International Journal of Peptide Research and Therapeutics
https://doi.org/10.1007/s10989-018-9797-3

In Silico Analysis for Determination and Validation of Iron-Regulated Protein from *Escherichia coli*

Fateme Sefid[1,2] · Armina Alagheband Bahrami[3] · Maryam Darvish[4] · Robab Nazarpour[5] · Zahra Payandeh[6]

Accepted: 8 December 2018

Abstract
The iron ion is an essential element in biological processes. Many of biological activities in cells, such as peroxide reduction, nucleotide biosynthesis, and electron transport, are helped via iron ions. Extra-intestinal localities have few iron content; so that, during the infection period, the ExPEC strain attempts to pick up iron from the host. The ireA gene is an iron-regulated gene and is involved in iron attainment in human pathogenic *E. coli* isolates. A better understanding of the essence of *ireA* as well as its role in serious E. coli infections will help to provide a new and more effective treatment for *E. coli* infections. Knowledge of the three-dimensional structure of proteins can contribute to the fraction of their function, as well as their interactions with other compounds such as ligands. In addition, rational modification and protein engineering depend on identification of their 3D structures. Thereafter, various bioinformatics tools were employed to predict their immunological, biochemical and functional properties. Our results indicated that this modeled protein form common beta barrel structures. Our immunological, biochemical and functional analysis have led us to select a region of each antigen harboring the highest immunogenic properties. Our strategy to employ 3D structure prediction and epitope prediction results could be deemed as an amenable approach for efficient vaccine design. Our strategy could pave the way for further structural, functional and therapeutic studies in the context of vaccine design investigations.

Keywords Urinary tract infections · Vaccine · Iron receptor · Bioinformatics · OMP

Introduction

Avian pathogenic *Escherichia coli* (APEC), a subgroup of extra-intestinal pathogenic *E. coli* (ExPEC) causes avian coli bacillosis and burden economic losses in the global poultry industry (Singer 2015). However, the APEC pathogenesis is not well known. Many of the virulence genes have been studied in order to identify viral agents in the APEC, consist of those that have been studied in adhesion, iron regulation, toxin/cytotoxin production and serum resistance (Singer 2015; Vincent et al. 2010). The iron ion is an essential element in the biological processes whereas many of biological activities in cells, such as peroxide reduction, nucleotide biosynthesis, and electron transport, are helped via iron ions. Extra-intestinal localities have less iron content; so that, during the infection period, the ExPEC strain attempts to pick up iron from the host (Agarwal et al. 2012). During natural infection, the onset, development and transmission of most bacterial infections depends on the ability of the invading pathogen to obtain iron from a complex circumstance (Schaible and Kaufmann 2004). Within the iron attainment, the cell should produce transmembrane receptors for siderophores that chelate iron ions (Schaible and Kaufmann 2004). There are several receptors that chelate iron ions which encoded by bacterial genes such as *chuA, SitABCD, iron, iha, iutA*, and *ireA* (Pilarczyk-Zurek et al.

✉ Zahra Payandeh
zpayandeh58@yahoo.com

1 Departeman of Medical Genetics, Shahid Sadoughi University of Medical Science, Yazd, Iran

2 Departeman of Biology, Science and Art University, Yazd, Iran

3 Department of Biotechnology, School of Advanced Technologies in Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

4 Departeman of Medical Biotechnology, Faculty of Medicine, Arak University of Medical Science, Arāk, Iran

5 Biotechnology Research Center, Tabriz University of Medical Science, Tabriz, Iran

6 Immunology Research Center, Tabriz University of Medical Sciences, Tabriz, Iran

Published online: 15 December 2018

2013). It was suggested that IreA interferes with in Fe inception and acts as an iron-regulated virulence gene in ExPEC *E. coli* that derived from human blood or urine (Russo et al. 2001). Nevertheless, its exact role in APEC strains is still unclear (Russo et al. 2001).

The ireA gene is an iron-regulated gene that involved in iron attainment in human pathogenic *E. coli* isolates, and studies affirmed the role of this protein in APEC. In addition, studies reported two new ireA functions that utilized deletion mutant. The *ireA* contribute in adhesion to the DF-1 cells. Furthermore, expression of multiple adhesion genes was tested and the results indicated that there is no significant difference between wild-type and mutated strain, suggest that in fact, the *ireA* gene affects adhesion (Tarr et al. 2000).

The IrgA Siderophore receptor has been reported to contribute to growth in the rabbit ileal loop model *in vivo* and increase the viral load in an infant mouse model, maybe play a role in colonization (Goldberg et al. 1990; Tashima et al. 1996). Studies have shown that the iron-regulated gene, *ireA* plays an important role in the adhesion of the APEC strains. The *ireA* gene also increased stress-resistance in alkaline and hyperosmolality conditions, also low temperature. Therefore, the redundancy of the siderophore receptors may reflect their multifunctional roles. The *ireA* gene was mainly distributed in the more virulent phylogenetic ECOR group B and D. In *ireA* deletion mutant, the adhesion and resistance to environmental stress in comparison to the wild-type strain were significantly reduced Implying that *ireA* is an iron-regulated gene that helps with adhesion and resistance to stress in the APEC strain DE205B (Li et al. 2016).

A better understanding of the essence of *ireA* as well as, its role in serious *E. coli* infections will help to provide a new and more effective treatment for *E. coli* infections. Knowledge of the three-dimensional structure of proteins can contribute to the fraction of their function, as well as their interactions with other compounds such as ligands (Floudas et al. 2006). In addition, rational modification and protein engineering depend on the identification of their 3D structures (Blundell et al. 1988).

The 3D protein structures can be used in drug and vaccine designs (Li et al. 2016) and conformational epitope predictions (Khalili et al. 2014). A large number of known protein sequences highlight the need to identify the tertiary protein structures compared to the insignificant number of structural annotation. The empirical determination of protein structures because of its high rate of failure is an important challenge. Since experimental determination of 3D protein structures is expensive and time-consuming, other approaches need to be considered (Floudas et al. 2006; Rahman and Zomaya 2005).

For outer membrane proteins, purification and crystallization, in addition to the common experimental determination of 3D protein structures, are further obstacles. Today, bioinformatics tools are of great interest to biologists. The prediction of the 3D protein structure is one of the broadest applications of these tools (Floudas et al. 2006).

To predict the protein structure, there are several methods and algorithms, that one of which is homology modeling. Homology modeling is an in silico method for predicting 3D protein structures based on well-known homologous protein structures as a template.

In this study, we aimed to determine the 3D structure of ireA proteins. The 3D structures of these proteins help us to predict that the linear and conformational B cell epitopes based on their sequences. According to this information, most immunogenic regions of the antigens could be identified and applied to design a multivalent vaccine associated with flexible linkers (Haddad et al. 2017). Due to the inappropriate amount of iron's free form in biological fluids, the bacteria obtain the iron in complex forms using various strategies. The immunological targeting of all involved antigens in iron metabolism is a new strategy that can block all possible mechanisms of iron attainment. This does not allow for any iron attainment mechanism to compensate the iron deficiency by conventional vaccines targeting an iron uptaking antigen.

Methods

Sequence Retrieval and Alignment

The protein sequence of IreA retrieved from NCBI at http://www.ncbi.nlm.nih.gov/protein in FASTA format (accession no. AMR36194.1) and served as a query for BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi against a non-redundant protein database. Probable putative conserved domains of the query were also searched for, at this address. Searching the template, the protein sequence as an input to PSI-BLAST against protein data bank (PDB) at http://blast.ncbi.nlm.nih.gov/Blast.cgito identifying its homologous structures (Fiser 2004; Gish 1993).

Analysis of Primary Sequence and Subcellular Localization

The online software, Protparam (Gasteiger et al. 2005) at http://expasy.org/tools/protparam.html employed for determination of properties such as molecular weight, theoretical pI, amino acid composition, instability index, aliphatic index and Etc.

Likewise, CELLO subcellular Localization predictor at http://cello.life.nctu.edu.tw/ (Yu et al. 2014) and PSLpred, a SVM based method for the subcellular localization of

prokaryotic proteins at http://crdd.osdd.net/raghava/pslpred/ employ to predict the position of vaccine candidate.

Secondary Structure Prediction

In order to prediction of protein secondary structure, two online server SOPMA at https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html and phyre2 at http://www.sbg.bio.ic.ac.uk/phyre2 applied by consensus prediction from multiple alignments (Geourjon and Deleage 1995).

Topology and Signal Peptide Prediction

Prediction of the integral transmembrane topology (TM) of a protein provides valuable information about the function of protein. TopCons (Tsirigos et al. 2015) (http://topcons.cbr.su.se/) predicts consensus topology of membrane proteins and signal peptides (SPs). This server uses different algorithms including SPOCTOPUS (predicts SPs and membrane protein topology), RHYllus, OCTOPUS (employs Hidden Markov Models in combination with artificial neural networks). TMHMM (Krogh et al. 2001) at http://www.cbs.dtu.dk/services/TMHMM/ and SPOCTOPUS at http://octopus.cbr.su.se/ predict trans-membrane helices in protein structure.

Moreover, PRED-TMBB (Bagos et al. 2004) (http://biophysics.biol.uoa.gr/PRED-TMBB/) localize transmembrane β-strands of gram-negative bacteria and also predicts the topology of the loops in protein structure. Signal P 4.1 server (Petersen et al. 2011) at http://www.cbs.dtu.dk/services/SignalP/ searches for the presence of signal peptide cleavage sites and localizes them in different organisms based on several combined artificial neural networks.

3D Structure Prediction Based on Homology Modeling and Threading

PS^2v^2 (Chen et al. 2009) (http://ps2.life.nctu.edu.tw/) is an automated homology modeling server using a combination of PSI-BLAST, IMPALA, and T-Coffee methods to select a template and align target-template. SWISS-MODEL (Schwede et al. 2003) at https://swissmodel.expasy.org/, a fully automated protein structure homology-modeling server, available on ExPASy web server, and DeepView (Swiss PDB-Viewer) program (Guex and Peitsch 1997).

LOMETS (Local Meta-Threading-Server) (Wu and Zhang 2007) a protein structure prediction server at http://zhanglab.ccmb.med.umich.edu/LOMETS/, uses ten threading programs (FUGUE, HHsearch, MUSTER, PPA, PROSPECT2, SAM-T02, SPARKSX, SP3, FFAS, and PRC) to generate the tertiary structure. Phyre 2 (Kelley et al. 2015) (Protein homology/ analogy recognition engine V2.0) at http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index is an updated version of Phyre's main server using new features including improved accuracy and applying stronger interface. Unlike Phyre which uses a profile–profile alignment algorithm, Phyre2 uses the alignment of hidden Markov models via HHsearch1 to improve the accuracy of alignment and detection rate. Furthermore, Phyre2 included Poing, a new ab initio folding simulation to predict the structure and generate a model for proteins which no homology has been detected for. Poing is also used to combine multiple templates. Distance constraints from individual models are treated as linear elastic springs. Poing then synthesizes your entire protein in the presence of these springs, whilst modeling unconstrained regions using its physics simulation.

Evaluations of Models

The quality of all 3D models was assessed via Rampage at http://mordred.bioc.cam.ac.ukrapper/rampage.php. So that, Ramachandran plots were depicted for each model (Carugo and Djinović-Carugo 2013).

Models Refinement

We applied protein structure refinement methods with the goal of taking template-based approximate models and bringing them closer to the native state (Xu and Zang 2011). To this end, we used Mod Refiner, (http://zhanglab.ccmb.med.umich.edu/ModRefiner/), is a high-resolution protein structure refiner that will significantly improve the physical quality of local structures.

Orientation of the Protein 3D Structure in Membrane

The OPM server (http://opm.phar.umich.edu/server.php) was applied for positioning the rotation of transmembrane and peripheral proteins in membranes that use the tertiary structure (PDB coordinate file) as an input. Many of the membrane-associated proteins from the PDB have already been calculated and can be found in the OPM database (Lomize et al. 2012).

Prediction of Ligand Binding Site

COFACTOR at http://zhanglab.ccmb.med.umich.edu/COFACTOR/ predicts the biological function of proteins based on their structure, sequence, and protein–protein interaction (PPI). COFACTOR is a protein threading algorithm that uses BioLiP protein function database to find the best structure matches to predict functional sites and homologies. Retrieved homology templates provide some information about function, Gene Ontology (GO), Enzyme

Commission (EC), and ligand-binding sites. For GO, more information can be obtained from UniProt-GOA by sequence and sequence-profile alignments and from STRING by protein–protein interaction inferrals (Roy et al. 2012).

Identification of Functionally and Structurally Important Residues

In order to predict the functional sites of the protein surface using InterProSurf at http://curie.utmb.edu/pattest9.html. In this regard, the Protein 3D structure served as an input file for this server (Negi et al. 2007).

Surface Accessible Pockets and Clefts Analyses

Pocket regions are defined using several online servers. The GHECOM (Grid-based HECOMi finder) server at http://strcomp.protein.osaka-u.ac.jp/ghecom/ is a program for finding multi-scale pockets on protein surfaces that use mathematical morphology. Structural pockets and cavities are related to the binding and functional regions of the proteins and nucleic acids (Kawabata 2010).

The CastP server (http://sts.bioe.uic.edu/castp/) uses the weighted Delaunay triangulation and the alpha complex for shape measurements. It provides information that identifies and measure the accessible and inaccessible pockets for proteins and other molecules (Dundas et al. 2006).

In addition, Depth (http://mspc.bii.a-star.edu.sg/tankp/help.html) is a server for calculating/predicting depth, cavity sizes, ligand binding sites and PKA. Depth measures the closest distance of a residue/atom to bulk solvent (Tan et al. 2013).

Single-Scale Amino Acid Properties Assay

According to the physicochemical properties of the protein including hydrophilicity, flexibility, accessibility, turns and antigenic propensity of the polypeptide, using two progressed servers IEDB (Vita et al. 2014) at http://tools.immuneepitope.org/tools/bcell/iedb and BCEPred (Saha and Raghava 2004) at http://webs.iiitd.edu.in/raghava/bcepred/ the position of the B cell epitopes has determined.

Prediction of Linear, Spatial Epitopes and Immunogenic Regions

The identification and characterization of B-cell epitopes play an important role in vaccine design, immunodiagnostic tests, and antibody production. Therefore, the computational tools for reliably predicting linear B-cell epitopes are very favorable. BepiPred at http://www.cbs.dtu.dk/services/BepiPred/ predicts the location of linear B-cell epitopes by combination of the hidden Markov model and the propensity scale method (Jespersen et al. 2017).

SVMTriP at http://sysbio.unl.edu/SVMTriP/prediction.php used to predict the antigenic epitope within sequence input. In this method, Support Vector Machine (SVM) has been utilized by combining the Tri-peptide similarity and Propensity scores (SVMTriP) in order to achieve the higher accuracy and specificity by leave-one-out test (Yao et al. 2012).

Protein 3D structure served as an input file for predicting B cell epitopes based on 3D structure. For prediction of discontinuous B cell epitopes, we used Discotope at http://www.cbs.dtu.dk/services/DiscoTope/. The method utilizes calculation of surface accessibility (estimated in terms of contact numbers) and a novel epitope propensity amino acid score. The final scores are calculated by combining the propensity scores of residues in spatial proximity and the contact numbers (Davies and Flower 2007).

Likewise, Ellipro server at http://tools.iedb.org/ellipro/, a method for identifying continuous epitopes in the protein regions protruding from the protein's globular surface. It is a new structure-based tool for predicting the antibody epitopes that is suggested. ElliPro executes three algorithms based on the following tasks: Estimating the protein shape as an ellipsoid, computing the residue protrusion index (PI) and clustering of neighboring residues based on their PI values (Ponomarenko et al. 2008).

Immunogenic Regions Selection

Regions with the largest collection of linear and conformational epitopes could be selected as vaccine candidates. These regions should be qualified as single-scale amino acid properties assay. Furthermore, some specifications such as probability of antigenicity, physiochemical properties average, and etc should also be considered in region selection. Therefore, two regions were selected as suitable antigenic candidates. Further analyses were conducted in selected regions to confirm the selection. The probability of antigenicity was predicted for selected regions using Vaxijen (Doytchinova and Flower 2007) at http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html.

Physiochemical properties average of the selected regions calculated by Protparam at http://expasy.org/tools/protparam.html that was employed for estimation and determination of properties such as molecular weight, theoretical pI, amino acid composition, the total number of negatively and positively charged residues, instability index and aliphatic index. All analyses performed in this section were also carried out on IreA in order to compare with those of selected regions.

Result

Sequence Availability, Homology Alignment and Template Search

The IreA protein sequence with 682 amino acid was saved in FASTA format. Protein sequences serving as a query for BLAST produced a set of sequences as the highest similar sequence. BLAST search revealed numerous hits to the IreA protein sequence. PSI-BLAST against protein data bank (PDB) results displayed several hits as homologous structures. The first hit possessing the highest score was selected as a template for homology modeling. This top hit for IreA sequence was proteins with PDB code 2HDI_A (36% identity). Putative conserved domains were detected within this sequence. Most of the sequences belong to putative iron-regulated outer membrane virulence protein super family.

This family includes Outer membrane receptor proteins, mostly Fe transport.

Primary Sequence Analysis and Subcellular Localization

The protein sequences served as input for the computation of various physical and chemical parameters. IreA possesses 682 residues. The total number of negatively charged (Asp + Glu) and positively charged (Arg + Lys) residues compute 81 and 76 respectively.

The computed parameters included the molecular weight, theoretical pI (isoelectric point), instability index, aliphatic index and grand average of hydropathicity (indicates the solubility of the proteins: positive GRAVY (hydrophobic), negative GRAVY (hydrophilic)) and Vaxijen score are summarized in Table 1. The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins. The instability index provides an estimate of the stability of the protein in a test tube. The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of all theamino acids, divided by the number of residues in the sequence. The protein sequence Subcellular localization predicted by CELLO was outer membrane with the highest reliability of 3.909. PSLpred predicted the protein sequence as outer membrane with 90.2% accuracy.

Secondary Structure Prediction

Coil, helix and strands are components constituting secondary structure of the protein. The secondary structure could be used to validate the tertiary structures. Attribution of secondary structure components in the protein is alpha helix (16.57%), extended strand (28.59%), beta turn (12.76%) and random coil (42.08%).

SOPMA confirm phyre2 Secondary structure predictions. Consensus Secondary Structure Prediction results for the protein sequence was shown in Fig. 1.

Topology and Signal Peptid Prediction

A 2D topology model of IreA was built based on predicted inside, transmembrane and outside regions of the protein (Fig. 2). This protein is composed of 22 trans membrane anti parallel β-strands. The model suggests that the protein has β-barrel structure in native form. Strands forming β-barrel are linked together through loops at the outside or turns at the inside.

TMHMM, TOPCONS and SPOCTOPUS predicte no trans membrane helixes (TMHs) all over the protein sequence. SignalP and SPOCTOPUS predict no signal peptide at the N terminal of the protein sequence.

Whole protein include two domains comprising a cork domain at N terminal of the protein and a trans membrane barrel at the C terminal. The barrel topology composed of 22 trans membrane beta sheet, 10 short periplasmic turns and 11 large extracellular loops.

Table 1 Protparam and Vaxijen primary sequence analysis

Protein name	VaxiJen score	Number of amino acids	Molecular weight	Theoretical pI	Instability index	Aliphatic index	GRAVY
IreA	0.6570	682	75291.25	6.15	32.38 (stable)	79.93	−0.487

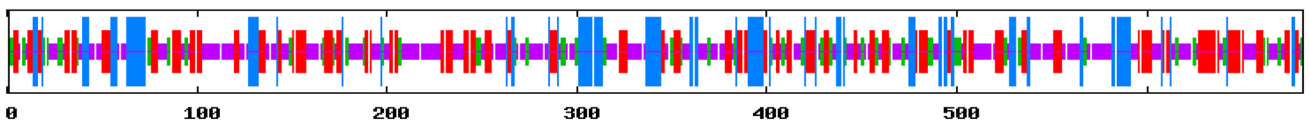


Fig. 1 SOPMA 2D structure prediction (blue: alpha helix, red: extended strand, green: beta turn, yellow: random coil). (Color figure online)

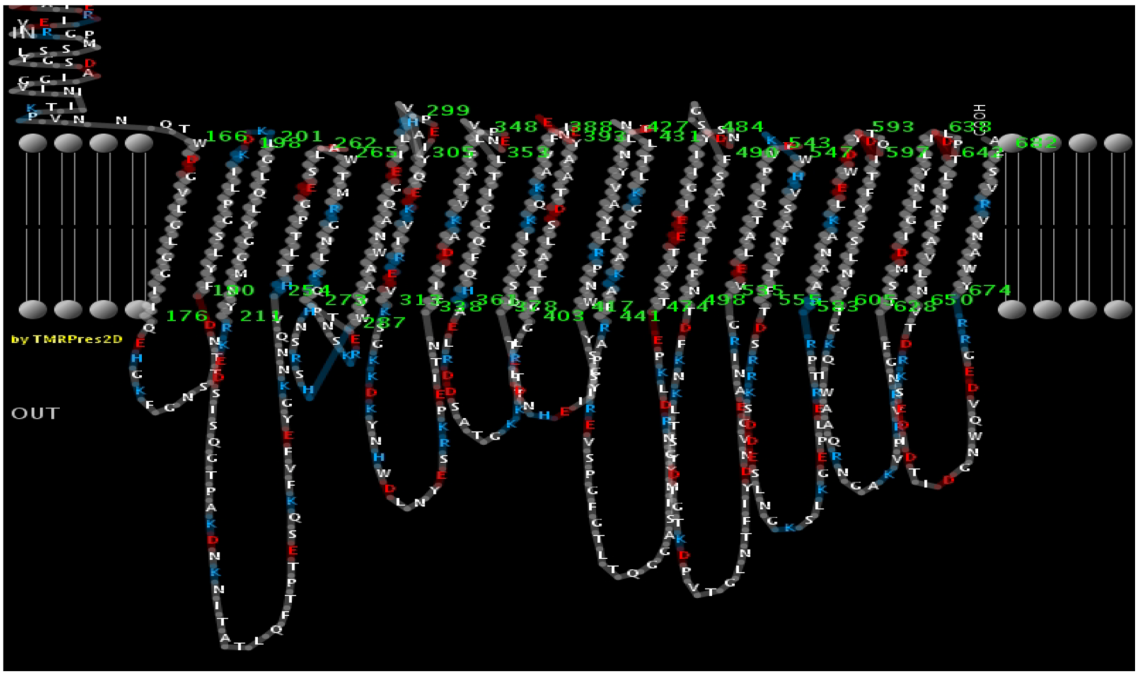


Fig. 2 A 2D topology model of IreA

3D Structure Prediction Using Homology Modeling and Threading

Swiss model and Ps^2v^2 (template-based protein structure prediction server) recruited for homology modeling identified 4 and 1 model respectively. Their models was selected for further scrutinizes. Phyre2 predicted one three-dimensional model and Lomets Meta server predicted 10 models for the protein. The models were taken for validation analyses. Some template proteins of similar folds (or super-secondary structures) were retrieved from the PDB library by LOMETS, a locally installed meta-threading approach.

Evaluations of Models

The 3D models estimated qualitatively by Rampage. To recognize the errors in the generated models, coordinates were supplied by uploading 3D structures in PDB format into Rampage, which is frequently employed in protein structure validation. In Ramachandran plot of models the percent residues were located in favored, allowed and outlier regions. The Ramachandran plot is the 2d plot of the φ–ψ torsion angles of the protein backbone. It provides a simple view of the conformation of a protein. The φ–ψ angles cluster into distinct regions in the Ramachandran plot where each region corresponds to a particular secondary structure. Amongst all the predicted models, the selected model was outstanding. With the maximum percent of favored residue (95.3%) and the minimum percent of outlier residue (0.9%). All the models validations summarized in Table 2.

Table 2 The Ramachandran plot structures validation represent the percent of residues located in favored, allowed and outlier regions

Program	Favoured region (%)	Allowed region (%)	Outlier region (%)
FFAS-3D	95.3	3.8	0.9
PRC	93.4	4.9	1.8
SP3	94.9	3.5	1.6
FFAS03	94.9	4.0	1.2
PROSPECT2	84.7	9.6	5.7
pGenTHREADER	81.0	13.2	5.7
Neff-PPAS	93.2	4.1	2.6
SPARKS-X	93.7	5.0	1.3
wdPPAS	95.0	3.8	1.2
MUSTER	93.5	4.4	2.1
Swiss model	94.1	4.6	1.2
Swiss model	89.4	8.0	2.6
Swiss model	90.4	6.3	3.3
Swiss model	90.0	6.9	3.1
PS2V2	91.8	6.5	1.8
Phyre2	90.6	4.6	4.8

Models Refinement

One model was selected as the best 3D model. This model was built by lomets meta threading server. Template based modeling (TBM) represents the most accurate method in protein structure prediction. In the contemporary TBM, multiple templates are often identified through meta threading techniques, and full-length combining the models are built by structural fragments/restraints from multiple templates.

Selected models serve as input for Mod refiner server. Mod refiner refine protein structure closer to the native while keeping the physically meaningful atomic details of local structure. The Ramachandran plot of initial and the final models after refinement are compared in Fig. 3. In initial model the percent residues in the favored region was 95.3% while 96.3% in the final model. Percent residues in allowed region for initial and final models are 3.8% and 2.8% respectively.

Orientation of the Protein 3D Structure in Membrane

The OPM database currently includes all unique structures of transmembrane protein complexes and selected monotopic, peripheral proteins and membrane-bound peptides from PDB with their calculated membrane boundaries (Fig. 4). OPM explores orientations of quaternary complexes formed by a number of interacting proteins, rather than orientations of individual subunits or domains. The precision of calculated hydrophobic thicknesses and tilt angles are ~ 1 Å and 2°, respectively, as judged from their deviations in different crystal forms of the same proteins. The fluctuations of these parameters calculated within 1 kcal/mol around the global minimum of transfer energy are usually smaller than 2 Å and 4°, respectively (± values in all tables of the database). The calculated tilt angles in homologous proteins differ by 2°–16° depending on the

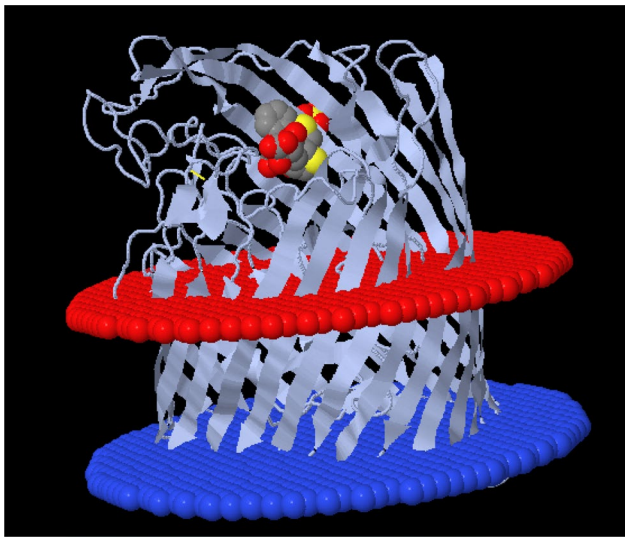


Fig. 4 Orientation of protein in cell membrane. The origin of coordinates corresponds to the center of lipid bilayer. Z axis coincides with membrane normal; atoms with the positive sign of Z coordinate are arranged in the "outer" leaflet as defined by the user-specified topology

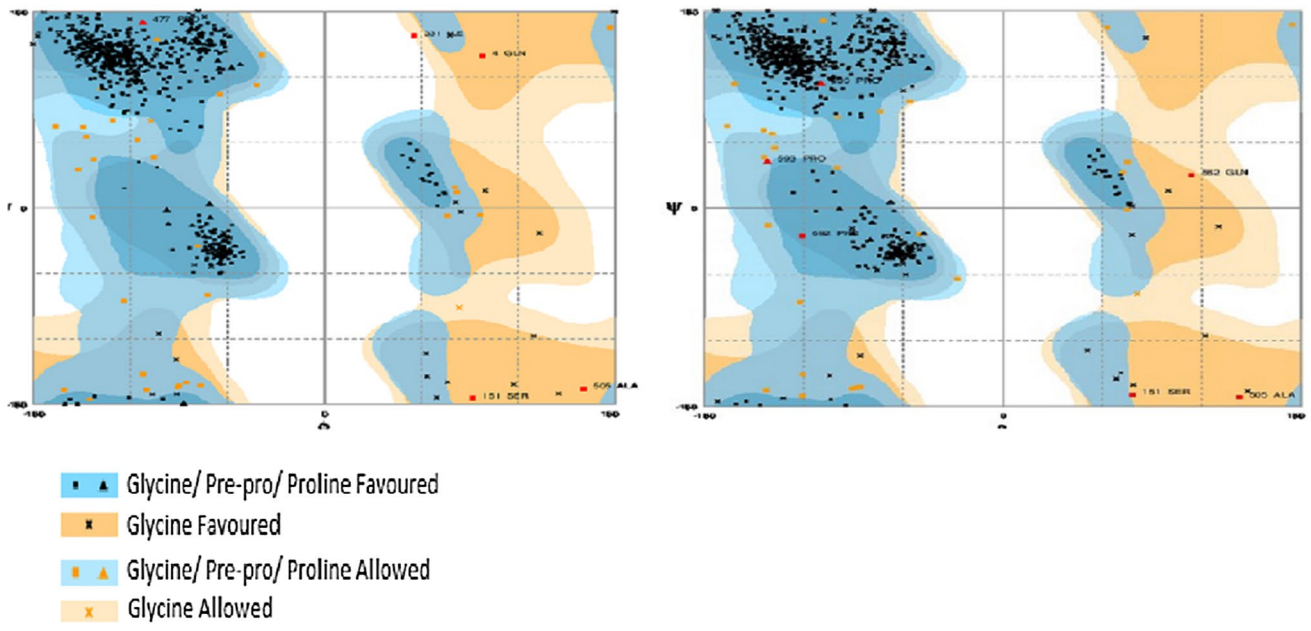


Fig. 3 Ramachandran plot of initial and the final IreA models after refinement. The percent residues in the favored region was 95.3% while 96.3% in the final model. Percent residues in allowed region for initial and final models are 3.8% and 2.8% respectively

size of the protein, its oligomeric state and percentage of sequence identity.

Ligand Binding Site Prediction

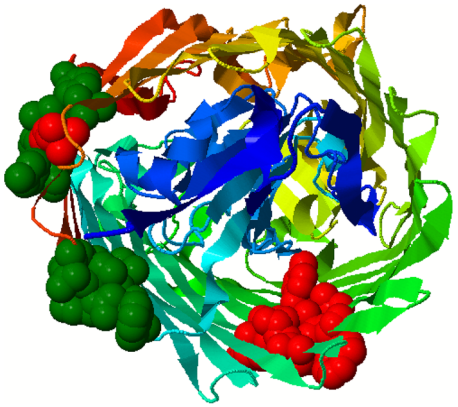
Ligand binding sites determined using COFACTOR software, indicate involvement of conserved residues between the crock domain and the large extracellular loops of barrel in iron binding site with the highest Cscore. Cscore is the confidence score of predicted binding site (Fig. 5).

Identification of Functionally and Structurally Important Residues

The protein was predicted as an iron transporter outer membrane protein. Interprosurf results show functional sites on protein structure surface. These results show that outer membrane loops and the cork domain are the most functionally site in the protein structure. Functional residues at the protein structure surface predicted by Interprosurf are shown in Fig. 6.

Surface Accessible Pockets and Clefts Analyses

GHECOM server finds five pockets on protein surfaces using mathematical morphology. In this regard, GHECOM computes a pockets score (sum of 1/[Rpocket] /(1/[Rmin] * [vol of shell])) for each residue. A residue in a deeper and larger pocket has a larger value of pockets. The pockets of small-molecule binding sites and active sites were higher than the averaged value; specifically, the values for the active sites were much higher. This suggests that pockets contribute to the prediction of binding sites and active sites from protein structures. GHECOM results are shown in Fig. 7.



Functional residues at the protein structure surface predicted by Interprosurf
243,244,245,287,289,290,291,303,304,305,597,599 ,632,634,636,646,166,168,169,170,641,680,681,682

Fig. 6 Top functional residues are highlighted in filling space model with red balls and the next top cluster highlighted in filling space model with green balls

CastP server predicts nine areas in protein structure. This server measures analytically the area and volume of each pocket and cavity, both in a solvent accessible surface (SA, Richards' surface) and molecular surface (MS, Connolly's surface). Biologically important functional residues annotated from three sources mapped to PDB structures and visualization is provided. Figure 8 shows the atoms of the charge relay system that resides in a functional pocket of the protein. The atoms of annotated residues that lie in the pocket are highlighted.

The potential binding sites (PBS) of proteins are those residues or atoms which bind to ligands directly on protein surface, they are near to the ligand binding sites. Binding cavity is a protein sub-structure of conserved geometrical

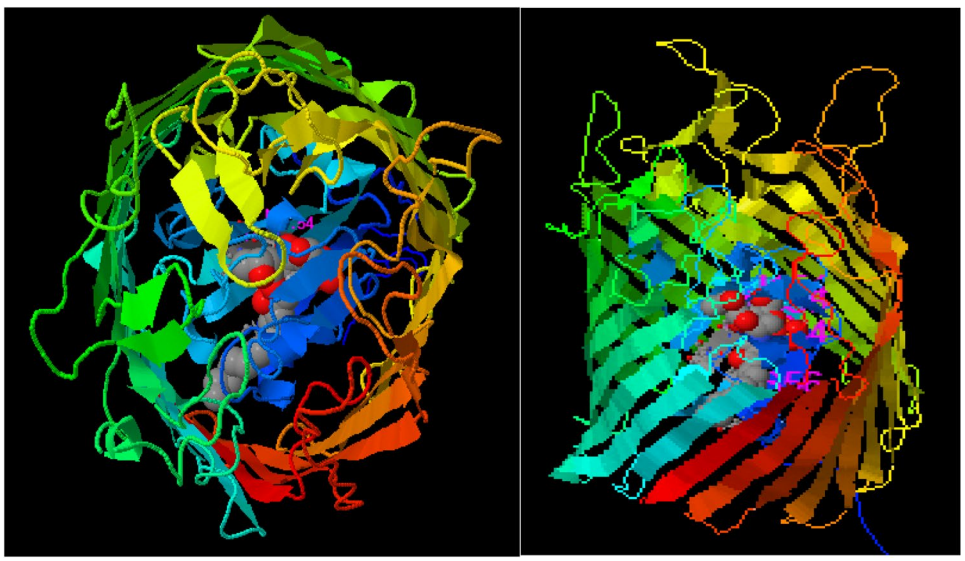
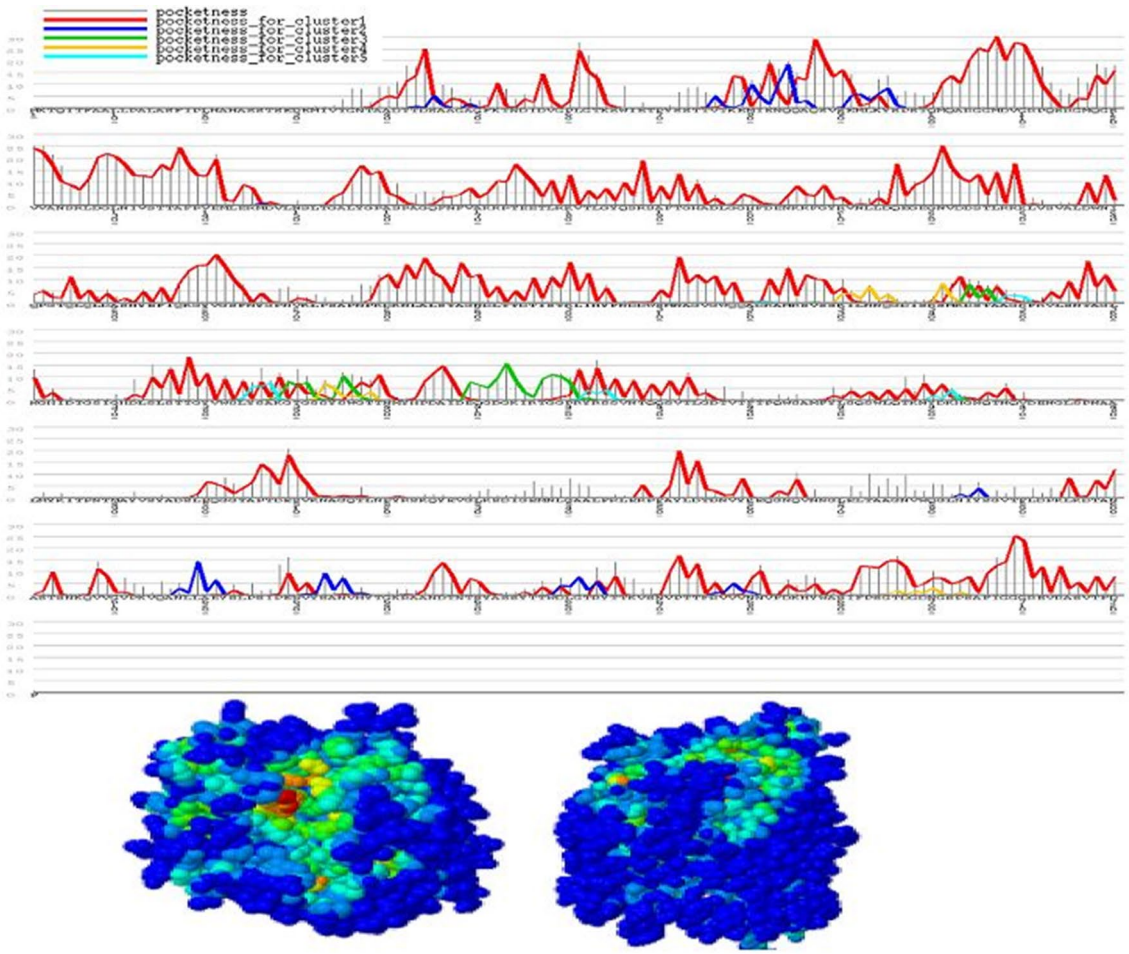


Fig. 5 IreA structure at contact with ligand from lateral and top view. The protein 3D structure is shown in ribbon and ligand in the space filling model



PocID	Area	Volume
1	3481.260	5108.663
2	1180.565	842.092
3	388.358	195.907
4	98.566	68.954
5	124.438	28.666
6	55.395	26.337
7	34.437	20.178
8	47.328	14.319
9	37.334	12.148

Fig. 7 GHECOM results showing graph residue-based pocketness and Jmol view of pocket structure. Top: graph residue-based pocketness. The height of the bar shows the value of pocketness [%] for each residue. The color of pocketness bar indicates cluster number of pocket (red: cluster 1, blue: cluster 2, green: cluster 3, yellow: cluster 4, cyan: cluster 5). Below: Jmol view of pocket structure based on pocketness color. (Color figure online)

and chemical properties complimentary to its bound ligand. Using a training-set of ligand bound high-resolution crystal structures of proteins, residue depth, and solvent-accessible area values were computed for all residues. The probability of individual amino acids to form part of the binding cavity is parametrized by the residue depth accessible area value pairs. The algorithm estimates the probability value of forming part of a binding cavity for every residue of the protein. The plot shows both mean and standard deviation of depth values. Probability of residue forming a binding site and

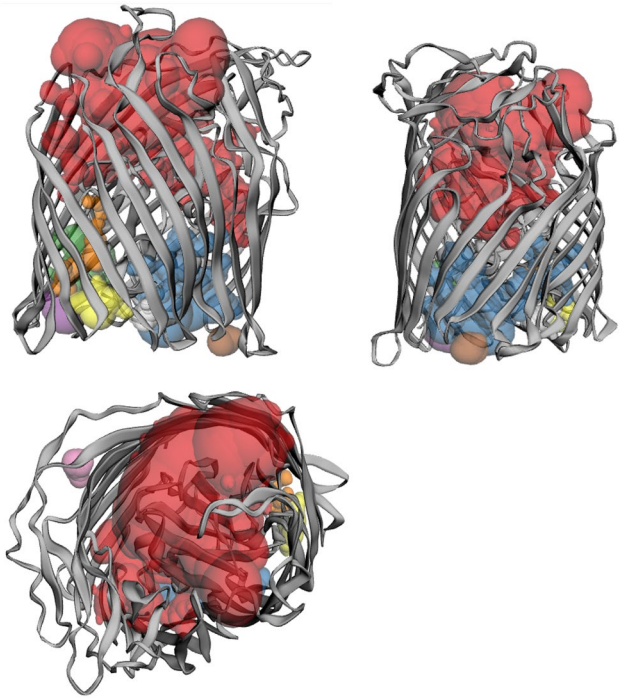


Fig. 8 CastP results showing surface accessible pockets as well as interior inaccessible cavities. Residues are colored based on area and volume size. The most important one illustrate in red and other are shown in blue, green, purple, orange, yellow, brown, pink and white respectively. (Color figure online)

residue depth plot and a 3D rendition of the cavity prediction is shown in Fig. 9.

Single-Scale Amino Acid Properties Assay

IEDB and BCEpred server predict several properties such as hydrophilicity, accessibility, antigenicity, flexibility and beta turn secondary structure in the protein sequence. Although single-scale amino acid properties were detectable in all sequence length, most saleint regions of higher probability was located in position 500–670 where a segment of beta barrel with large extracellular loops located. Peaks in the plot indicate putative susceptible epitope boundaries. Termini hits to the first 30 residues of the protein sequence were not remarkable these properties points of view. These servers predicted the most exposed amino acids in loop regions and the most buried amino acids in turns.

Prediction of Linear, Spatial Epitopes and Immunogenic Regions of Protein

Linear B cell epitops predicted by Bepipred server are more concentrated in the region of 500–600. The highest score is related to "RKSDDESLNGKSLKGEPLERTPR" sequence at position 560–582.

Svmtrip predicted 10 linear B cell epitopes ranking based on their scores. One of the best epitope recommended by this server is "LNVTDRKSEDIDTIDGNWQV" at position 649–668.

18 linear along with 5 discontinuous B cell epitopes were predicted by ElliPro software. The best discontinuous and linear epitopes with the highest PI (protrusion index) are shown in Fig. 10. "WDYTQDITF" at position 591–599 is the best linear epitope determined by Ellipro. Discontinuous B cell epitopes predicted from the 3D structure of protein by Discotope. This serve highlighted outer membrane loops as conformational B cell epitopes.

Immunogenic Regions Selection

A region covering residues 500–670 (inclusive a part of barrel) was selected as vaccine candidate and several properties were compared to its parent protein (IreA). Vaxijen antigenicity score, PI, instability index, solubility, hydrophilicity, accessibility, flexibility and secondary structure properties calculated for both vaccine candidate and IreA protein. All results for the vacine candidate and IreA protein were summarized in Table 3.

Discussion

Identification of new or undetectable bacterial agents in the pathogenesis of this infection may lead to the development of an effective vaccine or new therapies. An approach to this end is to identify genes with enhanced in vivo expression. Such genes are probably involved in pathogenesis (Mekalanos 1992). In addition, the bacterial traits that are exposed to the surface, regardless of their role in pathogenesis, are potential candidates for vaccine. Various methods have been successfully developed for this purpose (Young and Miller 1997).

An approach to identifying genes has been enhanced by ex vivo expression after exposures to body fluids (e.g., urine) or eukaryotic tissue cell culture (Zhang and Normark 1996). Further evaluations of the genes identified by this method in animal models have validated its advantage for identifying virulence traits. Using ex vivo human body fluids, we have identified a new ireA gene that has increased expression in urine, blood, and ascites. The sequencing analysis of its putative gene product showed significant identities (29–38%) and similarities (48 to 56%) with a diversity of siderophore receptors. Additionally, 5′ is suppressed into coding region for ireA, a fur box and ireA Fe expression. So, although we have no experimentally confirmation for IreA's performance, it is probably involved in Fe attainment (Russo et al. 2001).

Clearly, the presence of almost everywhere of Fe attainment systems among the human and animal pathogens has been evaluated to date, and the increase in the expression

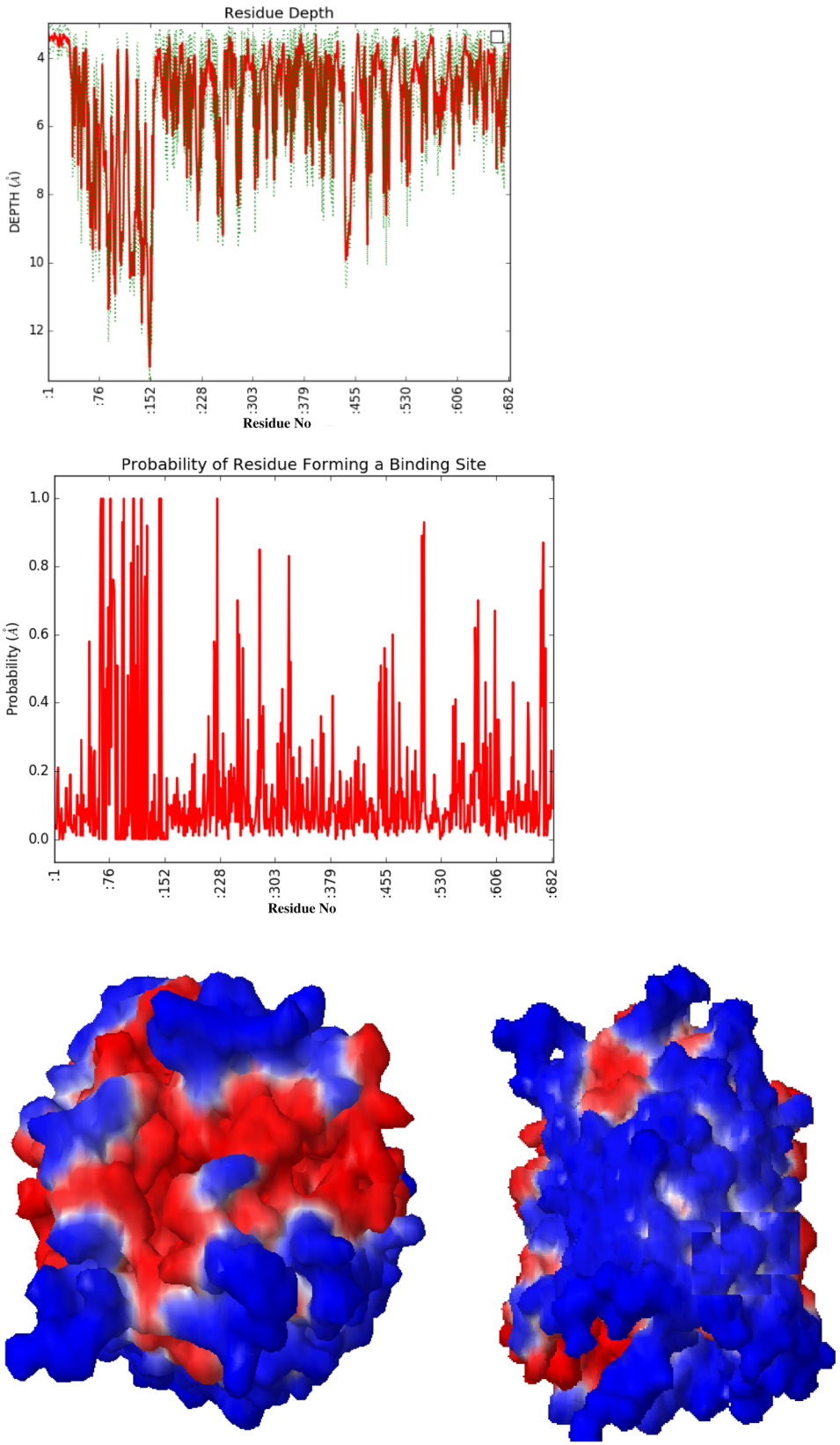


Fig. 9 Probability of residue forming a binding site and residue depth plot and a 3D rendition of the cavity prediction. Top: probability of residue forming a binding site and residue depth plot. Below: a 3D rendition of the cavity prediction is shown using Jmol. Residues of the predicted binding cavity are colored red while the rest of the protein is colored blue. (Color figure online)

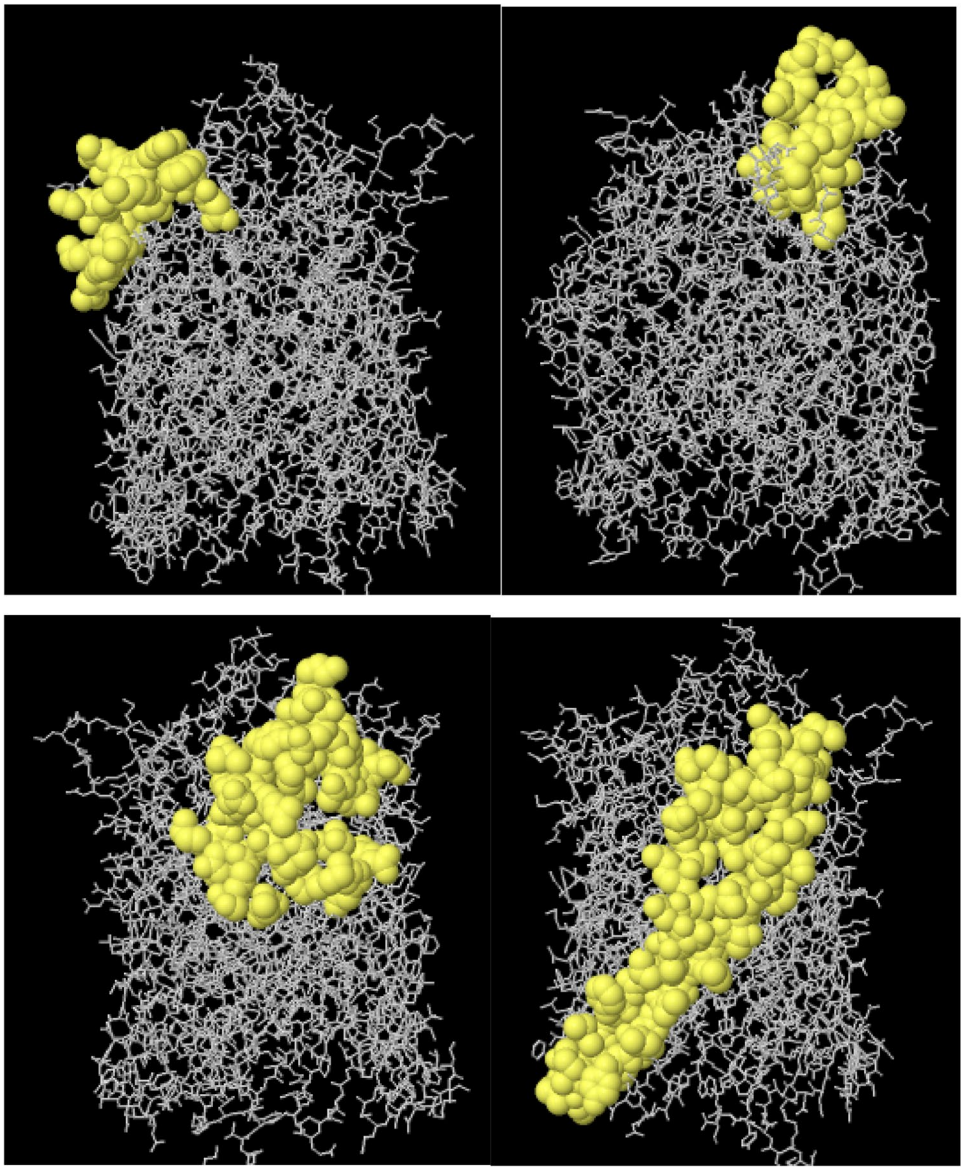


Fig. 10 The best linear (top) and discontinuous (below) epitopes with the highest PI score predicted by Ellipro server. Epitopes mapped on 3D models using Discovery Studio Visualizer 2.5.5 software

Table 3 Average physicochemical properties of two vaccine candidates and IreA

	Residue number	Weight	Vaxijen score	Instability index	pI	B cell epitope	Hydrophilicity	Flexibility	Beta turn	Accessibility
Vaccine candidate	170	19300.43	0.8625	21.12	5.20	0.526	1.964	1.0	1.027	1.00
IreA	682	75291.25	0.6570	32.38	6.15	0.514	1.963	1.0	1.018	1.00

Parameters such as hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity and antigenic propensity of polypeptides chains have been correlated with the location of continuous epitopes. This has led to a search for empirical rules that would allow the position of continuous epitopes to be predicted from certain features of the protein sequence. All prediction calculations are based on propensity scales for each of the 20 amino acids. Each scale consists of 20 values assigned to each of the amino acid residues on the basis of their relative propensity to possess the property described by the scale

of these systems in vivo or ex vivo (e.g., iroN and ireA) is strongly influenced by functional requirements for them within the host. However, it is not clear that if the defined siderophore systems are site specific or not. Test of this hypothesis pending the discovery of all Fe attainment systems for a certain pathogen, producing single and multi-isogeneic mutant for each of these and subsequent experiments in various in vitro and in vivo systems (Russo et al. 2001).

Moreover, IreA also has a significant homology with IrgA, a Fe-regulated virulence factor in *Vibrio Cholera*. IrgA contributes to in vivo growth in the rabbit ileal loop model and increases the virulence in an infant mouse model, shows the potential role in colonization (Tashima et al. 1996). Perhaps IreA and IroN also have the ability to serve as adhesins. In this case, especially if any siderophore receptor is known to have a particular ligand, the evolutionary advantage of such multifunctional proteins can be easily predicted. In this regard, our findings indicate that IreA's contribution to bladder colonization is consistent with such a role. Studies specifically designed to evaluate both IreA and IroN as adhesins are currently underway (Goldberg et al. 1990).

Our BLAST search results indicated that IreA sequences are homologous to many other molecules. Most of the sequences that obtained are belong to the TonB dependent/Ligand-Gated channels, the ligand-gated-channel protein family and the outer membrane channels superfamily. The highest homology is with 2HDI_A, is the crystalline structure of the Colicin I receptor Cir from *E. coli* in conjunction with the Receptor Binding Domain of Colicin Ia.

A successful homology modeling requires a reliable template that can be obtained by searching for similarity and sequence alignment. An acceptable template should bear low *E* value, high query coverage and high identity (more than 35%) against the target sequence. Therefore, a hit with the highest overall score can be the most reliable model for homology modeling. After modeling a protein, a model refinement run can improve the quality of predicted models. The refinement of the model can bring the initial models in terms of hydrogen bonds, backbone topology and side-chain position, closer to their native state. Two sets of criteria were considered to evaluate the results of full-atom refinement of the predicted model (Kleywegt and Jones 1998). The first set is based on the global topological similarity of the model to the experimental structure, consist of the root mean square deviation (RMSD) and a second set is template modeling (TM)-score. Specifically, the lower RMSD and higher TM-score/GDT-TS showed that the models are closer to their native state (Brillet et al. 2011).

The 2D IreA topology model is based on the predicted inside, transmembrane and outside regions of the proteins. Our results showed that this protein consists of several trans-membrane antiparallel β-strands. The predicted model shows that the protein has a β-barrel structure in its native form (Bagos et al. 2005). The strands that make the β-barrel are joined together through loops at the outside or turns at the inside. There are more than 11 external loops in these proteins, and the side chains of all residues are highly exposed to the environment, which shows its role in the initial binding events with complex Fe-siderophore.

Previously, the antigenicity and immunogenicity of the antigen have been shown to correlate directly with the epitope density (Liu and Chen 2005). Therefore, the collection of information on B-cell epitopes can play an important role in vaccine design studies, developing immunodiagnostic tests, and antibody production. The determined epitomic data can be subjected to select IreA regions with higher epitopes density. These areas can be used to design more effective immunogens that can induce humoral responses with a higher average avidity for epitope-specific mAb and polyclonal antibodies (pAb).

The best Linear B cell epitopes in the IreA protein are located in the largest extracellular loops. Interestingly, conformational B cell epitopes predict from the 3D protein structure include all of the extracellular loops. The results of epitope predictions were confirmed according to experimentally identified epitopes tested by approved antibodies. This concordance represents the precision of applied procedures for both 3D structures and epitope predictions. Comparison of antigenic scores in selected regions and the whole protein showed that the antigenicity of the selected regions is significantly higher than the total antigen (Chen et al. 2007).

In conclusion, it should be noted that the use of bioinformatics tools is a compelling strategy to close the gap between the number of protein sequences and the 3D protein structure. In silico data from the structural and immunological properties of the antigens can be used for vaccine design purposes. The design of the chimeric vaccine design from a set of immunogens can compensate for the limitations associated with antigen vaccines. Our strategy for utilizing the 3D structure prediction and the results of epitope predictions can provide a way for more structural, functional and therapeutic studies in the field of vaccine design research.

Acknowledgements The authors thank Yazd University of Medical Sciences and Tabriz University of Medical Sciences for support to conduct this work.

Compliance with Ethical Standards

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.

References

Agarwal J, Srivastava S, Singh M (2012) Pathogenomics of uropathogenic *Escherichia coli*. Indian J Med Microbiol 30(2):141

Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ (2004) PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. Nucleic Acids Res 32(suppl_2):W400–W404

Bagos PG, Liakopoulos TD, Hamodrakas SJ (2005) Evaluation of methods for predicting the topology of β-barrel outer membrane proteins and a consensus prediction method. BMC Bioinform 6(1):7

Blundell T, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T, Overington J, Singh DA, Sibanda BL, Sutcliffe M (1988) Knowledge-based protein modelling and design. FEBS J 172(3):513–520

Brillet K, Reimmann C, Mislin GL, Noël S, Rognan D, Schalk IJ, Cobessi D (2011) Pyochelin enantiomers and their outer-membrane siderophore transporters in fluorescent pseudomonads: structural bases for unique enantiospecific recognition. J Am Chem Soc 133(41):16503–16509

Carugo O, Djinović-Carugo K (2013) Half a century of Ramachandran plots. Acta Crystallogr D 69(8):1333–1341

Chen J, Liu H, Yang J, Chou K-C (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33(3):423–428

Chen C-C, Hwang J-K, Yang J-M (2009) 2-v2: template-based protein structure prediction server. BMC Bioinform 10(1):366

Davies MN, Flower DR (2007) Harnessing bioinformatics to discover new vaccines. Drug Discov Today 12(9–10):389–395

Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinform 8(1):4

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34(suppl_2):W116–W118

Fiser A (2004) Protein structure modeling in the proteomics era. Expert Rev Proteomics 1(1):97–110

Floudas C, Fung H, McAllister S, Mönnigmann M, Rajgaria R (2006) Advances in protein structure prediction and de novo protein design: a review. Chem Eng Sci 61(3):966–988

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. Humana Press, Totowa, pp 571–607

Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics 11(6):681–684

Gish W (1993) Identification of protein coding regions by database similarity search. Nat Genet 3(3):266–272

Goldberg MB, DiRITA VJ, Calderwood SB (1990) Identification of an iron-regulated virulence determinant in Vibrio cholerae, using TnphoA mutagenesis. Infect Immun 58(1):55–60

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-Pdb viewer: an environment for comparative protein modeling. Electrophoresis 18(15):2714–2723

Haddad J, Whitehead GF, Katsoulidis A, Rosseinsky MJ (2017) In-MOFs based on amide functionalised flexible linkers. Faraday Discussion 201:327–335

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

Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins Struct Funct Bioinform 78(5):1195–1211

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10(6):845

Khalili S, Jahangiri A, Borna H, Ahmadi Zanoos K, Amani J (2014) Computational vaccinology and epitope vaccine design by immunoinformatics. Acta Microbiol Immunol Hung 61(3):285–307

Kleywegt GJ, Jones TA (1998) Databases in protein crystallography. Acta Crystallogr D 54(6):1119–1131

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol 305(3):567–580

Li Y, Dai J, Zhuge X, Wang H, Hu L, Ren J, Chen L, Li D, Tang F (2016) Iron-regulated gene ireA in avian pathogenic *Escherichia coli* participates in adhesion and stress-resistance. BMC Veterinary Research 12(1):167

Liu W, Chen YH (2005) High epitope density in a single protein molecule significantly enhances antigenicity as well as immunogenicity: a novel strategy for modern vaccine development and a preliminary investigation about B cell discrimination of monomeric proteins. Eur J Immunol 35(2):505–514

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res 40(D1):D370–D376

Mekalanos JJ (1992) Environmental signals controlling expression of virulence determinants in bacteria. J Bacteriol 174(1):1

Negi SS, Schein CH, Oezguen N, Power TD, Braun W (2007) InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 23(24):3397–3399

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785

Pilarczyk-Zurek M, Chmielarczyk A, Gosiewski T, Tomusiak A, Adamski P, Zwolinska-Wcislo M, Mach T, Heczko PB, Strus M (2013) Possible role of *Escherichia coli* in propagation and perpetuation of chronic inflammation in ulcerative colitis. BMC Gastroenterol 13(1):61

Ponomarenko J, Bui H-H, Li W, Fusseder N, Bourne PE, Sette A, Peters B (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinform 9(1):514

Rahman A, Zomaya AY (2005) An overview of protein-folding techniques: issues and perspectives. Int J Bioinform Res Appl 1(1):121–143

Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40(W1):W471–W477

Russo TA, Carlino UB, Johnson JR (2001) Identification of a new iron-regulated virulence gene, ireA, in an extraintestinal pathogenic isolate of *Escherichia coli*. Infect Immun 69(10):6209–6216.

Saha S, Raghava G (2004) BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In: International conference on artificial immune systems. Springer, pp 197–204

Schaible UE, Kaufmann SH (2004) Iron and microbial infection. Nat Rev Microbiol 2(12):946

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31(13):3381–3385

Singer RS (2015) Urinary tract infections attributed to diverse ExPEC strains in food animals: evidence and data gaps. Front Microbiol 6:28

Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS (2013) Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res 41(W1):W314–W321

Tarr PI, Bilge SS, Vary JC, Jelacic S, Habeeb RL, Ward TR, Baylor MR, Besser TE (2000) Iha: a novel *Escherichia coli* O157: H7 adherence-conferring molecule encoded on a recently acquired chromosomal island of conserved structure. Infect Immun 68(3):1400–1407

Tashima KT, Carroll PA, Rogers MB, Calderwood SB (1996) Relative importance of three iron-regulated outer membrane proteins for in vivo growth of Vibrio cholerae. Infect Immun 64(5):1756–1761

Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res 43(W1):W401–W407

Vincent C, Boerlin P, Daignault D, Dozois CM, Dutil L, Galanakis C, Reid-Smith RJ, Tellier P-P, Tellis PA, Ziebell K (2010) Food reservoir for *Escherichia coli* causing urinary tract infections. Emerg Infect Dis 16(1):88

Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A (2014) The immune epitope database (IEDB) 3.0. Nucleic Acids Res 43(D1):D405–D412

Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35(10):3375–3382

Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 101(10):2525–2534

Yao B, Zhang L, Liang S, Zhang C (2012) SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PLoS ONE 7(9):e45152

Young GM, Miller VL (1997) Identification of novel chromosomal loci affecting Yersinia enterocolitica pathogenesis. Mol Microbiol 25(2):319–328

Yu C-S, Cheng C-W, Su W-C, Chang K-C, Huang S-W, Hwang J-K, Lu C-H (2014) CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. PLoS ONE 9(6):e99368

Zhang JP, Normark S (1996) Induction of gene expression in *Escherichia coli* after pilus-mediated adherence. Science 273(5279):1234–1236