	Region	Trans membrane sequence	Type
1	13-35	SGLIAAALCVLATAAVLITAPAV	Primary

No.	N terminal	transmembrane region	C terminal	type	length
0	1	MSGSARLRDSPRSGLIVAALCVLATAAVLITAPA	34	SignalPeptide(Primary)	34

Fig. 1 Cellular localization image of NHP1 protein showing a signal peptide with transmembrane structure using the server SOSUI. **a** N-terminal, transmembrane, and C-terminal domains of NHP1. **b** NHP1 hydropathy and charge plot. **c** Transmembrane sequence of NHP1

and also a cleavage site between 37 and 38 amino acids of NHP1 (Fig. 3). SignalP 5.0 predicted the presence of signal peptide for esxA protein used as positive control but did not predict any signal peptide for other positive and negative control proteins (Figs. S3 and S4). InterPro tool predicted the presence of signal peptide and a transmembrane helix from regions 13–35 at N-terminus. The TBpred server also predicted the presence of an N-terminal signal peptide with transmembrane helices. PFP-FunDSeqE server predicted the query protein folding pattern to have flavodoxin-like folds, based on the functional domain and sequential evolutionary information. Evidence for ATP interacting residues were also found in the protein. The query NHP1 was thus predicted to be confined to the membrane and was found to possess a transmembrane domain in the N-terminal region with a signal peptide. Since the protein possesses these properties, immunological studies were further done.

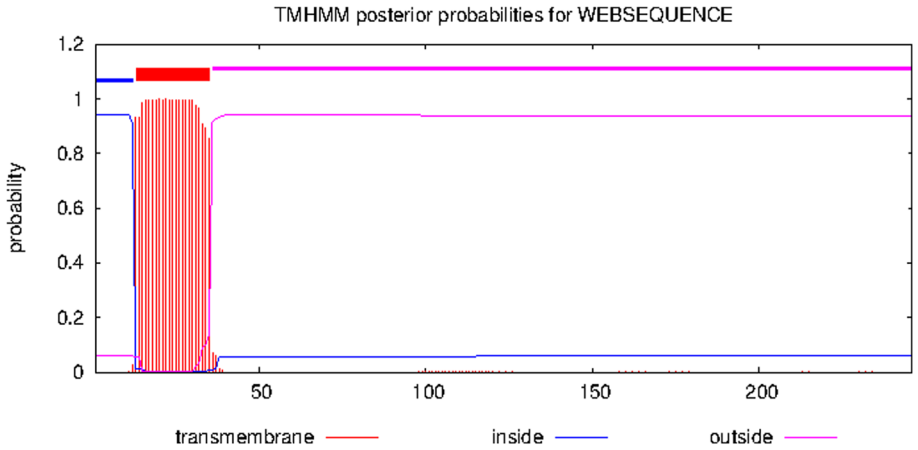
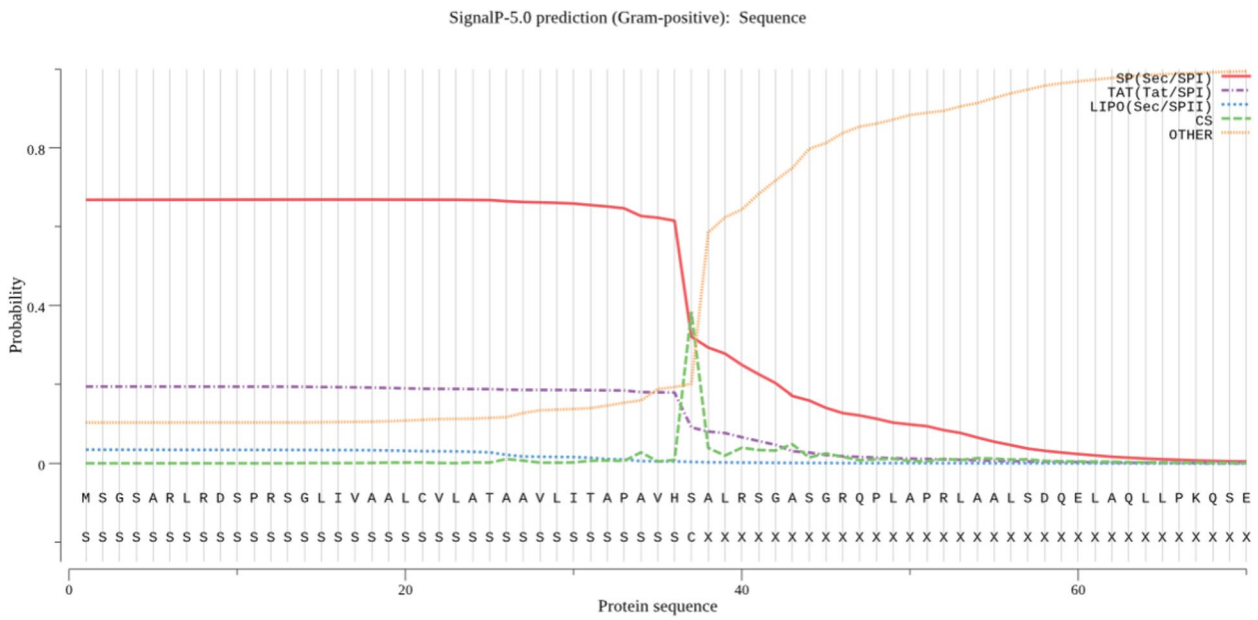


Fig. 2 Graphical output of a transmembrane domain between 13 to 37 amino acids of NHP1 protein predicted using the server TMHMM

Prediction of antigenicity, allergenicity, toxicity, and virulence

VaxiJen servers predicted the protein as antigenic based on the physical and chemical properties, where the default threshold of 0.4 was chosen as the antigenicity



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.6676	0.1952	0.0347	0.1025

Fig. 3 Graphical representation of the presence of signal peptide (sec/spl) with a cleave cite between 37 and 38 amino acids of the NHP1 protein using SignalP server

measure. The query protein had an antigenicity score of 0.4509. The positive control proteins gave a VaxiJen threshold of esxA-0.557, esxB-0.782, and fbp-0.584 and negative control proteins whiB2-0.356, tuf ef-0.398, and cyp144-0.37, respectively. ANTIGENPro server predicted the antigenicity of the query protein as 0.764607. AllerTOP v.2.0 and AllergenFP predicted the query protein to be a probable non-allergen. Among the positive controls used, esxA protein was predicted as a probable allergen and toxic in nature, whereas esxB and fbpB proteins were nontoxic and non-allergen in nature. According to ToxinPred, whiB2 protein was studied as toxin, and tuf ef protein according to AllergenFP was studied as a probable allergen candidate. ToxinPred data predicted NHP1 protein to be nontoxic. VirulentPred data has shown NHP1 to be a virulent protein. Table 2 shows the comparison of immunological properties of NHP1 protein along with positive and negative controls.

T-cell epitope prediction

HLAPred predicted 22 T-cell epitopes for the NHP1 protein. Predicted binders were scanned for genomic molecular mimicry against different species where no identical sequences were seen which can be used as a vaccine candidate. NetCTL server predicted the presence of seven MHC ligands with a threshold of 0.75 as shown in Table 3. IEDB MHC II binding predictions were used to predict the T-helper cell epitopes for allele HLA DR. The adjusted rank ranged from 0.02 to 100 where the recommended value is 2.2 where low adjusted rank ranging from 0.02 to 2.2 is considered good binders. All proteins chosen for positive and negative controls were studied for T-cell epitope binding property, and except fbpB protein, esxA and esxB proteins did not show the threshold value using NetCTL server and are shown in Table S3. T-cell epitope threshold values of all the negative controls using NetCTL are shown in Tables S4, S5, and S6. T-cell class I pMHC immunogenicity predictor depicted the immunogenicity of the protein to range from −0.02 to 1.09 threshold. T-cell class I pMHC immunogenicity threshold is shown in Table 4. pMHC class T-cell epitope studies of the positive and negative controls are given in Tables S7–S12.

B-cell epitope prediction

ABCpred was used to predict the B-cell epitope of the NHP1 protein using artificial neural networks. A threshold of 0.5 along with an epitope length of 16 amino acids was used in the ABCpred server. Eighteen epitopes were predicted by the ABCpred server above (0.92–0.56) the

Table 2 Comparison of the immunological properties of NHP1 protein with positive and negative controls

Prediction servers	NHP1	Positive control			Negative control		
		Rv3875	Rv3874	Rv1886c	Rv3260c	Rv0685	Rv1777
VaxiJen	0.450	0.557	0.782	0.584	0.356	0.398	0.370
AntigenPro	0.765	0.891	0.923	0.859	0.101	0.257	0.264
AllerTop	Non-allergen	Probable allergen	Non-allergen	Non-allergen	Non-allergen	Probable allergen	Non-allergen
AllergenFP	Non-allergen	Probable allergen	Non-allergen	Non-allergen	Non-allergen	Probable allergen	Non-allergen
ToxinPred	Nontoxic	Toxic	Nontoxic	Nontoxic	Toxic	Nontoxic	Non-toxic
VirulentPred	Virulent	Virulent	Virulent	Virulent	Virulent	Non-virulent	Virulent
HLAPred	Potential vaccine candidate	Potential vaccine candidate	Potential vaccine candidate	Potential vaccine candidate	Potential vaccine candidate	Can cause auto-immune disease	Potential vaccine candidate

Table 3 Sequence and threshold values of MHC ligands of NHP1 protein using NetCTL server

No	Sequence	Values
1	LSDQELAQL	0.8836
2	LSDTFGYFK	1.9021
3	FGSAVAGSY	0.8195
4	TAGPQRFRY	0.8411
5	PADETRTDY	1.6935
6	ETRTDYYSY	1.1664
7	RTDYYSYGR	0.7605

threshold level, where the default threshold value was 0.51. BcePred predicted the presence of six B-cell epitopes in the query protein with an accuracy of 58.7% at 2.3 threshold. Physicochemical characters like hydrophobicity, polarity, and flexibility were combined together at a threshold of 2.38 for linear B-cell epitope prediction by BcePred. BepiPred predicted the presence of B epitopes in the amino acid sequence of NHP1 with a threshold ranging from 0.2 to 0.722 (Fig. 4). BepiPred threshold of the positive and negative controls studied is depicted in Figs. S5 and S6.

Protein secondary structure prediction

The secondary structure of NHP1 was predicted using different computational tools (Table 5). SOPMA prediction is based on similar properties and evolutionary significance of homologous proteins in the databases with 69.5% prediction possibility. SOPMA predicted 33.47% alpha helical residues along with 18.55% extended strands, 3.23% beta turns, and 44.76% random coils. The GOR IV method is based on the probability of the amino acids to form a secondary structure. The GOR IV server predicted a similar structure as the SOPMA server. PSIPRED is an online secondary structure prediction tool which also predicts transmembrane topology, transmembrane helix, domain recognition sites, etc. PSIPRED also predicted strand, helix, and coiled secondary structure for the query protein (Fig. 5).

Protein tertiary structure modeling

The conventional method of determining the three-dimensional protein structure is by NMR spectroscopy and X-ray diffraction. 3D homology models of NHP1 protein were predicted using the homology protein structure modeling method. Homology modeling has led to a new breakthrough in the field of computational biology. This modeling is very effective in studying molecular evolution since evolutionarily-related proteins share a common structure. MODELLER server is based on homology and comparative modeling of the protein sequence. MODELLER selected the most appropriate structure similar to NHP1 protein in the PDB database. The template 4TMD_A showed a probability of 99.56%

Table 4 Immunogenicity threshold value of NHP1 protein using T-cell class I pMHC immunogenicity predictor

Peptide	Length	Score
LAQLLPKQSEFPASWTVDETTELSDTFGYFKYHVFDEGLGFDPIECFGVVGVASTGAFDA	60	1.09703
AQVFGHDPAAQVAVADGKDILLTVGREFDRSGFDAFVGLVSRCLRFGSAVAGSYTVRILE	60	1.08736
DSRPTAGPQRFRYSLTTTISAEPADETRTDYYSYGRTSGLILTGSAGSGHQQALDALFDI	60	0.31786
TLQRIVER	8	0.26744
MSGSARLRDSPRSGLIAAALCVLATAAVLITAPAVHSALRSGASGRQPLAPRLAALSDQE	60	0.02806

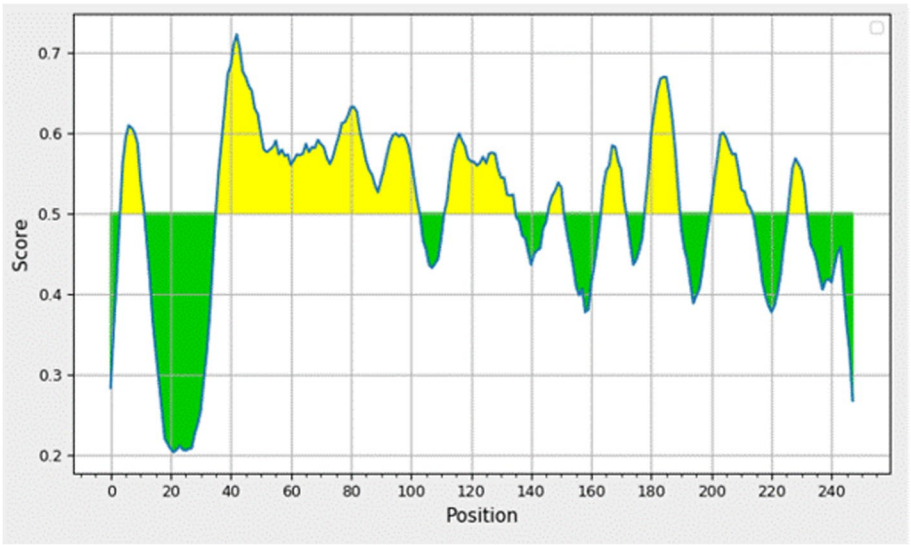


Fig. 4 BepiPred threshold graph of NHP1 protein predicting linear epitopes of NHP1 protein. Yellow color depicts the predicted epitope residues

Table 5 Secondary structure elements developed by SOPMA and GRO4 server

Secondary structure elements	Symbol	No. of amino acids		Percentage	
		SOPMA	GORIV	SOPMA	GOR IV
Alpha helix	Hh	83	70	33.47%	28.23%
3_{10} helix	Gg	0	0	0.00%	0.00%
Pi helix	Ii	0	0	0.00%	0.00%
Beta bridge	Bb	0	0	0.00%	0.00%
Extended strand	Ee	46	55	18.55%	22.18%
Beta turn	Tt	8	0	3.23%	0.00%
Bend region	Ss	0	0	0.00%	0.00%
Random coil	Cc	111	123	44.76%	49.605

with an E-value 1.8e-13 with a target length of 195. The predicted model was saved as PDB format. 3D structure of the protein was predicted using the Robetta server. Robetta is an automated tool for protein structure prediction using homology modeling and de novo structure prediction method.

The predicted structures were analyzed using PROCHECK, VERIFY3D, and ProSA. Ramachandran plot was analyzed for the predicted model. The Ramachandran plot of the Robetta structure indicated 92.7% of residues with 188 residues in favorable regions and 20 residues in allowed regions. The number of non-glycine and non-proline residues was 212. ProSA web server was used for Z-score validation. The Z score of the 3-D sequence was estimated as −6.06. Analysis of the predicted structure using the VERIFY3D server showed that 80.24% of the residues have an average 3D-1D confirmation score. The refined image of the protein was predicted by the GalaxyWEB Refine server. GalaxyWEB Refine is used to improve the accuracy of the predicted model using side-chain perturbation. Refined 3D homology model of NHP1 had 95.1% residues as favored in the Ramachandran plot, with an RMSD value of 0.362 (Fig. 6).

Discussion

Tuberculosis is one of the oldest infectious diseases which still causes extensive mortality and morbidity globally. The currently popular BCG vaccine is efficacious in imparting protection against childhood tuberculosis and TB meningitis, but is not very effective in preventing adult respiratory TB, and hence, there is an urgent need for identifying newer vaccine candidates. In silico studies of RD regions of *M. tuberculosis* depicted immunodominant properties. Genes encoded by RD1, RD7, and RD9 regions were found to be immunodominant antigens against T cells [48]. In the present in silico study, the hypothetical protein NHP1 mapped to be present in the RD7 region of *M. tuberculosis* was shown to encode a protein of molecular weight of 26.34 kDa, with highly acidic pH of 4.90 and an isoelectric point of 5.09. Physicochemical characterization of NHP1 protein showed a GRAVY score of 0.035 and an instability index of 44.06 making the protein hydrophobic. This also points to the fact that this could be a membrane protein of the bacteria. A high aliphatic index of 86.65 makes NHP1 thermostable as the greater the aliphatic index of a protein, the higher is its thermostability.

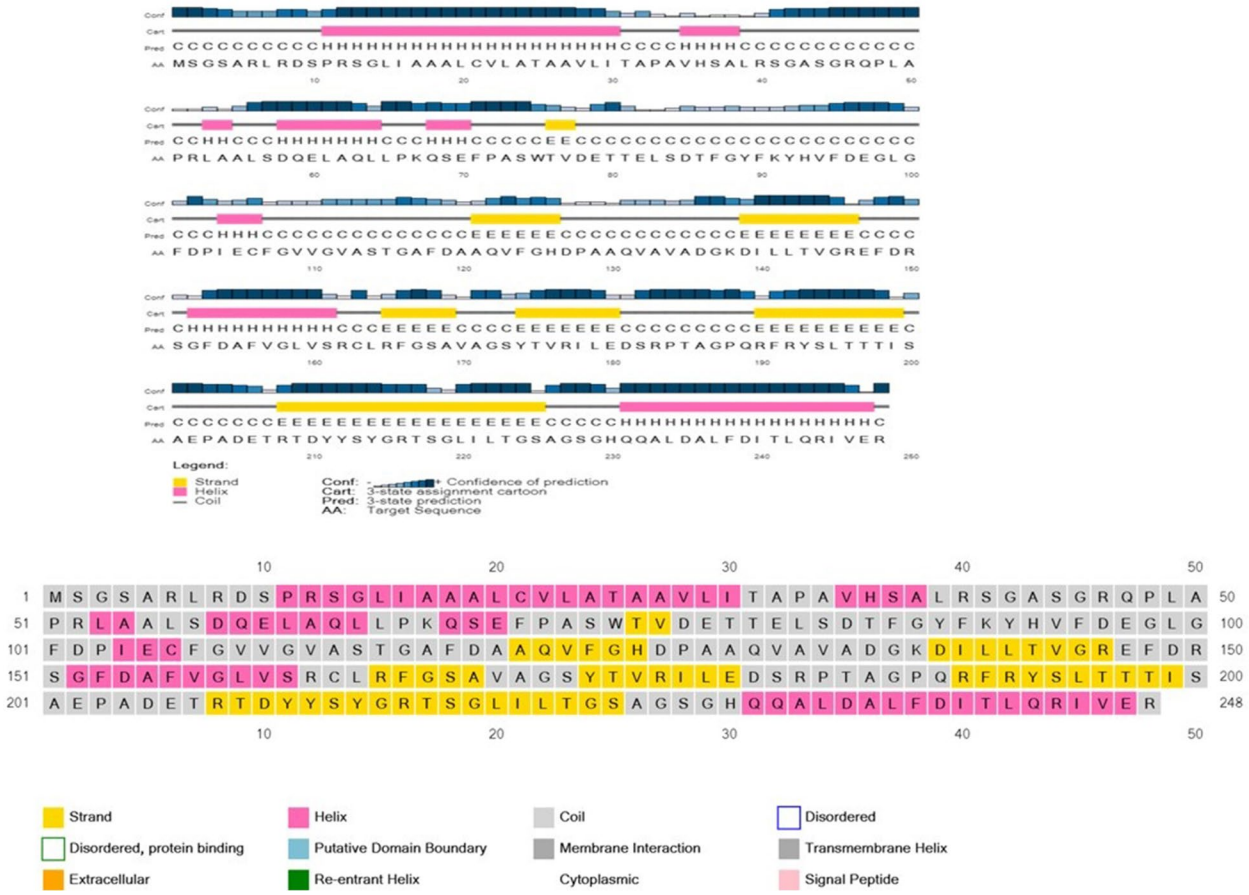


Fig. 5 NHP1 protein secondary structure predicted by PSIPRED server displaying the presence of helix, strands, and random coils

Three negative and three positive controls were used in the study for comparison of the physical, chemical, and immunological properties of the query protein. esxA (6 kDa early secreted antigen, encoded by Rv3875), esxB (protein activates human neutrophils, encoded by Rv3874), and fbpB (Ag85B elicits cellular immunity, encoded by Rv1886c) which were extensively studied for immunological properties both in silico and in vivo and further subjected to vaccine studies were selected as positive control. whiB2 (Rv3260c), tuf elongation factor (Rv0686), and cyp144 (Rv1777) proteins that are important for the biochemical functions of the bacteria were used as negative control.

Subcellular localization studies play an important role in identifying proteins as drug targets, and most of the prediction tools are validated against mycobacterial proteins [49]. NHP1 protein was predicted to contain a signal sequence with transmembrane helices attached to the membrane. Previous studies conducted on twelve P450 enzymes of *M. tuberculosis* using different computational tools like CELLO, TMHMM, and HMMTOP predicted these proteins as cytoplasmic transmembrane proteins with no signal peptides. Homology modeling of the proteins was done using RaptorX, SWISS-MODEL server, and Phyre 2. Molecular docking of the proteins was done with azole drugs using AutoDock Vina and revealed that the P450 enzymes were found to act as a target site for novel drugs. Different azole drugs docked against P450 enzymes were predicted to be good drug candidates against *M. tuberculosis* [50]. According to the PSIPRED tool, NHP1 had an extracellular C-terminal region, a transmembrane helix, and an N-terminal region attached to the membrane. *M. tuberculosis* membrane proteins are known to be immunogenic. Surface membrane proteins are considered as potential vaccine targets, while cytoplasmic proteins are possible drug targets [51]. Since membrane-associated proteins of *M. tuberculosis* are generally potent activators of human T-cell responses [3], NHP1 also could be thought of as a potential vaccine candidate.

A study of 999 hypothetical proteins of *M. tuberculosis* H37Rv was published in 2016 which can be used

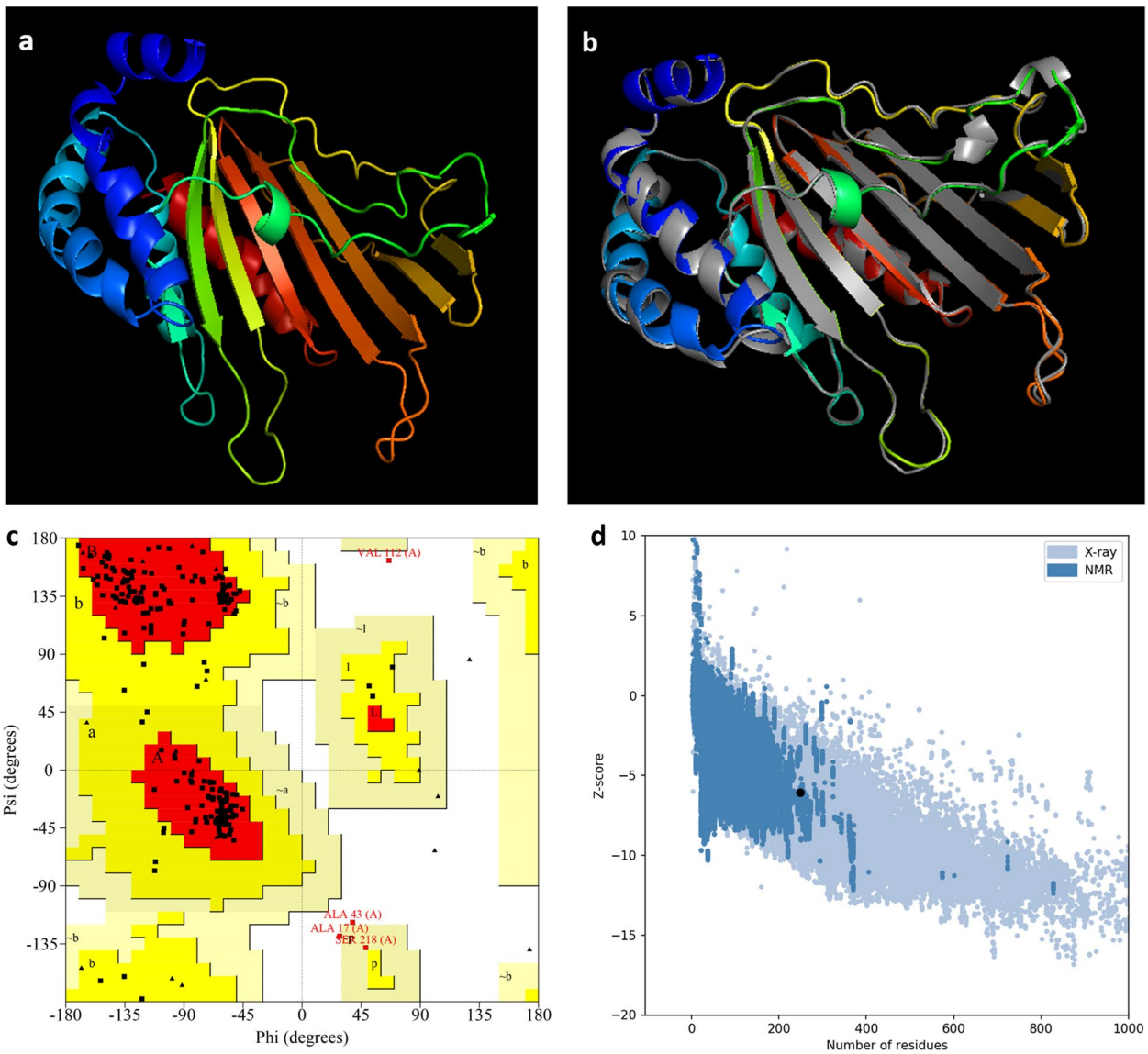


Fig. 6 **a** The 3D model of NHP1 protein obtained from Robetta server following homology modeling. **b** Refined model after refinement superimposition using Galaxy Refine-GalaxyWEB server. **c** Ramachandran plot for structural validation with 92.7% residues in the favored region. **d** ProSA-web, with a Z score of −6.06

for comparative studies. These 999 hypothetical proteins were further studied for conserved domains using CDD Blast and ScanProsite. Ninety-eight proteins were selected from this set and further studied for structural and functional characterization. Proteins with plasma membrane localization and transmembrane topology prediction were studied for antigenicity using VaxiJen server. Proteins having an antigenic score of > 5 were considered and selected for epitope prediction of which two proteins were shown to be having 15 linear epitopes and 21 conformational epitopes. This study has shown these epitopes can be candidates for developing new drugs or vaccines against tuberculosis [15].

ATPint: tool predicted the ATP binding residues of the NHP1 protein (https://webs.iiitd.edu.in/cgibin/atpint/chkres?8477). PFP-FunDSeqE server results depict NHP1 protein to have a flavodoxin-like folding. Proteins with flavodoxin-like folding patterns are considered as ancient origin structures. Flavodoxin family proteins of *M. bovis* BCG, *M. canetti*, *M. pinnipedii*, and few other species of *Mycobacterium* were found to be deposited in EMBL UniProt (https://www.uniprot.org/uniref/UniRef100_

A0A0K2HZ71). Flavin cofactors were present in flavodoxin folding proteins in the electron transport chain, and in few bacteria, these proteins were induced in low iron conditions [52].

Antigenicity refers to the ability to recognize a specific antigen followed by an immune response, whereas immunogenicity is the ability to induce cellular and humoral immune response. An ideal vaccine candidate should be both antigenic and immunogenic [53]. Immunological properties of NHP1 protein were studied, predicting the protein to be highly antigenic, having epitopes to both T-cell and B-cell receptors. T-helper cells play a major role in providing immunity against *M. tuberculosis*. T-cell epitope prediction of proteins including ESAT-6, CFP10, Ag85B, and MPT70 was studied using ProPred server [54]. PPE65 family proteins were studied using immunoinformatics and were predicted to be vaccine candidate proteins [55]. For VaxiJen, all the positive control proteins should possess an antigenic score of > 0.4 and can be subjected to vaccine studies, but the negative control proteins should have an antigenic score of < 0.4 making them nonantigenic. The antigenic threshold of NHP1 protein was 0.45, and hence when compared to the positive and negative controls, the query protein can be used as a vaccine candidate. The antigenic threshold of the positive controls was as follows: esxB (0.782), fbpB (0.584), and esxA (0.557), and negative controls were whiB2 (0.35), tuf elongation factor (0.398), and cyp144 (0.37). The antigenic threshold of NHP1 found using the server Antigen Pro was 0.765.

Allergenicity studies provide clues regarding the possible allergic reactions stimulated by the vaccine. AllerTOP data has shown that NHP1 is nonallergenic. According to AllerTOP and AllergenFP, the positive control proteins esxB and CFP10 protein were nonallergenic and nontoxic, whereas esxA protein was a probable allergen and toxic in nature. In the present study, NHP1 was predicted to have high binding affinity towards B and T cells, showed allergenicity and antigenicity, and was also nonallergenic. Hence, the protein could be considered as a potential vaccine candidate. B-cell epitopes generally play an important role in eliciting humoral immunity and hence are significant in vaccine design too. All positive control proteins showed the presence of more epitopes binding to both T and B cells. According to NetCTL server, NHP1 protein showed 7 MHC ligands above the threshold value, whereas fdpB (positive control) showed 6 MHC ligands. Only 2 MHC ligands above the threshold value were found in all three negative controls under study. IEDB MHC II predicted the presence of T-helper cell epitopes of which fbpB and cyp144 were having T-helper cell epitopes. BepiPred predicted the presence of B epitope alleles for positive and negative controls. In comparison with all positive and negative controls, NHP1 protein showed immunological properties which tend to make it a vaccine candidate.

The secondary structure of NHP1 was predicted using SOPMA, PSIPRED, and GOR-IV. SOPMA and PSIPRED were earlier used to predict the structure of the hypothetical protein CGM946K2_146 of *M. tuberculosis*. The 3D structure of NCGM946K2_146 gene of *M. tuberculosis* was studied using tools like SWISS-MODEL server, Phyre2, and MODELLER, of which MODELLER was found to be the most efficient tool for homology protein structure prediction. The structural and functional studies of NCGM946K2_146 pointed to those of a potential therapeutic drug candidate [56]. PSIPRED was used to study the secondary structure of Rv1907c gene of *M. tuberculosis H37Rv* [57]. Structural studies of alpha-phosphoglucomutase of *M. tuberculosis* were carried out using GOR-IV server. SOPMA and GOR IV predicted 28–33% of alpha helical structure, 18–22% extended strand structure and 44–49% random coil structure, and 3% beta turn structure. PSIPRED predicted the secondary structure similar to the prediction of GOR-IV and SOPMA with high confidence of prediction. The tertiary structure of the NHP1 protein was studied using homology modeling. Membrane proteins including NP_216679, NP_218309, and NP_218312 of *M. tuberculosis* were studied for physical and chemical characterization, secondary structure prediction, and functional characterization using different computer tools. These proteins were found to be membrane proteins with transmembrane helices.

3D structural studies of NHP1 were done using homology protein modeling. Proteins with high sequence similarity were selected as templates using the MODELLER server. Along with the homology template model, a position-specific scoring matrix (PSSM) and hidden Markov model (HMM)-based server Robetta were considered to develop the 3D structure of the query protein. The developed structures were refined and validated using different quality assessment tools including PROCHECK. The refinement of the developed 3D structure of the protein was also done using the Galaxy web server which is based on molecular dynamic simulation.

The physical and chemical parameters of a multi-epitope vaccine developed using reverse vaccine technology were studied using Expasy ProtParam which predicted the vaccine developed as acidic, hydrophilic, and stable in nature [16]. Another study using computational tools had revealed that NCGM946K2_146 protein with 455 amino acids was acidic and unstable in nature as per the physical and chemical parameters [56]. 4TMD_A (PDB ID), a hypothetical protein of *M. smegmatis*,

showed 99.56% sequence similarity with 195 amino acids among the 248 long amino acid NHP1 sequence. Robetta server was used to predict the structure of NHP1 using 4TMD_A as model protein structure. 3D structure of a multiepitope vaccine using Ag85A (Rv3804c), Mtb32A (Rv0125), Rv2684, and Rv2608 developed previously using ITASSER showed 85.9% favored region, 8.9% allowed region, and 5.2% disallowed region in Ramachandran plot [16]. Ramachandran plot for NHP1 protein structure favored 84% of the peptide residues in the favorable region, and 8% of the residue in the allowed regions confirming the phi and psi bond angles and secondary structure prediction was complementary. The VERIFY3D server was used to compare the one-dimensional and three-dimensional structure of the predicted protein, giving a result of 80.02%. GalaxyWEB Refine server was used to refine the predicted model structure, allowing a 3% refinement further showing 95.1% residues in the Ramachandran plot to be favorable. If at least 90% of its residues lie in the favored region, the protein model can be considered reliable [58]. Since the total residues in the favored region of NHP1 model were more than 90%, it can be considered a valid structure. In another study of modeled structures, NP_218312 was docked against different ligands in search of potential drug candidates against tuberculosis [59].

Homology protein modeling has led to a new breakthrough in the field of computational biology. In homology modeling, new protein structures are designed using a known protein structure. This method is very effective in studying evolutionarily related proteins since they share common structural domains. Crystallographic studies of membrane proteins are challenging, since they can be affected by detergents and other physical and chemical properties of the protein. Total number of protein structures deposited in Protein Data Bank (PDB) until 2021 is 183,584 of which 2037 are membrane proteins. Generally, a sequence identity more than 50% can give reliable models with only slight errors in the loops and side-chain positions. Homology protein modeling is a time-saving method when compared to X-ray crystallography technique.

Cell-mediated response is significant in the progress of the disease in a patient. Bacterial membrane proteins play a significant role in the host-pathogen interaction. Hence, the protein under study offers a promise of immunogenicity and hence important as a candidate vaccine. The 3D structure of NHP1 was predicted with 99% confidence using the Robetta server. NHP1 protein has also shown 99% sequence similarity with *M. canetti*, which is an ancestral smooth tubercle bacillus. Studies on the nature of protein folds have given us insights on the presence of a flavodoxin-like fold in NHP1 pointing to the ancient nature of this mycobacterial protein, a corroboration of our earlier research on the N4.5 region's close similarity with the corresponding genomic locus in the smooth tubercle bacillus, *M. canettii*. Thus, the evidence collected through various in silico approaches to better understand NHP1 has led us to the conclusion that it is a transmembrane, signal peptide of *M. tuberculosis* with sufficient antigenicity and immunogenicity as depicted by prediction servers like VaxiJen, AntigenPro, AllerTOP, and HLAPred. An ideal vaccine to put a curb on the spread of tuberculosis should target multiple immune system components. As it is very difficult to create a vaccine with all these attributes, multiple vaccines or combination approaches would be the ideal strategy to look forward to. In this context, the current study exploring the possibility of NHP1 as a potential vaccine candidate assumes importance as it has immunogenic features comparable with those of esxA, esxB, Ag85B, etc. which are currently explored vaccine candidates.

Conclusion

The structural and functional properties of a putative protein, NHP1, in the RD7 region of the genome were studied using in silico approaches. This was found to be an ORF in a subtracted genomic region N4.5 in clinical isolates in India. The initial studies on this protein depicted it to be a novel multiepitope possible subunit vaccine candidate. NHP1, predicted to be a membrane protein with signal sequence, showed 99% similarity with the counterpart in *M. canetti* and showed the presence of flavodoxin-like folding patterns with ATP binding sites stressing the evolutionarily ancient nature of the protein. Evidence from our earlier studies of the sequences upstream and downstream of NHP1 loci has also revealed almost 100% similarity with the corresponding regions of *M. canetti*, the ancient smooth tubercle bacillus. These two facts drive home the ancient nature of this locus. This revelation is particularly important as India has a lot of clinical isolates of the pathogen falling under the category of ancestral strains. NHP1 protein showed high antigenic threshold along with epitopes of T cells and B cells making the protein a vaccine candidate similar to esxB and fbpB proteins and other proteins in the RD7 region. The comparative analysis of NHP1 protein with positive and negative controls showed that NHP1 protein has comparable immunological properties and hence is likely to have vaccine potential. Data from these computational studies have revealed NHP1 to be a membrane protein with good antigenicity and non-allergenicity. Thus, the current approach has helped in identifying a multiepitope protein present in the RD7 region of *M.*

tuberculosis clinical strains as a novel vaccine candidate using immunoinformatics tools which can be further analyzed using animal models for its efficacy.

Abbreviations

BCG: Bacille Calmette-Guérin; ESAT-6: Early secretory antigenic target-6; RD: Region of difference; HIV: Human immunodeficiency virus; TB: Tuberculosis; mce3: Mammalian cell entry; SMRT: Single-molecule, real-time; MTBC: *Mycobacterium tuberculosis* complex; GTP: Guanosine triphosphate; PCR: Polymerase chain reaction; NCBI: National Center for Biotechnology Information; BLAST: Basic local alignment search tool; ORF: Open reading frame; SMS: Sequence manipulation suite; CDD-BLAST: Conserved domain database; SMART: Simple modular architecture research tool; TMHMM: Tied mixture hidden Markov model; SOSUI: Software systems for prediction of signal peptide and membrane protein; CELLO: SubCellular localization predictor; HLA Pred: Human leukocyte antigen; MHC-I: Major histocompatibility complex-I; HTL: Helper T lymphocytes; pMHC: Peptide-bound MHC; IEDB: Immune epitope database; HMM: Hidden Markov model; PSIPRED: PSI-blast-based secondary structure prediction; SOPMA: Self-optimized prediction method with alignment; GOR IV: Garnier–Osguthorpe–Robson 4; PDB: Protein Data Bank; ProSA: Protein structure analysis; CYS: Cysteine; Asp: Aspartic acid; Glu: Glutamic acid; Arg: Arginine; Lys: Lysine; GRAVY: Grand average of hydropathicity.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43141-022-00340-5.

Additional file 1.

Acknowledgements

The authors would like to acknowledge Karnataka Science Technology Academy (KSTA), Bangalore, Karnataka, for the short-term research grant. Both authors acknowledge the Department of Life Sciences, CHRIST (Deemed to be University), Bangalore, Karnataka, India, for the research fellowship and infrastructure facilities.

Authors' contributions

Both authors participated in this study. SS conceived the idea. KPK did the analysis and interpretation of the data. SS and KPK wrote the manuscript and made substantial corrections. Both authors have read and approved the final manuscript.

Funding

The short-term research grant provided by the Karnataka Science and Technology Academy (KSTA) was used to carry out the study.

Availability of data and materials

All the data and material generated and analyzed in this study have been included in this manuscript

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 7 September 2021 Accepted: 30 March 2022
Published online: 08 April 2022

References

1. Chakaya J, Khan M, Ntoumi F, Aklillu E, Fatima R, Mwaba P, Kapata N, Mfinanga S, Hasnain SE, Katoto PDMC, Bulabula ANH, Sam-Agudu NA, Nachega JB, Tiberi S, McHugh TD, Abubakar I, Zumla A (2021) Global Tuberculosis Report 2020–reflections on the Global TB burden, treatment and prevention efforts. Int J Infect Dis 11(21):1201–9712. https://doi.org/10.1016/j.ijid.2021.02.107
2. Andersen P, Doherty TM (2005) The success and failure of BCG—implications for a novel tuberculosis vaccine. Nat Rev Microbiol 3(8):656–662 https://doi.org/10.1038/nrmicro1211
3. Ndiaye B, Thienemann F, Ota M, Landry B, Camara M, Dièye S, Esmail H, Goliath R, Huygen K, January V, Ndiaye I, Qni T, Raine M, Romano M, Satti I, Sutton S, Thiam A, Wilkinson KA, Mboup S, Wilkinson RJ, Mcshane H (2015) MVA85A 030 trial investigators safety, immunogenicity, and efficacy of the candidate tuberculosis vaccine MVA85A in healthy adults infected with HIV-1: a randomised, placebo-controlled, phase 2 trial. Lancet Respir Med 3(3):190–200. https://doi.org/10.1016/S2213-2600(15)00037-5 Epub 2015 Feb 26
4. Mortimer TD, Weber AM, Pepperell CS (2018) Signatures of selection at drug resistance loci in *Mycobacterium tuberculosis*. mSystems 3(1):e00108–e00117. https://doi.org/10.1128/mSystems.00108-17
5. Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CR, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 396(6707):190–190 https://doi.org/10.1038/31159
6. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A 99(6):3684–3689. https://doi.org/10.1073/pnas.052548299
7. Forrellad MA, Klepp LI, Gioffré A, Sabio Y, Garcia J, Morbidoni HR, Santangelo MDLP, Bigi F (2013) Virulence factors of the *Mycobacterium tuberculosis* complex. Virulence 4(1):3–66 https://doi.org/10.4161/viru.22329
8. Mazandu GK, Mulder NJ (2012) Function prediction and analysis of *Mycobacterium tuberculosis* hypothetical proteins. Int J Mol Sci 13(6):7283–7302 https://doi.org/10.3390/ijms13067283
9. Marmiesse M, Brodin P, Buchrieser C, Gutierrez C, Simoes N, Vincent V, Glaser P, Cole ST, Brosch R (2004) Macro-array and bioinformatic analyses reveal mycobacterial 'core'genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. Microbiology 150(2):483–496 https://doi.org/10.1099/mic.0.26662-0
10. Mustafa A (2005) Mycobacterial gene cloning and expression, comparative genomics, bioinformatics and proteomics in relation to the development of new vaccines and diagnostic reagents. Med Princ Pract 14(Suppl. 1):27–34 https://doi.org/10.1159/000086182
11. Ahmad S, El-Shazly S, Mustafa A, Al-Attiyah R (2004) Mammalian cell-entry proteins encoded by the mce3 operon of *Mycobacterium tuberculosis* are expressed during natural infection in humans. Scand J Immunol 60(4):382–391 https://doi.org/10.1111/j.0300-9475.2004.01490.x
12. Panigada M, Sturniolo T, Besozzi G, Boccieri MG, Sinigaglia F, Grassi GG, Grassi F (2002) Identification of a promiscuous T-cell epitope in *Mycobacterium tuberculosis* Mce proteins. Infect Immun 70(1):79–85 https://doi.org/10.1128/IAI.70.1.79-85.2002
13. Macalino SJY, Billones JB, Organo VG, Carrillo MCO (2020) In silico strategies in tuberculosis drug discovery. Molecules 25(3):665 https://doi.org/10.3390%2Fmolecules25030665
14. Gomez M, Johnson S, Gennaro ML (2000) Identification of secreted proteins of *Mycobacterium tuberculosis* by a bioinformatic approach. Infect Immun 68(4):2323–2327 https://doi.org/10.1128/IAI.68.4.2323-2327.2000
15. Gazi MA, Kibria MG, Mahfuz M, Islam MR, Ghosh P, Afsar MN, Khan MA, Ahmed T (2016) Functional, structural and epitopic prediction of hypothetical proteins of *Mycobacterium tuberculosis H37Rv*: an insilico approach for prioritizing the targets. Gene 591(2):442–455 https://doi.org/10.1016/j.gene.2016.06.057
16. Bibi S, Ullah I, Zhu B, Adnan M, Liaqat R, Kong W-B, Niu S (2021) In silico analysis of epitope-based vaccine candidate against tuberculosis

using reverse vaccinology. Sci Rep 11(1):1–16 https://doi.org/10.1038/s41598-020-80899-6
17. Sarojini S, Mundayoor S (2020) An ancestral genomic locus in *Mycobacterium tuberculosis* clinical isolates from India hints the genetic link with *Mycobacterium canettii*. Int. Microbiol 23(3):397–404 https://doi.org/10.1007/s10123-019-00113-0
18. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. The proteomics protocols handbook, vol 571-607 https://doi.org/10.1385/1-59259-890-0:571
19. Stothard P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28(6):1102–1104 https://doi.org/10.2144/00286ir01
20. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48(D1):D265–D268 https://doi.org/10.1093/nar/gkz991
21. El-Gebali S, Mistry J, Bateman A, Eddy S, Luciani A, Potter S, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD (2019) The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432 https://doi.org/10.1093/nar/gky995
22. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P (2000) SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 28(1):231–234 https://doi.org/10.1093/nar/28.1.231
23. Rashid M, Saha S, Raghava GP (2007) Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinform 8(1):1–9 https://doi.org/10.1186/1471-2105-8-337
24. Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 13(5):1402–1406 https://doi.org/10.1110/ps.03479604
25. Möller S, Croning MD, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17(7):646–653 https://doi.org/10.1093/bioinformatics/17.7.646
26. Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics (Oxford, England) 14(4):378–379 https://doi.org/10.1093/bioinformatics/14.4.378
27. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29(1):37–40 https://doi.org/10.1093/bib/3.3.225
28. Shen H-B, Chou KC (2009) Predicting protein fold pattern with functional domain and sequential evolution information. J Theor Biol 256(3):441–446 https://doi.org/10.1016/j.jtbi.2008.10.007
29. Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinform 8(1):1–7 https://doi.org/10.1186/1471-2105-8-4
30. Dimitrov I, Bangov I, Flower DR, Doytchinova I (2014) AllerTOP v. 2—a server for in silico prediction of allergens. J Mol Model 20(6):1–6 https://doi.org/10.1007/s00894-014-2278-5
31. Dimitrov I, Naneva L, Doytchinova I, Bangov I (2014) AllergenFP: allergenicity prediction by descriptor fingerprints. Bioinformatics 30(6):846–851 https://doi.org/10.1093/bioinformatics/btt619
32. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Consortium OSDD, Raghava GP (2013) In silico approach for predicting toxicity of peptides and proteins. PLoS One 8(9):e73957. https://doi.org/10.1371/journal.pone.0073957
33. Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinform 9(1):1–12 https://doi.org/10.1186/1471-2105-9-62
34. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinform 8(1):1–12 https://doi.org/10.1186/1471-2105-8-424
35. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M (2015) Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics 67(11):641–650 https://doi.org/10.1007/s00251-015-0873-y
36. Saha S, Raghava GPS (2004) BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. Paper presented at the International Conference on Artificial Immune Systems. https://doi.org/10.1007/978-3-540-30220-9_16.
37. Saha S, Raghava GPS (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 65(1):40–48 https://doi.org/10.1002/prot.21078
38. Jespersen MC, Peters B, Nielsen M, Marcatili P (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 45(W1):W24–W29 https://doi.org/10.1093/nar/gkx346
39. Buchan DW, Jones DT (2019) The PSIPRED protein analysis workbench: 20 years on. Nucleic Acids Res 47(W1):W402–W407 https://doi.org/10.1093/nar/gkz297
40. Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics 11(6):681–684 https://doi.org/10.1093/bioinformatics/11.6.681
41. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120(1):97–120 https://doi.org/10.1016/0022-2836(78)90297-8
42. Šali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815 https://doi.org/10.1006/jmbi.1993.1626
43. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32(suppl_2):W526–W531 https://doi.org/10.1093/nar/gkh468
44. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26(2):283–291 https://doi.org/10.1107/S0021889892009944
45. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164–170. https://doi.org/10.1126/science.1853201
46. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35(suppl_2):W407–W410. https://doi.org/10.1093/nar/gkm290
47. Ko J, Park H, Heo L, Seok C (2012) GalaxyWEB server for protein structure prediction and refinement. Nucleic Acids Res 40(W1):W294–W297 https://doi.org/10.1093/nar/gkt458
48. Al-Attiyah RA, Mustafa AS (2010) Characterization of human cellular immune responses to *Mycobacterium tuberculosis* proteins encoded by genes predicted in the RD15 genomic region that is absent in *Mycobacterium bovis BCG*. FEMS Immunol Med Microbiol 59(2):177–187 https://doi.org/10.1128/IAI.00199-08
49. Restrepo-Montoya D, Vizcaíno C, Niño LF, Ocampo M, Patarroyo ME, Patarroyo MA (2009) Validating subcellular localization prediction tools with mycobacterial proteins. BMC Bioinform 10(1):1–9 https://doi.org/10.1186/1471-2105-10-134
50. Kumar S (2020) In silico identification of novel tuberculosis drug targets in MycobacteriumtuberculosisP450 enzymes by interaction study with azole drugs. Malays J Med Health Sci 16(1):24–30
51. Shanmugham B, Pan A (2013) Identification and characterization of potential therapeutic candidates in emerging human pathogen *Mycobacterium abscessus*: a novel hierarchical in silico approach. PLoS One 8(3):e59126 https://doi.org/10.1371/journal.pone.0059126
52. Harold LK, Antoney J, Ahmed FH, Hards K, Carr PD, Rapson T, Cook GM (2019) FAD-sequestering proteins protect mycobacteria against hypoxic and oxidative stress. J Biol Chem 294(8):2903–5814 https://doi.org/10.1074/jbc.RA118.006237
53. Ilinskaya AN, Dobrovolskaia MA (2016) Understanding the immunogenicity and antigenicity of nanomaterials: past, present and future. Toxicol Appl Pharmacol 299:70–77 https://doi.org/10.1016/j.taap.2016.01.005
54. Mustafa AS (2013) In silico analysis and experimental validation of *Mycobacterium tuberculosis*-specific proteins and peptides of *Mycobacterium tuberculosis* for immunological diagnosis and vaccine development. Med Princ Pract 22(Suppl. 1):43–51 https://doi.org/10.1159/000354206

55. Elhag M, Sati AOM, Saadaldin MM, Hassan MA (2019) Immunoinformatics prediction of epitope based peptide vaccine against *Mycobacterium tuberculosis* PPE65 family protein. bioRxiv:755983 https://doi.org/10.1101/755983
56. Saikat ASM, Islam R, Mahmud S, Imran M, Sayeed A, Alam MS, Uddin M (2020) Structural and functional annotation of uncharacterized protein NCGM946K2_146 of *Mycobacterium tuberculosis:* an in-silico approach. Paper Present Multidisciplinary Digit Publishing Inst Proc. 66(1):p13 https://doi.org/10.3390/proceedings2020066013
57. Beg M, Shivangi TS, Meena L (2019) Systematic analysis to assist the significance of Rv1907c gene with the pathogenic potentials of *Mycobacterium tuberculosis H37Rv*. J Biotechnol Biomater 8(287):2. https://doi.org/10.4172/2155-952X.1000287
58. Bhagat CB, Tank SK, Dudhagara PR, Trivedi ND, Trivedi UN (2014) In silico study of target proteins for Mycobacterium tuberculosis. Am J Phytomed Clin Ther 2:455–462
59. Fiser A (2010) Template-based protein structure modeling. Methods Mol Biol 673:73–94 https://doi.org/10.1007/978-1-60761-842-3_6

Publisher's Note