Journal of Microbiological Methods 90 (2012) 36–45

Contents lists available at SciVerse ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

An in silico chimeric multi subunit vaccine targeting virulence factors of enterotoxigenic *Escherichia coli* (ETEC) with its bacterial inbuilt adjuvant

Shahram Nazarian [a], Seyed Latif Mousavi Gargari [a,*], Iraj Rasooli [a], Jafar Amani [b], Samane Bagheri [a], Masoome Alerasool [a]

[a] Department of Biology, Shahed University, Tehran-Qom Express Way, Opposite Imam Khomeini's Shrine, Tehran-3319118651, Iran
[b] Applied Biotechnology & Environmental Research Center, Baqiyatallah Medical Science University, Tehran, Iran

ARTICLE INFO

Article history:
Received 5 March 2012
Received in revised form 4 April 2012
Accepted 8 April 2012
Available online 14 April 2012

Keywords:
Colonization factors
Fusion protein
Bioinformatics
Immunogenicity
Diarrheal diseases

ABSTRACT

Enteric infections resulting in diarrheal diseases remain as major global health problems. Among bacteria, enterotoxigenic *Escherichia coli* (ETEC) causes the largest number of diarrheal cases. There is a great interest in developing an effective ETEC vaccine. An ETEC vaccine could focus on virulence factors present in ETEC pathogens and nontoxic Heat-labile B subunit (LTB). Chimeric proteins carrying epitopes, or adjuvant sequences increase the possibility of eliciting a broad cellular or humoral immune response. In-silico tools are highly suited to study, design and evaluate vaccine strategies. Colonization factors are among the virulence factor studied in the present work employing bioinformatic tools. A synthetic chimeric gene, encoding CfaB, CstH, CotA, and LTB was designed. Modeling was done to predict the 3D structure of protein. This model was validated using Ramachandran plot statistics. The predicted B-cell epitopes were mapped on the surface of the model.

Validation result showed that 97.2% residues lie in favored or additional allowed region of Ramachandran plot. VaxiJen analysis of the protein showed high antigenicity. Linear and conformational B-cell epitopes were identified. The identified T-cell epitopes are apt to bind MHC molecules. The epitopes in the chimeric protein are likely to induce both the B-cell and T-cell mediated immune responses.

1. Introduction

Enteric infections are major global health problems. Of 2–4 billion diarrheal episodes in developing countries, more than 2 million deaths per year take place among children under 5 years (Kosek et al., 2003; Sommer et al., 2010; Yano et al., 2007). Diarrheal diseases are also major health problems threatening travelers to Africa, Asia and Latin America (Qadri et al., 2005; Svennerholm and Qadri, 2008). Bacterial watery diarrheal diseases as a result of enterotoxins constitute almost half of all diarrheas. Enterotoxigenic *Escherichia coli* (ETEC) is responsible for the largest number of cases and is the most common cause of noninflammatory diarrhea in the developing world (Qadri et al., 2005; Svennerholm and Qadri, 2008). 280–400 million diarrheal episodes of ETEC origin affect children under 5, and 100 million cases are seen in children aged 5–14 years per year (Svennerholm and Qadri, 2008). Higher risk of acquiring ETEC infections is expected in malnourished children (Fleckenstein et al., 2010). Heat-labile [LT] and/or a heat-stable [ST] enterotoxin, one or more colonization factors (CFs), are the major virulence mechanisms in ETEC (Svennerholm and Tobias, 2008). LT is a heterohexameric molecule composed of a pentameric B subunit and a single A subunit. Increased levels of intracellular cAMP are the result of enzymatically active A1 portion of the subunit A. This brings about activation of cystic fibrosis transmembrane regulator (CFTR) chloride channel. This process is followed by the secretion of electrolytes and water leading to diarrhea (Sears and Kaper, 1996). The LTB consists of five identical 103 amino acid peptide responsible for binding to its GM1gangliosides receptors on eukaryotic cells surface (Fingerut et al., 2005). This non-toxic subunit induces both systemic and mucosal antibody responses. It is a suitable candidate potent immunogenic antigen to provide anti-LT immunity. Colonization factors, either non-fimbrial, fimbrial, helical or fibrillar mediate ETEC engagement with epithelial cells of the small bowel thus allowing expression of either or both of LT and ST in close proximity to the intestinal epithelium (Kosek et al., 2003; Sommer et al., 2010; Yano et al., 2007).

Of 25 CFs identified so far, seven i.e. CFA/I, CS1, CS2, CS3, CS4, CS5 and CS6 account for 50–70% of all clinical ETEC isolates (Petri et al., 2008). An ETEC vaccine could contain fimbrial antigens for the proven efficacy of CF antigens and the LT antigen, particularly on LT + ST or ST alone producing strains, in order to provide broad-spectrum protection. Thus, "a multivalent ETEC vaccine containing CFA/I, CS1–6 and an LT may provide protection against ca 80% of ETEC strains world-wide" (Svennerholm and Tobias, 2008).

Because live attenuated or killed bacterial cell vaccines have potential for transient side-effects, the use of subunit vaccines rather

* Corresponding author. Tel.: +98 21 51212600; fax: +98 21 51212601.
E-mail addresses: nazarian@shahed.ac.ir (S. Nazarian), slmousavi@shahed.ac.ir (S.L. Mousavi Gargari).

0167-7012/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2012.04.001

than those based on whole bacterial cells has been promoted (Baxter, 2007).

Subunit vaccines are considerably safer, more specific with less adverse reactions, the ability to target the site where immunity is required, and large scale production of recombinant proteins by biotechnological revolution (Baxter, 2007; Soria-Guerra et al., 2011). More complete protective response is generated by multi-component vaccine than a single component. Hence, more efficient immunoprotection could be achieved against ETEC infection by multisubunit vaccine. The identification of vaccine subunits is often a lengthy process and bioinformatic approaches have recently been used to identify candidate protein vaccine antigens. Such methods ultimately offer more rapid advance towards preclinical studies with vaccines. Comparative structural and immunological analysis of antigens could lead to the judicious selection of a combination of immunogens for a multi subunit chimeric vaccine (Jacob and Vert, 2008). In the present study, we designed a novel multi subunit antigen that provides a suitable and safe vaccine candidate against ETEC infection. A chimeric protein containing the major subunit from ETEC virulence factors of CFA/I, CS2 and CS3 and LTB was constructed together. Finally, structure of the chimeric protein was analyzed through an in silico approach. The results are discussed in the following paragraphs.

2. Methods

2.1. Sequences, databases and construct design

Major subunit genes with structural and functional properties in CFA/I, CS2 and CS3 colonization factors are *cfaB*, *cotA* and *cstH*. These major subunits and B subunit from LT toxin were selected for the current study. Related sequences were obtained from publicly available sequence databases, primarily from the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov). The sequences were retrieved in FASTA format for analysis. Multiple alignments were carried out using clustalw2 software of the European Bioinformatics Institute website (http://www.ebi.ac.uk/Tools/clustalw2). Coordinates for model building were taken from protein data bank library. Gene order in the gene cassette was optimized from antigenicity perspective. The sequences were fused together by amino acid linker in order to find the best epitope exposing chimeric antigen.

The in silico gene analysis and multi parameter gene optimization of the synthetic chimeric gene was performed using Stand-alone Leto gene optimization software (www.entelechon.com), OPTIMIZER server (Puigbo et al., 2007), Kazusa codon usage database (http://www.kazusa.or.jp/codon), and Swissprot reverse translation online tool (http://www.bioinformatics.org/sms2/rev_trans.html). The chimeric gene was designed for cloning and expression in *E. coli*. VaxiJen server was used to predict the immunogenicity of the whole antigen and its subunit vaccine (Doytchinova and Flower, 2007).

2.2. The physico-chemical parameters

The physico-chemical parameters, theoretical isoelectric point (pI), molecular weight, total number of positive and negative residues, extinction coefficient, half-life, instability index, aliphatic index and grand average hydropathy (GRAVY) were computed using the Expasy's ProtParam (http://us.expasy.org/tools/ protparam.html).

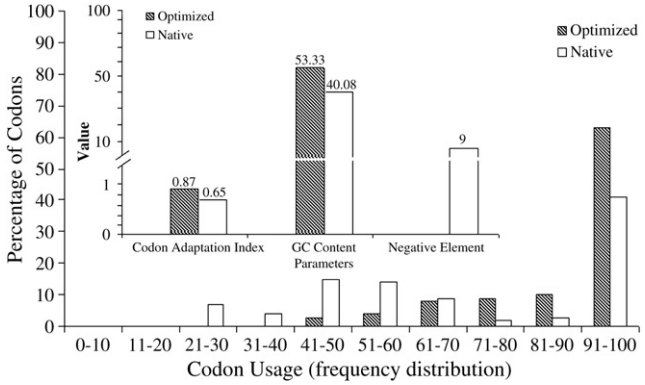


Fig. 2. Comparative analysis of various parameters employed for codon optimization. The bar graph shows frequency distribution of codon usage. The native gene sequence had 41% of codons most preferred in *E. coli*, while the optimized gene possessed 65%. The graph in the inset shows Codon adaptation index, overall GC content and CIS negative elements present in the native and codon optimized DNA sequences. All the parameters show a significant improvement over the native DNA sequence.

2.3. Prediction of antigenic B-cell epitopes

For prediction of B-cell epitopes, each full-length protein sequence was subjected to BCPreds analysis (El-Manzalawy et al., 2008) and all predicted B-cell epitopes (20 mers) having a BCPreds cutoff score >0.8 were selected. Discotope Server was used for predicting discontinuous B-cell epitopes from three-dimensional protein structures (Haste et al., 2006). Prediction of conformational B-cell epitope from primary sequence was done with web server CBTOPE (Ansari and Raghava, 2010). Surface exposed B-cell epitope sequences having the cutoff value for BCPreds (>0.8) were selected and further analyzed using VaxiJen (threshold = 0.4, ACC output) to check the antigenicity. In addition location of conformational epitopes on protein surface was defined by Episearch software (Negi and Braun, 2009).

2.4. Prediction of T-cell epitopes

Propred-1 (47 MHC Class-I alleles) (Singh and Raghava, 2003) and Propred (51 MHC Class-II alleles) (Singh and Raghava, 2001) servers that utilize amino acid position coefficients, were used to identify common epitopes that bind to both the MHC class molecules as well as to count the total numbers of interacting MHC alleles. The half maximal (50%) inhibitory concentration (IC_{50}) and antigenicity of common epitopes predicted by Propred-1 and Propred were calculated using MHCPred server (Guan et al., 2003) and VaxiJen, respectively.

2.5. Prediction of RNA secondary structure

The messenger RNA secondary structure of the chimeric gene was analyzed by Mfold web server (Zuker, 2003). RNA secondary structure was compared before and after gene optimization. The results were confirmed by other online servers such as CentroidFold Web Server (Hamada et al., 2009).

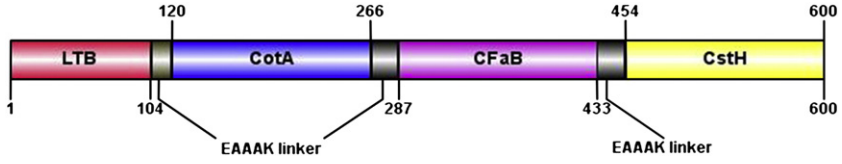
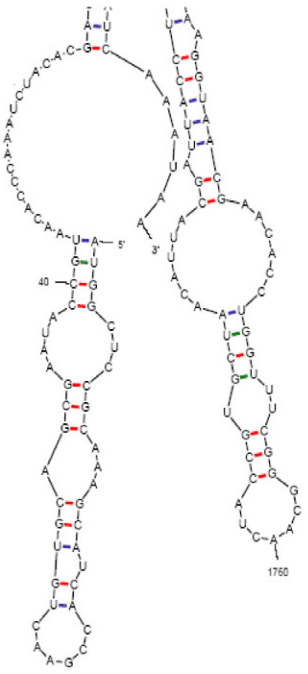


Fig. 1. Schematic representation of ETEC antigenic construct consists of *ltb*, *cotA*, *cfaB* and *cstH* genes bound together by appropriate linkers for expression in *E. coli*.

A



B

Helix	-4.80	4 base pairs.
Hairpin loop	4.90	Closing pair is A_{70}-U_{77}
Stack	-1.40	External closing pair is A_1-U_{42}
Stack	-1.40	External closing pair is U_2-G_{41}
Stack	-3.30	External closing pair is G_3-C_{40}
Helix	-6.10	4 base pairs.
Interior loop	2.70	External closing pair is G_4-C_{39}
Stack	-2.40	External closing pair is C_8-G_{34}
Stack	-3.40	External closing pair is G_9-C_{33}
Helix	-5.80	3 base pairs.
Interior loop	2.70	External closing pair is C_{10}-G_{32}
Stack	-3.40	External closing pair is G_{14}-C_{30}
Stack	-2.10	External closing pair is C_{15}-G_{29}
Helix	-5.50	3 base pairs.
Bulge loop	1.60	External closing pair is A_{16}-U_{28}
Stack	-2.10	External closing pair is C_{18}-G_{27}

C

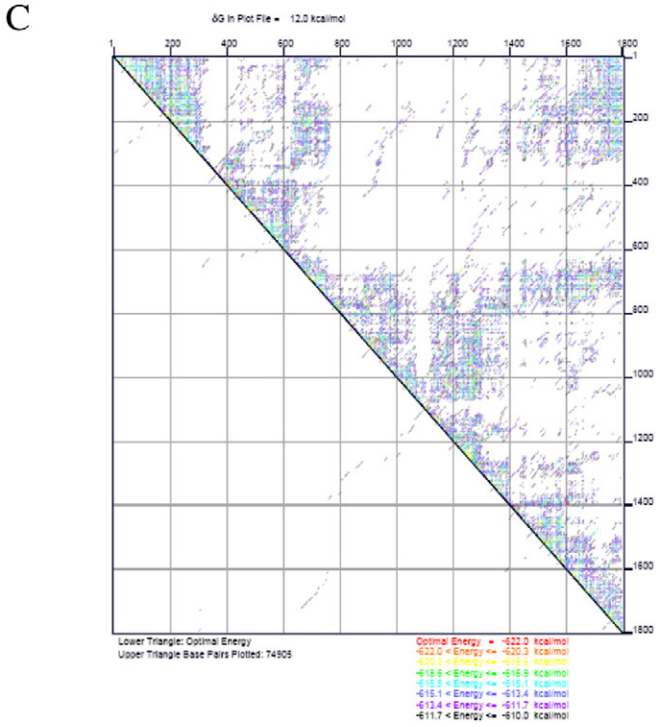


Fig. 3. Prediction of RNA secondary structure of the L2C3 chimeric gene using Mofld algorithm. Predicted structure has no long stable hairpin and pseudo knot at 5′ site of mRNA. A: The one predicted folding for the sequence. B: Free energy details for mRNA structure. This layout mimics the html output. C: The *energy dot plot* for mRNA structure. This is a triangular graph showing bases which can form complementary pairs. The colored, smaller dots represent the superposition of all possible suboptimal folding within *p%* of the minimum free energy, where *p* is the maximum percent deviation from the minimum free energy. The color ranges are red, blue and yellow; representing base pairs that are in foldings within *p*/3%, 2*p*/3% and *p*% of the minimum free energy, respectively.

2.6. Protein's secondary structure

The protein secondary structure prediction was performed by GOR secondary structure prediction method (Sen et al., 2005). PredictProtein server (Rost et al., 2004) was used for sequence analysis and the prediction of protein structure and function such as low-complexity regions, regions lacking regular structure, secondary structure, solvent accessibility, transmembrane helices, coiled-coil

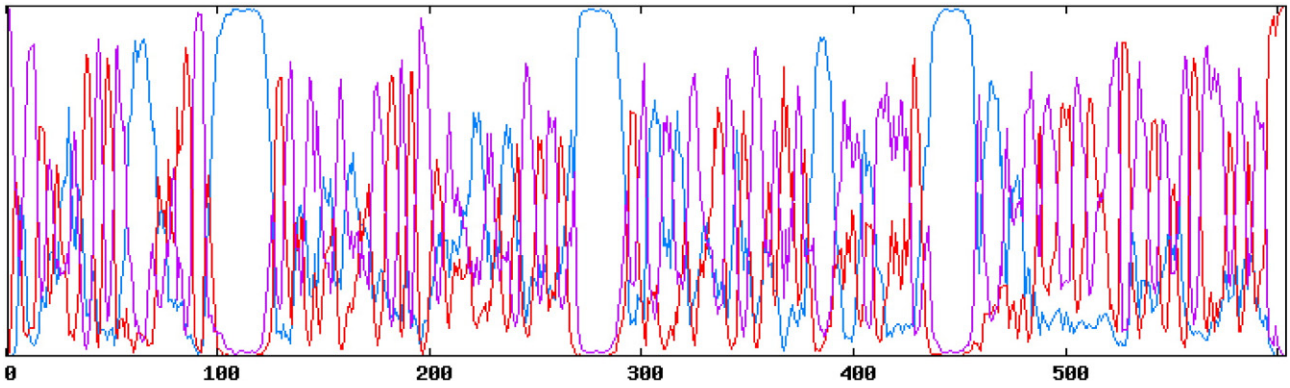


Fig. 4. Graphical results for secondary structure prediction of chimeric protein. Extended strand: purple, Coil: red, Helix: blue. Three helix peaks corresponding to the linker fragments are shown.

regions, disulfide-bonds, sub-cellular localization as well as functional annotations.

2.7. Tertiary structure prediction

For 3D structure prediction, I-TASSER ab initio online software (Zhang, 2009), the database of three-dimensional protein models calculated by comparative modeling such as SWISS-MODEL work space (Kiefer et al., 2009) were used. The Swiss model is the automated modeling software, which develops the 3D structure model of unknown structure protein based on the sequence homology with the known structured protein. The tool Rasmol and Accelrys Discovery Studio 2.5 were used to visualize the modeled 3D structures.

2.8. Tertiary structure validation

To recognize the errors in the generated models, coordinates were supplied by uploading 3D structures in PDB format into ProSA-web, which is frequently employed in protein structure validation (Wiederstein and Sippl, 2007). The structure was validated to see the quality of the resulting stereochemistry of structure by Ramachandran plot in PROCHECK software (Laskowski et al., 1993). The energy minimization of modeled structure was done by GROMOS96 implemented in SWISS-MODEL software (Christen et al., 2005). The GROMOS96 helps in minimization of bond stretch energy of the modeled protein. It incorporates both bonded and non-bonded forms of energy occupied in the protein molecule.

2.9. Ligand binding sites prediction

3DLigandSite program was used for the prediction of protein ligand binding residues in Critical Assessment of protein Structure Prediction experiment (CASP) (Wass et al., 2010). The basis for this is the identification of the binding sites using a combination of both residue conservation and the homologous structures that bound ligand location with the predicted structure of the targets. I-Tasser software was employed for predicting the binding sites on 3D structures. The software automatically defined structural analogs with binding sites similar to the model built for I-Tasser.

Table 1
Percentage of secondary structure elements of chimeric and single proteins.

Sequence	Alpha helix	Extended strand	Random coil
LTB	25.24%	33.01%	41.75%
CfaB	27.89%	23.81%	48.30%
CotA	30.61%	23.13%	46.26%
CstH	7.53%	32.19%	60.27%
L2C3	30.55%	23.37%	46.08%

2.10. Protein solubility prediction

The probability of the protein to form inclusion bodies was analyzed by pep server. Protein solubility also, was evaluated using recombinant protein solubility prediction (Davis et al., 1999).

2.11. Allergenic sites prediction

Presence of possible allergenic sites was analyzed by AlgPred. AlgPred allows prediction of allergens based on similarity of known epitope with any region of protein. The allergenicity was further analyzed by homology search in SDAP database (Ivanciuc et al., 2003) for confirmation.

3. Results

3.1. Design and construction of chimeric gene

The membrane proteins viz, CfaB, CstH,CotA and LTB, were selected for the present study. Sequence comparison by ClustalW, showed that the major subunits of colonization factors and B subunit of LT toxin were highly conserved among different strains of ETEC. Schematic diagram of protein domain structures with linker's sites designed with DOG1.0 software (Yao and Xue, 2009) is shown in Fig. 1.

Both the wild type and the synthetic chimera were analyzed for their codon bias (Fig. 2) and GC content (Fig. 2). Codon adaptation index on the native chimeric L2C3 gene was 0.65, while the optimized chimeric L2C3 gene sequence had a codon adaptation index of 0.87 (Fig. 2).

Percentage of codon having a frequency distribution of 91–100 in the native chimeric L2C3 gene was 41%, which was significantly improved to 65% in the optimized gene sequence (Fig. 2). The overall GC content, which is a measure of transcriptional and translational efficiency, was improved from 40.09% to 53.33% upon codon optimization. There were nine Negative *cis* elements in the native L2C3 gene sequence, which were removed after optimization. The *Eco*RI and *Hin*dIII restriction sites for cloning in prokaryotic vectors were successfully introduced at the N and C-terminal of sequence.

3.2. mRNA structure prediction

Combination of genetic algorithm-based RNA secondary structure prediction with sequence analysis showed that the 5′ terminus of the gene was folded in the similar pattern typical to the bacterial gene structures. The minimum free energy for secondary structures formed

Table 2
B-cell epitopes from full length proteins using BCPred (BCPred + AAP). Antigenicity of full length proteins as-well-as all B-cell epitopes was calculated using VaxiJen.

AAP predictions	Amino acid positions	BCPred scores	VaxiJen scores	BCPred predictions	Amino acid positions	BCPred scores	VaxiJen scores
ISAAPKTAGTAPTAGNYSGV	405	1	0.9483		406	1	0.7422
ATTISTDNANITLDKNAGNT	537	1	0.9158	EAAAKEAAAKAAGPTLTKEL	444	0.995	0.9133
ASVDPAIDLLQADGNALPSA	295	1	0.4584	VHTNNKTKGIQIKLTNDNVV	168	0.99	1.4987
KTIPLEVSFAGTKLSTAATS	196	1	0.9103	TQEPEAAAKEAAAKEAAAKE	263	0.984	1.1874
YSPASKTFESYRVMTQVHTN	319	1	0.3372	TTISTDNANITLDKNAGNTI	538	0.968	0.8280
INTQVHTNNKTKGIQIKLTN	164	1	1.5955	ATWAPQDNLTLSNTGVSNTL	475	0.955	0.8375
SDTSKNGTVTFAHETNNSAS	516	1	0.8992	TFKSGATFQVEVPGSQHIDS	42	0.94	0.9086
APQSITELCSEYRNTQIYTI	2	1	0.2826	SPASKTFESYRVMTQVHTND	320	0.877	0.2485
PAALDATWAPQDNLTLSNTG	470	1	0.6172	TELCSEYRNTQIYTINDKIL	7	0.842	0.6036
VQMPISVSWGGQVLSTTAKE	362	1	0.5044	TDVLNSTVQMPISVSWGGQV	355	0.804	0.4216
VSATKELVINAGSTQQTNIV	231	0.999	0.6245				
QVEVPGSQHIDSQKKAIERM	50	0.998	0.4287				
LTQEPEAAAKEAAAKEAAAK	262	0.99	1.2550				
DKILSYTESMAGKREMVIIT	23	0.967	0.1897				
ITVTASVDPTIDLMQSDGTA	124	0.802	0.8568				
NGSQLPTNLPLKFITTEGNE	562	0.437	0.9655				

by RNA molecules was also predicted (Fig. 3B). All 46 structural elements obtained in this analysis revealed folding of the RNA construct. ΔG of the best-predicted structure was −574.05 kcal/mol and the first nucleotides at 5′ did not have a long stable hairpin or pseudoknot (Fig. 3A), whereas in native mRNA, first nucleotides formed pseudoknot and the structure ΔG was −407.05 kcal/mol. The computed foldings contained 3599 base pairs out of 74,905 (4.8%) in the energy dot plot (Fig. 3C).

3.3. The physico-chemical parameters

The average molecular weight of L2C3 calculated was 62.05 kDa. Isoelectric point (pI) was defined as the pH at which the surface of protein is covered with charge but net charge of the protein is zero. Acidity of the protein was indicated by the pI value pI<7.

Extinction coefficient of L2C3 at 280 nm was 33,015 M^{-1} cm^{-1}. The biocomputed half-life was greater than 10 h. On the basis of

Table 3
Conformational B-cell epitopes from full length proteins using CBTOPE server. CBTOPE has been developed for predicting B-cell epitope from its amino acid sequence.

Amino acid	Position	Probability scale	Amino acid	Position	Probability scale
M	1	6	VE	228–9	5
A	2	5	ELVINA	236–41	4
PQSI	4–6	4	ST	243–4	4
G	34	4	QQ	245–6	5
MVIITFKSGA	38–47	4	T	247	4
TFQVE	48–52	5	NIVAG	248–252	5
VP	53–4	6	NYQG	253–6	4
G	55	8	S	259	4
S	56	7	T	263	4
QH	57–8	8	E	267	4
I	59	7	VEKNITV	287–293	4
D	60	8	KL	316–7	4
L	86	4	YR V	329–31	4
SIAAISM	96–102	4	VH	335–6	4
D	140	4	A TKKV	340–44	4
AVNI	147–150	4	Q	353	4
E	160	4	LN	358–9	4
ARIN	162–5	4	M	364	4
Q	167	4	K	380	4
V	168	5	LV	401–2	4
HTN	169–71	4	IS	403–4	5
G	176	4	AA	405–6	4
D	184	4	TFAHETN	525–31	4
N	185	5	ASFAT	534–38	4
VV	186–7	4	I STDNANITLD	540–50	4
SD	192–3	4	TI	556–7	4
PL	199–200	4	E	578	4
T	211	4	NEHLVSGN	580–87	4
AG	226–7	4	TIK	597–99	4

Table 4
Conformational B-cell epitopes from full length proteins using DiscoTope server.

Amino acid	Position	Contact number	DiscoTope score
Q	17	12	−6.208
I	18	15	−6.758
NDKILS	22–27	14-13-19-12-10-12	−6.684, −6.415, −4.530, −6.778, −6.061, −6.541
TES	29–31	10-11-10	−5.512, −6.242, −5.775
K R	35–36	16-13	−7.306, −5.884
P	54	10	−7.306
HIDS	58–61	18-13-18-25	−7.533, −5.191, −3.909, −7.015
KAIERM K	64–70	20-23-18-13-7-8-11	−6.587, −7.407, −3.683, −2.010, −1.705, −2.320, −4.067
DTLRITYL	71–78	8-9-8-9-11-11-12-20	−2.928, −3.729, −4.379, −4.141, −5.370, −5.767, −4.664, −6.912
NNK	90–92	15-12-13	−6.304, −4.865, −6.856
SA	119–120	16-12	−7.338, −4.228
T	142	12	−6.430
VHTN	168–171	21-22-20-23	−7.228, −6.161, −3.679, −5.736
SDP	192–194	11-13-13	−5.230, −6.863, −5.995
AK	270–271	12-10	−7.116, −7.562
K	276	6	−6.600
Y	329	11	−7.683
DATK	339–342	10-11-9-11	−4.514, −5.160, −4.448, −5.525
TKLA	460–464	6767	−4.944, −6.112, −6.719, −7.373
DATWAP	474–479	7-7-7-7-9-11	−7.699, −6.640, −6.000, −5.413, −7.050, −7.478
AHET	527–530	12-8-8-10	−7.399, −4.742, −4.742, −6.056
STD	541–543	11-10-11	−7.483, −5.979, −6.402
TTNGS	560–569	15-15-13-9-10	−7.643, −6.551, −5.685, −3.677, −4.119
QLPTN	565–569	8-6-14-11-13	−3.240, −2.562, −6.320, −5.285
G	586	14	−7.054
R	589	15	−7.570
N	591	14	−7.624

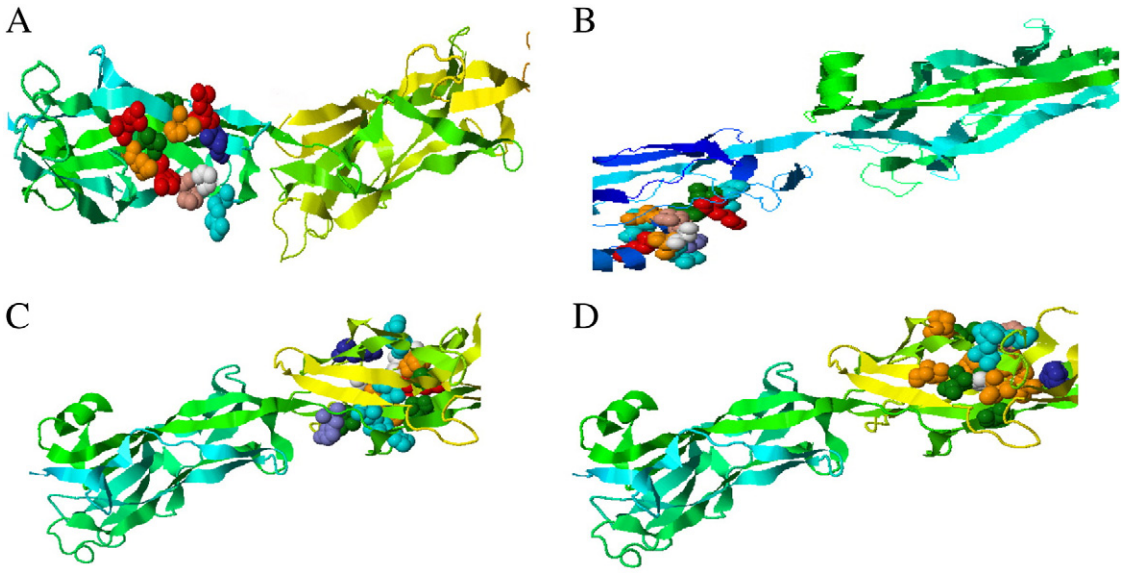


Fig. 5. Localization of potential B-cell epitopes on the protein structure. The residues predicted by the episearch method in the highest scoring patch and present in the input peptide sequences are shown. A, B, C, and D show graphical results of four top scoring patches centered on residues Q50 (Score: 0.813), N90 (Score: 0.803), T503 (Score: 0.846), and K 573 (Score: 0.884) respectively.

instability index, Expasy's ProtParam classifies the L2C3 protein as stable (Instability index<40). Aliphatic index of L2C3 was high.

3.4. Secondary structure prediction

The secondary structure prediction of the protein is shown in Fig. 4A. In order to validate our method of secondary structure prediction, first the LTB and CfaB were used as test sequences. Using the software PredictProtein and GOR IV, we obtained correctly predicted structural elements. Such prediction results lead to our confidence of the reliability of the L2C3 prediction.

The results show that random coil, extended strand and alpha helix are structural contents of protein. Composition of Predicted secondary structure for chimeric protein was 4.67% (H), 31.89% (E), and 63.44% (L) (Table 1). The total residues are made up of, 6 sheets, 7 hairpins, 1 psi loop, 1 beta bulges, 25 strands, 8 helices50 beta turns and 27 gamma turns. GOR analysis results showed three helix peaks located between positions 105–124, 267–289 and 434–453 corresponding to the linker fragments.

Other sequence analyses and the prediction of protein structure and function such as low-complexity regions (SEG), regions lacking regular structure (NORS), transmembrane helices, and coiled-coil regions showed that the sequence does not contain long regions with any regular secondary structure. Analysis of the amino acid composition demonstrated three regions with a low sequence complexity. These regions have linker sequences.

Table 5
Common epitopes from protein that can produce both the B- and T-cell mediated immunity are represented along with their various parameters.

Predicted epitopes	Amino acid	IC50 value	VaxiJen scores	Number of MHC Class I binding alleles	Number of MHC Class II binding alleles	Total number of MHC binding alleles
YRVMTQVHT	329	4.49	0.0671	5	20	25
LLQADGNAL	303	17.8	0.9200	16	4	20
GTAPTAGNY	413	19.4	0.6000	10	0	10
YRNTQIYTI	13	22.65	0.6529	8	23	31
KTKGIQIKL	173	23.50	2.8633	13	0	13
KTIPLEVSF	196	31.8	1.4034	12	1	13
TDNANITLD	542	38.99	0.9154	9	2	11
IQIKLTNDN	177	58.08	1.2649	4	23	27
WAPQDNLTL	477	94.54	0.5643	16	3	19
LTLSNTGVS	483	94.95	1.0683	7	8	15

3.5. Antigenic B-cell and T-cell epitopes

An antigen should be hydrophilic and produce both the B-cell and T-cell mediated immunity for becoming a good vaccine candidate. Therefore, full length protein was first subjected to B-cell epitope prediction using BCpreds and AAPpreds. Best epitopes were selected based on the criteria as mentioned in the methods. In general, epitopes having BCpreds, AAPpreds and VaxiJen cutoff values of >0.8, >0.8 and >0.4 respectively, were selected (Table 2). Furthermore, the conformational epitopes for B cells were predicted by the CBTOPE (Table 3) and Discotope servers (Table 4). EpiSearch method predicted four top scoring patches centered on residues Q50 (Score: 0.813), N90 (Score: 0.803), T503 (Score: 0.846), and K 573 (Score: 0.884). The patches Q50 and N90 were located in the N-terminal while the patches centered at T503 and K 573 were located in the C-terminal region of the protein (Fig. 5). Each selected B-cell epitope was analyzed for identification of T-cell epitopes within the B-cell epitope sequence. For screening T-cell epitopes, Propred-I (47 MHC Class-I alleles), Propred (51 MHC Class-II alleles), and MHCPred were used to identify common T-cell epitopes that share B-cell epitope sequence, which can interact with both MHC class I & II with the highest number (Table 5).

3.6. Tertiary structure prediction

Comparative and ab initio modeling of the synthetic sequence was exploited to produce 3D models of the chimeric protein. The 3D modeled structure for protein was generated by Swiss model and I-TASSER software. Results of tertiary structure of the fusion protein construction using I-TASSER showed a protein with four main domains linked together with linkers (Fig. 6). ProSA is a tool used to check 3D models of protein structures for potential errors. The z-score of the input structure was within the range of scores typically found for native proteins of similar size (Fig. 7). The confidence score (C-score) for estimating the quality of predicted models by I-TASSER was 0.93. C-score is typically in the range of [−5 to 2], where a C-score of higher value signifies the model with a high confidence. In addition, the expected TM-score for this model was 0.56±0.05. The expected RMSD was 2.9±042.

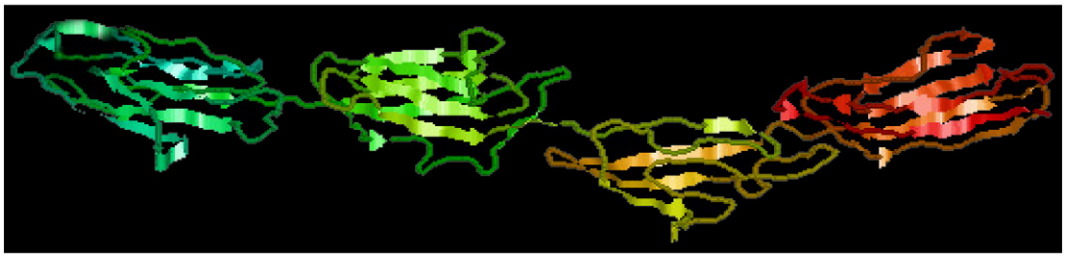


Fig. 6. Modeled structure of chimeric protein by I-TASSER software. The 3D modeled structure generated by I-TASSER software showed a protein with four main domains linked together with linkers. The structure is represented in ribbon model.

3.7. Evaluation of model stability

The Ramachandran plot analysis revealed that 84.8% of amino acid residues from modeled structure generated by I-TASSER were incorporated in the favored regions (A, B, and L) of the plot. Apart from that 9.6% of residues were in allowed regions (a, b, l,and p) of the plot. On the other hand, modeled structure by Swiss model projected 71.2% of amino acid residues in favored regions (A, B, and L) of the Ramachandran plot and 10.03% were in allowed regions (a, b, l, and p) of the plot (Table 6). Individual study of all 20 amino acid residue distribution on Ramachandran plot revealed that most of the amino acid residues were located in the shaded area (favored regions) of plot. Some residues covered the unflavored regions of the plot such as EA281, VA51, LA26, and SA512 (Fig. 8). The Ramachandran plot quality assessment analysis showed that at $2.0A^0$ most residues were within >90% (favored + allowed) regions and bad contacts were 2 residues per 100.

The profile of energy minimization calculated by Swiss-PdbViewer (19747.738 kJ/mol) indicated that the recombinant protein had acceptable stability compared to that of the original structure of each domain.

3.8. Ligand binding sites

3DLigandSite program was used for the prediction of protein ligand binding residues in Critical Assessment of protein Structure Prediction experiment (CASP). The tertiary model of the predicted protein was subjected to the more sensitive structure alignment program MAtching Molecular Models Obtained from Theory (MAMMOTH). The result identified six ligand clusters; among them, the sixth one was most significant predicting 1 ligand (n-acetyl-D-glucosamine) as well as 1 structure with average mammoth score of 37. Predicted Residues in the binding site, numbers of contacts, average distance and JS divergence were depicted in Table 7. Also JS divergence was measured in 0 to 1 scale and higher score mean more conserved residue.

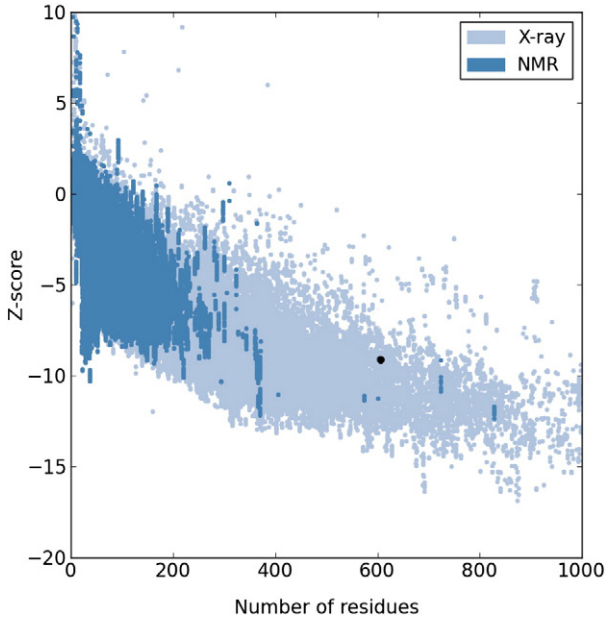


Fig. 7. ProSA-web z-score chimeric protein plot. The z-score indicates overall model quality. ProSA-web z-scores of all protein chains in PDB determined by X-ray crystallography (light blue) or NMR spectroscopy (dark blue) with respect to their length. The plot shows results with a z-score ≤ 10. The z-score L2C3 is highlighted as a large dot. The value is in the range of native conformations.

3.9. Other properties of the construct

The allergenicity of the sequence was predicted using AlgPred tool and SDAP allergen library. Based on different allergenicity prediction approaches in AlgPred tool, this protein was not detected as potential allergen. Finally, search for allergens showed no significant similarities between any region and SDAP allergen library. The construct sequence has a solubility chance of 49.5% when over expressed in *E. coli*.

4. Discussion

Vaccine candidate molecules have to be safe and immunogenic and could induce protective immunity against a broad spectrum of ETEC strains. The main strategy in the present study which was to design a chimeric vaccine as chimeric proteins carrying epitopes from different serotypes, linkers, adjuvant sequences offers not only increased immunogenicity of the recombinant antigen, but also the possibility to elicit a broad cellular or humoral immune response. Theoretically, our DNA fragment consists of four putative antigens that could be synthesized as a unique construct optimally suited for expression in *E. coli* system. Owing to this fact, heat labile enterotoxin B subunit (LTB) and three major fimbrial antigens (CfaB, CotA, CstH) present on the most prevalent ETEC pathogens were used in the present study for in silico designing of a chimeric subunit vaccine against ETEC. LTB is nontoxic subunit of LT molecule that plays an important role in ETEC virulence and pathogenesis and possesses adjuvant properties. The selection of linkers is particularly relevant in the design of functional chimeric proteins

Table 6
Comparison of Ramachandran plot statistics for modeled structure of chimeric protein from I-TASSER and Swiss model software.

Properties	I-TASSER		Swiss model	
Residues in most favored regions [A, B, L]	453	84.8%	381	71.2%
Residues in additional allowed regions [a, b, l, p]	51	9.6%	55	10.03%
Residues in generously allowed regions [~a, ~b, ~l, ~p]	15	2.8%	58	10.1%
Residues in disallowed region	15	2.8%	45	8.4%
Number of non-glycine and non-proline residues	534	100%	534	100%
Number of end-residues (excl. Gly and Pro)	2		2	
Number of glycine residues	42-30		42-30	
Number of proline residues	21		21	
Total number of residues	599		599	

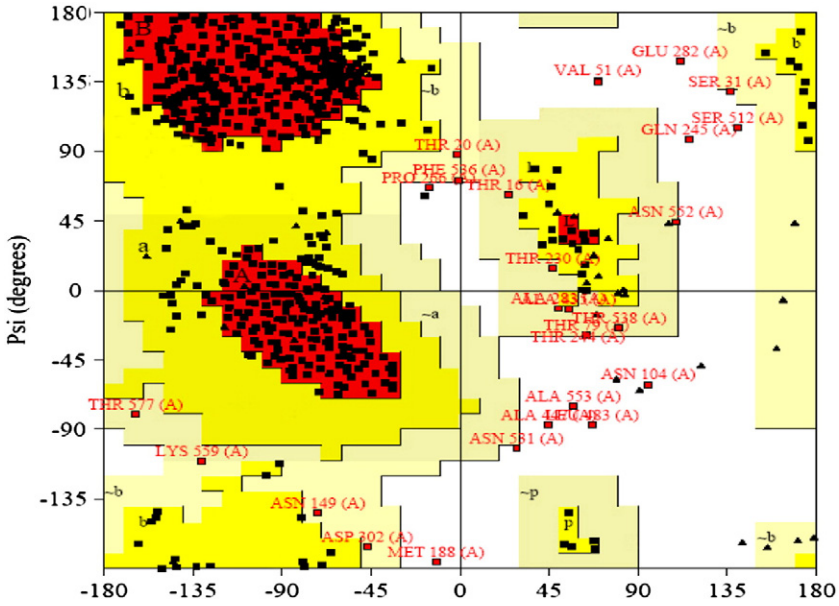


Fig. 8. Validation of protein structure using Ramchandran plot. The Ramachandran plot revealed that 84.8% of amino acid residues from modeled structure were incorporated in the favored regions (A, B, and L) of the plot. 9.6% of the residues were in allowed regions (a, b, l, and p) of the plot.

since linkers can play an important role in displaying specific epitopes in the overall structure of the fusion protein. Besides appropriate amino acid composition, the overall folding of the linker must be considered (Arai et al., 2001, 2004; Xue et al., 2004).

In order to separate different domains of our chimeric protein, linkers consisting of EAAAK repeats, expected to form a monomeric hydrophobic α-helix, were designed.

It has been shown that the salt bridge Glu^{-}-Lys^{+} between repeated Ala can stabilize helix formation (Arai et al., 2001). The synchrotron X-ray small-angle scattering experiments demonstrated that short helical linkers (n = 2, 3) cause multimerization, while the longer linkers (n = 4, 5) solvate monomeric chimeric proteins. In addition, chimeric proteins with a helical linker assumed a more elongated conformation compared to those with a flexible linker (Arai et al., 2004). Hence, these four repeated EAAAK sequences were introduced between different domains for more flexibility and efficient separation. Our successful experience of using four repeated EAAAK sequences in chimeric gene has shown that it could lead to logically acceptable results (Amani et al., 2010).

Codon optimization was performed to improve the transcription efficiency and transcript stability. This was achieved by improving the overall GC content of the gene, codon adaptation index and codon frequency distribution and removing negative elements that may form unfavorable secondary structures on mRNA. The optimized gene sequence had a codon adaptation index of 0.87, indicating that the optimized gene sequence could be expressed well.

Table 7
List of amino acid residues observed in cluster 1 of predicted protein with the number of contacts of ligand, Average distance and JS divergence.

Residue	Amino acid	Contact	Av distance	JS divergence
163	ARG	2	0.00	0.00
164	ILE	2	0.21	0.00
165	ASN	3	0.00	0.00
166	THR	2	0.52	0.00
234	THR	2	0.00	0.00
235	LYS	2	0.35	0.00
236	GLU	3	0.02	0.00
237	LEU	2	0.32	0.00

A genetic algorithm-based RNA secondary structure prediction was combined with comparative sequence analysis to determine the potential folding of the chimeric gene.

One of the software used for prediction of RNA secondary structure was Mfold. Advantages of Mfold are that it employs a theoretically tractable DP algorithm which can find the minimum ΔG structure within its thermodynamic model and high ability to predict true positive base pairs. Comparison of the synthetic gene with the original one revealed no major difference between these two molecules and their structures were compatible with each other. The data from mRNA structure prediction showed that the mRNA was stable enough for efficient translation in the host.

The study of protein secondary structure plays an important role in the prediction of protein tertiary structure with the ab initio method or protein fold recognition by providing additional constraints (Soria-Guerra et al., 2011). One of the methods used in studying secondary structure was GOR method because this method takes into account not only the probability of each amino acid having a particular secondary structure, but also the conditional probability of the amino acid assuming each structure given the contributions of its neighbors. The approach is both more sensitive and more accurate than that of Chou and Fasman because amino acid structural propensities are only strong for a small number of amino acids such as proline and glycine. In the present study, mainly GOR method has been employed.

The physico-chemical parameters of chimeric proteins were analyzed. The pI value of protein (pI < 7) showed acidic nature of the protein. Extinction coefficient of L2C3 at 280 nm was high, indicating the presence of high concentration of Cys, Trp and Tyr. On the basis of instability index, Expasy's ProtParam classifies the L2C3 protein as stable (Instability index < 40). High aliphatic index of L2C3 indicates that the protein may be stable for a wide range of temperature.

Functional characterization of a protein sequence is one of the most frequent problems in biology. This task is usually facilitated by accurate three-dimensional structure of the protein in question. The problems can be partially bypassed in comparative and ab initio protein structure prediction methods (Eswar et al., 2007; Petsko, 2006). Both comparative and ab initio methods were used for predicting three-dimensional structure of chimeric protein. LTB and CfaB have templates in the PDB library on the basis of which the 3D modeled structure of the protein was generated by Swiss model software. On

the other hand, because chimeric templates, CstH and CotA are absent from the PDB library, the models had to be built from scratch, i.e. ab initio folding. Our result showed that ab initio I-TASSER software could predict the folds and good resolution model for chimeric L2C3 protein. Analysis concluded that Zhang software predicted the 3D modeled structure of protein more accurately as compared to Swiss model. Assessment of the accuracy and reliability of experimental and theoretical models of protein structures is necessary. For the evaluation of the predicted models, both RMSD and TM-score were used. The best RMSD value was the result of superimposing our model on template which consisted of 599 amino acids; 100% of total protein residues. Expected TM-score of 0.56 ± 0.05 validates the accuracy of the n model. A TM-score >0.5 indicates a model of correct topology. Other scores including z-score and C-score suggest its confidence. The *z*-score indicates overall model quality and measures the deviation of the total energy of the structure with respect to an energy distribution derived from random conformations. *Z*-scores outside a range characteristic for native proteins indicate erroneous structures. The ProSA-web results indicate that L2C3 has features which are the characteristics of native structures.

Our chimeric structure showed desirable protein stability based on Ramachandran plot predictions. In Ramachandran plot analysis, nearly, a negligible 2.8% of the residues were found to be in outliner region that could probably be due to the presence of chimeric junctions.

To be effective, a vaccine candidate should be able to induce strong B cell and T cell responses. For this reason, the ability to map T cell and B cell epitopes is important for optimal vaccine design and development. Strategically, B-cell epitope identification can be adopted as first step in vaccine designing. BCpreds that uses a novel method of subsequence kernel was used to predict linear B-cell epitopes from each protein. The methods employed for prediction of linear epitopes led to almost identical results with some minor differences (Table 2).

The identification of conformational epitopes in antibody-antigen interaction is a crucial step for the rational design of novel drugs and vaccines (Negi and Braun, 2009). The present predicted conformational epitopes with structure-based method and sequence information methods. The results (Tables 3 and 4) show that the structural epitopes derived from CBTOPE and DiscoTope were almost similar. Some epitopes were recognized by single software only. Most of the epitopes recognized by CBTOPE had lower scores than those predicted by DiscoTope. CBTOPE predicts conformational epitopes by having amino acid sequences, while DiscoTope uses tertiary structure which seems to be more reliable.

MHCPred, Propred and Propred-1 were used for the purpose of prediction of T-cell epitopes and binding affinity of MHC molecules. The present study identified potential T-epitopes derived from antigenic B-cell epitopes of chimeric proteins. Selected T-epitopes are antigenic with potential to interact with human HLA alleles. Therefore, this chimeric protein has epitopes likely to induce both the B-cell and T-cell mediated immune responses. Finally search for allergens showed no significant similarities between any region and SDAP allergen library. Ligand binding site analysis showed that chimeric protein could be attached to n-acetyl-D-glucosamine. This means that each domain of chimeric protein has a correct and independent folding. The solubility chance of protein (49.5%) showed that it could be purified under native condition when expressed in *E. coli*. Such a protein retaining its tertiary structure could be a good candidate for immunogenic studies or trials.

5. Conclusion

Our data indicates that epitopes of the chimeric protein, L2C3, designed from colonization factors and toxin subunit of ETEC could induce both B-cell and T-cell mediated immune responses. LTB also plays adjuvant role, enhancing the immunogenicity of chimeric antigen. It can function as a carrier for mucosal delivery of the antigen. Therefore, the L2C3 with antigenic potential of the most important ETEC colonization factors is suggested as a vaccine candidate against various ETEC strains. The protein could be produced in microcapsulated form for oral immunization purpose.

Acknowledgment

The authors wish to thank Shahed University for the sanction of grants to conduct the present study. We also declare no conflict of interests.

References

Amani, J., Salmanian, A.H., Rafati, S., Mousavi, S.L., 2010. Immunogenic properties of chimeric protein from espA, eae and tir genes of *Escherichia coli* O157:H7. Vaccine 28, 6923–6929.
Ansari, H.R., Raghava, G.P., 2010. Identification of conformational B-cell epitopes in an antigen from its primary sequence. Immunome Res. 6, 1–9.
Arai, R., Ueda, H., Kitayama, A., Kamiya, N., Nagamune, T., 2001. Design of the linkers which effectively separate domains of a bifunctional fusion protein. Protein Eng. 14, 529–532.
Arai, R., Wriggers, W., Nishikawa, Y., Nagamune, T., Fujisawa, T., 2004. Conformations of variably linked chimeric proteins evaluated by synchrotron X-ray small-angle scattering. Proteins 57, 829–838.
Baxter, D., 2007. Active and passive immunity, vaccine types, excipients and licensing. Occup. Med. (Lond) 57, 552–556.
Christen, M., Hunenberger, P.H., Bakowies, D., Baron, R., Burgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Krautler, V., Oostenbrink, C., Peter, C., Trzesniak, D., van Gunsteren, W.F., 2005. The GROMOS software for biomolecular simulation: GROMOS05. J. Comput. Chem. 26, 1719–1751.
Davis, G.D., Elisee, C., Newham, D.M., Harrison, R.G., 1999. New fusion protein systems designed to give soluble expression in *Escherichia coli*. Biotechnol. Bioeng. 65, 382–388.
Doytchinova, I.A., Flower, D.R., 2007. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinforma. 8, 1–7.
El-Manzalawy, Y., Dobbs, D., Honavar, V., 2008. Predicting linear B-cell epitopes using string kernels. J. Mol. Recognit. 21, 243–255.
Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., Sali, A., 2007. Comparative protein structure modeling using MODELLER. Curr. Protoc. Protein Sci. John Wiley & Sons, Inc. Chapter 2: Unit 2.9.
Fingerut, E., Gutter, B., Meir, R., Eliahoo, D., Pitcovski, J., 2005. Vaccine and adjuvant activity of recombinant subunit B of *E. coli* enterotoxin produced in yeast. Vaccine 23, 4685–4696.
Fleckenstein, J.M., Hardwidge, P.R., Munson, G.P., Rasko, D.A., Sommerfelt, H., Steinsland, H., 2010. Molecular mechanisms of enterotoxigenic *Escherichia coli* infection. Microbes Infect. 12, 89–98.
Guan, P., Doytchinova, I.A., Zygouri, C., Flower, D.R., 2003. MHCPred: a server for quantitative prediction of peptide-MHC binding. Nucleic Acids Res. 31, 3621–3624.
Hamada, M., Kiryu, H., Sato, K., Mituyama, T., Asai, K., 2009. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics 25, 465–473.
Haste, A.P., Nielsen, M., Lund, O., 2006. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 15, 2558–2567.
Ivanciuc, O., Schein, C.H., Braun, W., 2003. SDAP: database and computational tools for allergenic proteins. Nucleic Acids Res. 31, 359–362.
Jacob, L., Vert, J.P., 2008. Efficient peptide-MHC-I binding prediction for alleles with few known binders. Bioinformatics 24, 358–366.
Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., Schwede, T., 2009. The SWISS-MODEL Repository and associated resources. Nucleic Acids Res. 37, D387–D392.
Kosek, M., Bern, C., Guerrant, R.L., 2003. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. Bull. World Health Organ. 81, 197–204.
Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291.
Negi, S.S., Braun, W., 2009. Automated detection of conformational epitopes using phage display Peptide sequences. Bioinf. Biol. Insights 3, 71–81.
Petri Jr., W.A., Miller, M., Binder, H.J., Levine, M.M., Dillingham, R., Guerrant, R.L., 2008. Enteric infections, diarrhea, and their impact on function and development. J. Clin. Invest. 118, 1277–1290.
Petsko, G.A., 2006. An introduction to modeling structure from sequence. Curr. Protoc. Bioinformatics. John Wiley & Sons, Ltd. Chapter 5: Unit 5.1., Unit.
Puigbo, P., Guzman, E., Romeu, A., Garcia-Vallvé, S., 2007. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. 35, W126–W131.
Qadri, F., Svennerholm, A.M., Faruque, A.S.G., Sack, R.B., 2005. Enterotoxigenic *Escherichia coli* in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. Clin. Microbiol. Rev. 18, 465–483.
Rost, B., Yachdav, G., Liu, J., 2004. The PredictProtein server. Nucleic Acids Res. 32, W321–W326.
Sears, C.L., Kaper, J.B., 1996. Enteric bacterial toxins: mechanisms of action and linkage to intestinal secretion. Microbiol. Rev. 60, 167–215.
Sen, T.Z., Jernigan, R.L., Garnier, J., Kloczkowski, A., 2005. GOR V server for protein secondary structure prediction. Bioinformatics 21, 2787–2788.
Singh, H., Raghava, G.P., 2001. ProPred: prediction of HLA-DR binding sites. Bioinformatics 17, 1236–1237.

Singh, H., Raghava, G.P., 2003. ProPred1: prediction of promiscuous MHC Class-I binding sites. Bioinformatics 19, 1009–1014.
Sommer, U., Petersen, J., Pfeiffer, M., Schrotz-King, P., Morsczeck, C., 2010. Comparison of surface proteomes of enterotoxigenic (ETEC) and commensal *Escherichia coli* strains. J. Microbiol. Methods 83, 13–19.
Soria-Guerra, R., Moreno-Fierros, L., Rosales-Mendoza, S., 2011. Two decades of plant-based candidate vaccines: a review of the chimeric protein approaches. Plant Cell Rep. 30, 1367–1382.
Svennerholm, A.M., Qadri, F., 2008. Mucosal immune responses against enterotoxigenic *Escherichia coli* [ETEC] in humans. In: Vajdy, M. (Ed.), Immunity Against Mucosal Pathogens. Springer, Netherlands, pp. 153–171.
Svennerholm, A.M., Tobias, J., 2008. Vaccines against enterotoxigenic *Escherichia coli*. Expert Rev. Vaccines 7, 795–804.
Wass, M.N., Kelley, L.A., Sternberg, M.J., 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res. 38, W469–W473.
Wiederstein, M., Sippl, M.J., 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 35, W407–W410.
Xue, F., Gu, Z., Feng, J.A., 2004. LINKER: a web server to generate peptide sequences with extended conformation. Nucleic Acids Res. 32, 562–565.
Yano, A., Ishimaru, R., Hujikata, R., 2007. Rapid and sensitive detection of heat-labile I and heat-stable I enterotoxin genes of enterotoxigenic *Escherichia coli* by loop-mediated isothermal amplification. J. Microbiol. Methods 68, 414–420.
Yao, X., Xue, Y., 2009. DOG 1.0: illustrator of protein domain structures. Cell Res. 19, 271–273.
Zhang, Y., 2009. Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. 19, 145–155.
Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415.