Journal of **A**pplied **B**iotechnology **R**eports

Original Article

In Silico Analysis of a Multi-subunit Immunogen, Targeting Virulence Factors of Enterohemorrhagic *Escherichia coli*

Emad Kordbacheh[1], Shahram Nazarian[1*], Milad Amerian[1]

Abstract

Enterohemorrhagic *Escherichia coli* (EHEC) strains are foodborne pathogens with importance in public health. The lack of effective clinical treatment, sequelae after infection and mortality rate in humans confirms the essential need for prophylactic and vaccines approaches. EspA, Tir, HcpA and Stx2 are major virulence factors for adherence and toxicity of EHEC, so an appropriate tetravalent immunogen consist of toxin subunit and crucial colonization factors was selected and constructed. Bio-informatic analyses of recombinant construction such as sequences choosing and optimizing, mRNA folding, physicochemical property in 2D and 3D structures, besides other immunoinformatics data like B-cell and T-cell epitopes and allergenicity of chimera were some reported according to the reliable servers. *In silico* assessment of the chimeric proteins demonstrated the desired model has a proper mRNA features, besides acceptable stability and solubility. This model is close to native proteins topologically, and all domains were found to have a high antigenic competency and surface accessibility. These results can be beneficial for the development of a chimeric immunogen against adherence and toxicity of EHEC in an animal model application.

3. Department of Biological Sciences,, Faculty of Sciences, Imam Hossein University, Tehran, Iran

*** Corresponding Author**
Shahram Nazarian
Department of Biological Sciences,, Faculty of Sciences, Imam Hossein University, Tehran, Iran
E-mail: kpnazari@ihu.ac.ir

Submission Date: 7/08/2017
Accepted Date: 10/7/2017

Keywords: Bioinformatics; Chimeric Protein; Virulence Factors; Enterohemorrhagic *E. coli*

Introduction

Enterohemorrhagic *Escherichia coli* (EHEC) is a Gram-negative bacillus, and it causes severe infectious diseases both in humans and animals [1]. EHEC is an important human pathogen causing diarrhea and in some cases hemolytic-uremic syndrome (HUS), leading to kidney failure and even death [2]. Secretion of Shiga toxins (Stxs) and formation of attaching and effacing (A/E) lesions in colonizing sites are crucial factors in the pathogenesis of EHEC infection [3]. The Stx proteins are divided into two groups according to their antigenic and genetic differences, Stx1 and Stx2. Immunization with mutant Stx1 protected mice against a challenge with lethal dose of wild-type Stx1 [4]. The Stx1B subunit conjugated with LPS elicited bactericidal antibodies to EHEC and neutralization antibodies to Stx1 in mice [5]. Nevertheless, Stx2 is more attractive than Stx1 for vaccine exploration because it is produced by nearly all the EHEC serotypes and it is associated with HUS too. Locus of enterocyte effacement (LEE) contains three major domains with the known role. One of the LEE regions encodes several proteins that are secreted via the type III secretion system (TTSS) which delivers these factors directly into the host cells, contains EspA, EspB, and EspD. These factors are essential for signal transduction in mammalian host cells and also for A/E lesion formation. EspA is a protein with a structural role and it is believed to be the major component of a large lamentous organelle. It has a transient expression on the bacterial surface and delivers EspB and EspD directly to the host cell membrane [6, 7].

During the early stage of A/E lesion formation, this protein is found to interact with epithelial cells, and also is involved in forming a bridge with bacteria surface. Through this bridge, Tir protein is transferred into the host cell and acts as a receptor for an integral outer membrane protein of EHEC called Intimin [8, 9]. Additional studies demonstrate EHEC O157:H7 can express and assemble Type 4 Pili (T4P), which is named Hemorrhagic Coli Pilus (HCP) in EHEC. These pili included a major pilin subunit encoded by the prepilin peptidase-dependent gene (ppdD) that is present in most *E. coli* strains, containing commensal and pathogenic strains.

Papers showed that HCP plays a role in adherence to cultured human colonic epithelial cells and porcine and bovine intestinal models and those HCP-specific antibodies are produced in patients with the hemolytic uremic syndrome, suggesting that HCP is produced *in vivo* [10]. So, a good candidate for vaccination against this pathogen is the right combination of different virulence factors involved in its pathogenicity. Subunit vaccines are considerably safer, more accurate with less adverse reactions, targeted immunity stimulation, and large-scale production of recombinant proteins by biotechnological revolution. In this article, we designed a novel multi-subunit antigen that provides an appropriate vaccine candidate against EHEC infection. So a new fusion respectively inclusive of HcpA-EspA-Tir-Stx2B proteins with four repeat of EAAAK rigid linker between each subunits was selected for best stability and subunit segregation. Moreover, other property of the chimeric protein structure analyzed through an *in silico*

approach and these results can be beneficial for the development of a chimeric immunogen against EHEC.

Materials and Methods

Sequence analyses and construct design

Major subunit genes (with final structural and functional properties) for HcpA, EspA, Tir and Stx2B virulence factors are hcpa, espa, tir, and stx2B, respectively. These apparatus subunits and B subunit of Stx toxin were chosen for the current study. Related sequences were selected from online sequence databases, primarily from GenBank database. The sequences were retrieved in FASTA format for analysis. Multiple alignments were carried out by using Clustal Omega tool, in the EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute) database. So the sequences were fused together by the proper rigid linker. The *in silico* gene analysis and multi-parameter gene optimization of the synthetic chimeric gene was performed using GenScript Rare Codon Analysis Tool(http://www.genscript.com/cgibin/tools/rare_codon_analysis), Rare Codon Calculator tool (http://nihserver.mbi.ucla.edu/RACC/) and WebDSV translate tool (http://www.molbiotools.com/WebDSV/). The synthetic gene was constructed for cloning and expression in *E. coli* k12. Finally, VaxiJen server was used to estimate the overall immunogenicity of the whole sequence and its protein subunits however it was utilized for specific short antigen in the following [11].

Prediction of RNA secondary structure

The messenger RNA secondary form and other different parameters of the chimeric gene were predicted by The Vienna RNA Web Servers (RNAfold) and RNA Predict Secondary Structure Server. RNA secondary structure was compared before and after gene optimization [12,13].

The physico-chemical parameters

The physicochemical parameters such as theoretical isoelectric point (pI), molecular weight (Mw), extinction coefficient, half-life, instability index, aliphatic index also grand average hydropathy (GRAVY) of each subunit, total number of positive and negative residues, beside different combinations of them were computed by using the Expasy's ProtParam tool [14]. Also Pep server analyzed the probability of the protein concerning forming inclusion bodies. Protein solubility also was evaluated using recombinant protein solubility prediction and PROSO II server [15].

Secondary and tertiary structures of Protein

Earlier the secondary structure analysis, construction of disulfide bond examined by DISULFIND server. The protein secondary structure prediction was performed by PHD secondary structure prediction method and GOR algorithm [16, 17]. Solvent accessibility, secondary structure, alpha beta transmembrane, domains and disulfide bonds were some data which obtained from Scratch protein predictor server [18]. Also, PSI-blast based secondary structure prediction (PSIPRED) program was used to searching homology alignment for the chimeric protein with its subunit, HcpA, EspA, Tir and Stx2B proteins [19]. Comparative modeling according to homology and threading, also *ab initio* modeling of the synthetic sequence was used to produce 3D models of the chimeric protein. 3D structure predicted by I-TASSER with its *ab initio* and comparative algorithm [21].

The tool Rasmol and Accelrys Discovery Studio 2.5 were used to visualize the modeled 3D structures [22, 23].

Tertiary structure validation

Besides TM-score and root mean square deviation (RMSD) which are some index for measuring the structural similarity of some 3D models topologically. To identify the errors in the produced models, coordinates were supplied by uploading PDB code or format into ProSA and ERRAT2 servers; some reliable programs for approve topology of predicted models [24, 25]. Also, previously the structure was accredited to see the quality of stereochemistry results by Ramachandran plot in RAMPAGE online software [26].

Antigenic propensity

Vaxijen server has developed to allow antigen classification based on the physicochemical properties of proteins. It was exploited initially to predict the antigenicity probability of the single and assembled proteins forms [27]. "Predicted antigenic peptides" server also was recruited on the field. This program predicts those segments from within one protein sequence that are likely to be antigenic by eliciting an antibody response. In this software, antigenic peptides are determined using the method of Kolaskar and Tongaonkar. Predictions are based on a table that reflects the occurrence of amino acid residues in experimentally known segmental epitopes.

B-cell and T-cell epitopes prediction

For prediction of B-cell epitopes, there are five or six different algorithms in the various databases such as Immune Epitope Database (IEDB) and Bcepred, which are according to hydrophilicity, flexibility, accessibility, turns, exposed surface polarity and the antigenic propensity of polypeptides chains. Full-length protein sequence was submitted to BCPreds analysis tool, and all B-cell epitopes (20 mers) which have a proper BCPreds cutoff score (>0.8) were picked up. Discotope 2.0 server was applied for predicting conformational B-cell epitopes from three-dimensional protein structures with threshold of -3.7 [28, 29]. prediction of the conformational epitope was accomplished by CBTOPE online software [30]. Concerning T-cell epitopes which are related to MHC molecules, we refer to some consensus database such as IEDB, nHLAPred, MHCPred and propred tool in imtech server [31–33]. Most of them utilize the same algorithm and will recognize similar HLA alleles. However, in this study, IEDB server was preferred as reliable software with more relevant results. This website predicts peptide binding to MHC class I and MHC class II molecules, based on IEDB prediction method. Hence among different haplotypes of the main histocompatibility molecules (MHC) in human, those of the common alleles with highest specific affinity cutoff were selected.

Allergenic sites prediction

Allergen-specific immunoglobulin E (IgE) can predict by Algpred online software in each sequence. This software allows predicting allergen using SVMc + IgE epitope + ARPs (allergen-representative peptides) BLAST + MAST

in a hybrid method [34]. Also, AllerTOP 1.0 online software was used for this purpose which the principle of this server is apart from former, and it is based on auto cross-covariance (ACC) algorithm [35].

Results

Entire EspA, HcpA and B subunit of Stx2 holotoxin without their signal peptide, furthermore, 110 amino acids from the middle of Tir protein were selected for the present study. Four repeats of an empirical linker (EAAAK) that included five amino acids with rigid structure have been chosen. To find the best subunits separation and epitope exposing we checked chimera sequence by domain linker predictor DLP and extremum of the diagram with minimum Z-score proved the ability of four repetition linker to make the appropriate distance. Schematic figure of protein substance with its linkers exhibits by a sequence editor tool like WebDSV which shown in Fig. 1. Codon adaptation index or CAI for this native fusion gene was 0.75, and for the optimized cassette it turned to 0.90, both of the native and the synthetic types were analyzed for their GC content and codon bias. GC content is necessary for transcription and translation operations.

By accomplished optimization, the GC content increased from 49.92% to 56.19%; the percentage of codon with 91–100 frequency distributions, in the intact cassette sequence was 53% and considerably increased to 74% after optimization (Fig. 2). According to the RaCC results, before codon optimization 3, 4, 2 and 1 codons respectively belong to Ilu, Leu, Pro and Arg which due to optimization turned to appropriate *E. coli* codon preferentially. The *HindIII* and *XhoI* restriction sites for cloning in pET vectors family were considered at the N and C-terminal of the sequence. The bar graph shows frequency distribution of codon usage. The native gene sequence had 53% of codons most preferred in *E. coli*, while the optimized gene possessed 74%. 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.

mRNA structure prediction

RNAfold web server calculates mRNA structure according to two algorithms, base pairing probability matrix and minimum free energy (MFE). After optimizing, it was determined that 5′ end of the sequence (consist of the start codon) be folded in the desired pattern with minimum interior loop size and pseudoknot especially in initial of structure (Fig. 3).

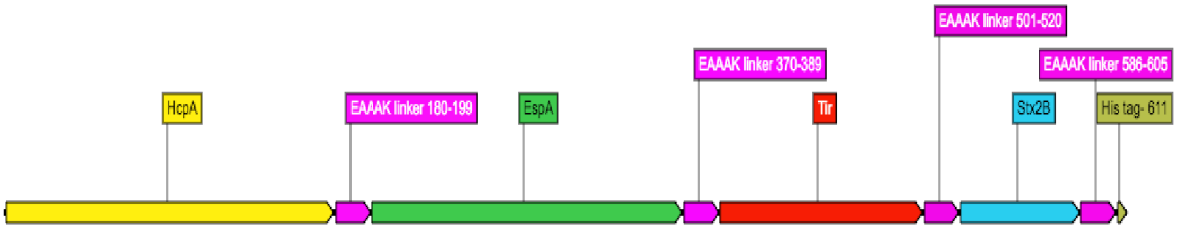

Figure1. Schematic drawing of cassette construct includes of HcpA, EspA, Tir and Stx2B subunit proteins connect each other by repetition of four EAAAK linkers, furthermore His tag termination.

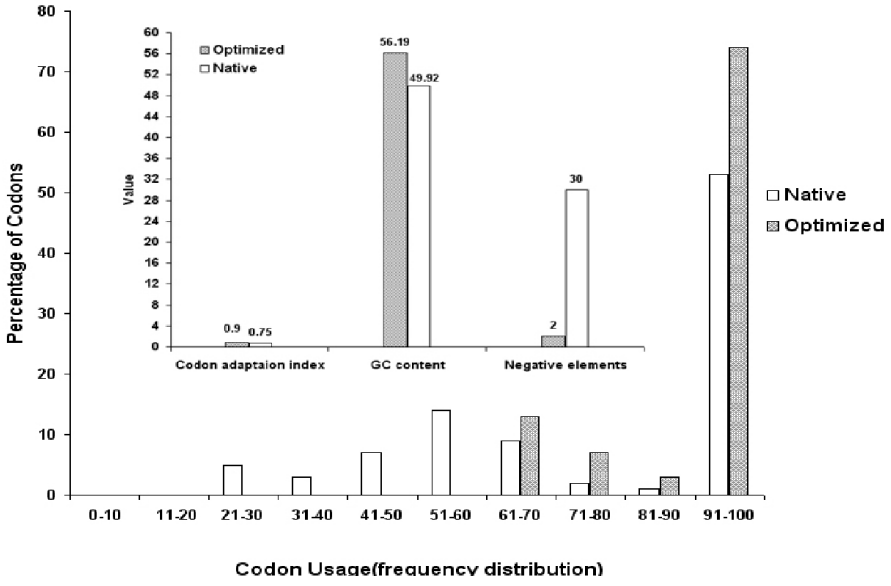

Figure 2. Comparative analysis of various parameters employed for codon optimization.

ΔG of the native sequence prediction calculated -562.60 kcal/mol compare with -630.70 for optimized sample. Ensemble diversity which is the average base pair distance between all structures was respectively 369.68 and 425.63, for native and optimized structure respectively.

The physico-chemical parameters

Sequence molecular weight with 611 amino acids was estimated 64850.3 Da. Respectively HcpA-EspA-Tir-Stx2B genes arrangement predicted as most permanent with a 40.31 stability score (data not shown). Protparam server was classified desired protein as stable; also this tool estimated pH of 5.29 for isoelectric point (pI). The pI value tool (pIb7) indicated theoretical acidity pI of the protein. The extinction coefficient of protein computed 38640 M^{-1} cm^{-1} at 280 nm. The half-life was greater than 10 hours while over expressing in *E. coli*. Furthermore Grand average of hydropathicity (GRAVY) calculated -0.406 for the chimera.

The solubility of the fusion protein was interpreted by using the PROSO II based on a classifier exploiting subtle differences among soluble proteins from target DB and the PDB and notoriously insoluble proteins from Target DB. The protein classified as soluble protein at threshold 0.734 with 71% accuracy.

Although Innovagen server predicted, protein has pH 5.05 in Iso-electric point, but both servers indicate good water solubility. Residues which form the hydrophobic and polarity of the desired protein are used to characterize the solvent accessibility distributions.

Secondary and tertiary structures

Prior to obtain close secondary and tertiary structure absence of disulfide bridges confirmed by DISULFIND tool, then comparative and de novo prediction was exploited for produced model (Fig. 4).

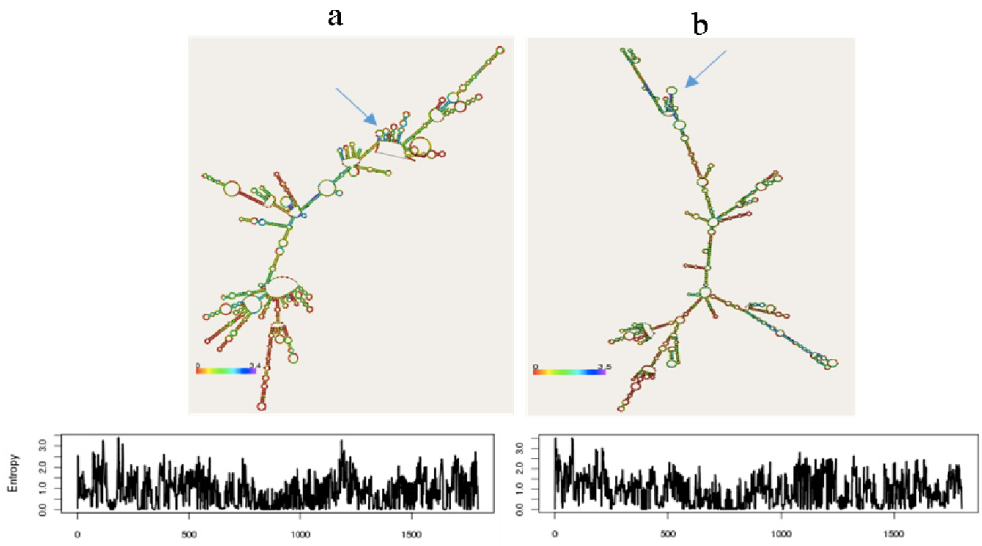

Figure 3. Schematic shape and diagram related to RNA structure. Prediction of RNA secondary structure before (a) and after (b) optimization which exhibits 5′ terminus and entropy chart.

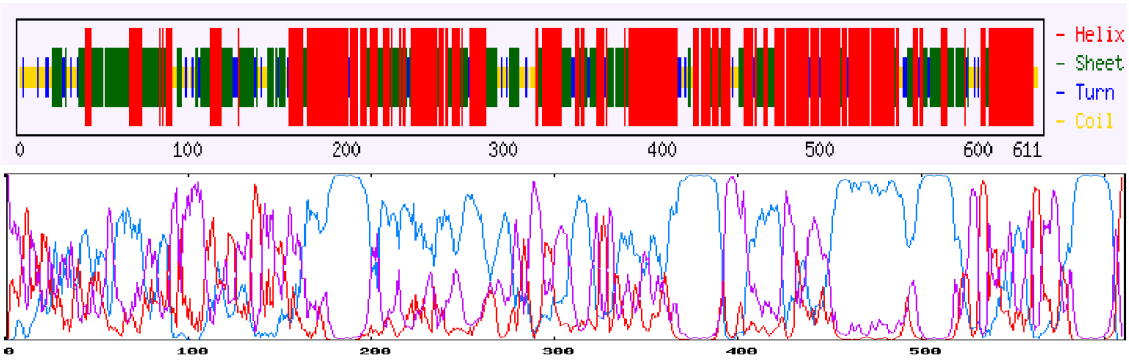

Figure 4. Graphical demonstrationof the secondary structure of fusion, consist of HcpA-EspA-Tir-Stx2B.

In the level of secondary structure, GOR IV and PSIPRED and CFSSP servers were recruited, to get a reliable prediction. GOR IV calculated structural contents of protein such as the alpha helix, extended strand, and random coil. The compound of secondary structure for chimeric protein was 58.92% (Hh), 10.15% (Ee), and 30.93% (Cc). Estimation of other residues in Pi helix, Beta Bridge, and Beta-turn calculated approximately 0% in PHD server. There were seven sheet structures as PSIPRED indicated. Also, we recruited PSIPRED to compare the secondary form of single native genes and their structure in the cassette. This Software demonstrates there was some difference just in 4 positions, and approximately they had corresponded with linker fragment. Analysis consequences showed sheets located among positions 115-121, 124-129, 136-142, 296-297, 534-538 and 563-566. In summary, all database predicted alpha helix are dominant in secondary structure; also all linkers took this shape to their selves.

For the tertiary structure of the fusion protein, I-TASSER server predicted aprotein with 6.5nm aperture diameter, 11.5 nm structure diameter and approximately 3 nm width. Furthermore, it exhibits four common domains and cassette tag, joined each other with stable linkers (Fig. 5). The confidence score (C-score) of first models was -0.76 which had a significant difference with next four predictions. C-score is an indicator between [−5 to 2], where a C-score of the higher number, approve more modeling confidence.

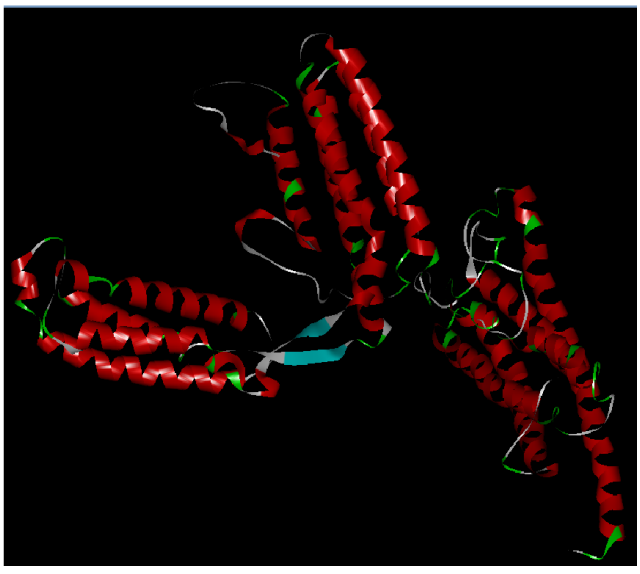

Figure 5. Predicted Tertiary structure of the protein by I-TASSER tool. Besides of initial vector sequence, last linker and His tag, there was four domain include HcpA, EspA, Tir and Stx2B which demonstrated by Accelrys software. Also the structure confirmed separating function of the rigid linker.

Evaluation of model confidence and stability

The expected template modeling score (TM-score) was measured 0.62 ± 0.14, and the supposed root mean squared deviation (RMSD) was 9.5 ± 4.6Å. Moreover Ramachandran plot in RAMPAGE server can estimate potential of errors in tertiary structure prediction. It evaluated 82.6% of residues are in the favored region (A, B, and L) of the plot, 12.0% in allowed area (a, b, l, and p), and 5.4% in outlier region. Overall quality factor is an index in ERRAT2 server that measured 90.033 for the produced model (Fig. 6). Furthermore ProSA tool calculated Potential errors of 3D models via Z-score index (-3.46), both of these indexes were in the span of scores generally which found for native proteins of similar size.

B-cell, T-cell and allergenic site prediction

Every antigen shall have hydrophilicity and another parameter to motivate both of the B and T cells to be considered as a good vaccine candidate. Among some software with different algorithm BCPred server and ABCPred were selected. For consensus predictions besides of BCpred, AAP was recruited (Table 1). Hence ABCPred best scores were selected for the linear epitope, also their vaxijen score examined for confidence (Table 1). It is well known B-cells can recognize discontinues and conformational epitopes which they predicted by CBTOPE. Briefly predictions demonstrated the appropriate distribution of nonlinear epitopes and their various amino acid lengths (Table 2). Among different T-cell epitope predictors, IEDB is reliable software which was updated with various HLA (human leukocyte antigen) alleles. Besides, it has more option in its interface. For screening MHC I class, five most frequency allele as its manual said was selected (HLA-A*0201, HLA-A*2402, HLA-A*0101, HLA-A*0301, HLA-B*0702). Also in this class, nine residues long are more common than others (Table 4). For predicting MHC II, we picked out HLA alleles with maximal population coverage which they were 27 number and 15-mers long (Table 3). For allergenicity prediction, just methods that based on SVM algorithm recognized this fusion, as allergen sequence. However other methods of Algpred also other servers like AllerTOP and AllergenFP did not prove allergenicity.

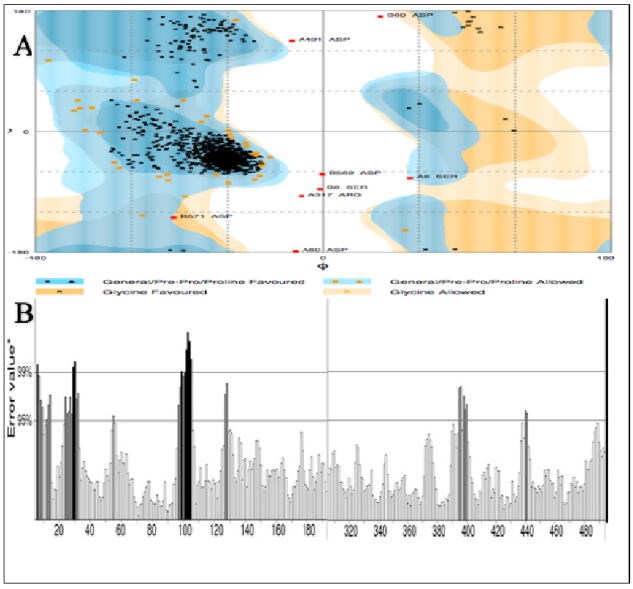

Figure 6. Evaluation of protein prediction stability based on Ramachandran plot in RAMPAGE (A) and ERRAT2 (B) online tools.

Table 1. linear B-cell epitopes prediction by BCPred server and ABCPred servers.

AAP Predictions	A.A Positions	BCPred scores	VaxiJen scores	ABCPred Predictions	A.A Positions	ABCPred scores	VaxiJen scores
DSPTTTDPDAAASATETATR	399	1	1.3132	KVDGKEYWTSRWNLQP	537	0.95	1.4437
KYDEQQAKRQEELKVSSGAG	481	1	1.6979	TPEPDSPTTTDPDAAA	395	0.94	0.9980
LTTVVNNSQLEIQQMSNTLN	326	1	0.5644	TGGQQMGRGMDKQRGF	25	0.94	0.9013
TGQESLNGLSVVMTPGWDNA	128	1	0.5036	LSVVMTPGWDNANGVT	136	0.93	0.4733
HGGLDTCDGGSNGIPSPTTT	93	1	0.8378	EHGGLDTCDGGSNGIP	92	0.89	0.8833
TGMTVTIKSSTCESGSGFAE	560	1	1.2414	ETATRDQLTKEAFQNP	414	0.88	0.5173
ASMTGGQQMGRGMDKQRGFT	22	1	1.0116	TTAQKMANLVDAKIAD	252	0.88	0.8870
EEQAKAAGEEAKQQAIENNA	457	1	1.0995	AQAQKKYDEQQAKRQE	476	0.87	1.5987
RNDISVTGIRDLSGDLSAGD	292	1	1.0224	EEAKQQAIENNAQAQK	465	0.87	1.0509
LTPEPDSPTTTDPDAAASAT	394	1	1.3132	SVTGIRDLSGDLSAGD	296	0.87	0.8022
NAQAQKKYDEQQAKRQEELK	475	0.996	1.5196	FSKYNEDDTFTVKVDG	525	0.84	1.1535
HGGLDTCDGGSNGIPSPTTT	93	0.991	0.8378	LGVFQAAILMFSYMYQ	216	0.84	0.4128
ASMTGGQQMGRGMDKQRGFT	22	0.98	1.0116	SSTCESGSGFAEVQFN	568	0.83	1.3826
SVVMTPGWDNANGVTGWARN	137	0.972	0.1122	AQLTGMTVTIKSSTCE	557	0.83	0.9741
KFADMNEASKASTTAQKMAN	240	0.959	0.7855	YDEQQAKRQEELKVSS	482	0.83	1.7408
KVDGKEYWTSRWNLQPLLQS	537	0.926	0.8872				
VQSSTDKNAKAKLPQDVIDY	268	0.908	0.9951				
NNLTTVVNNSQLEIQQMSNT	324	0.848	0.4734				

Table 2. Conformational B-cell epitopes from full-length proteins using CBTOPE server.

Amino acid	Position	Probability scale	Amino acid	Position	Probability scale
MDKQ	33-37	4,4,4,4	K	322	4
I	43	4	NL	325-326	4,4
A	83-84	4	VVNNSQLEI	229-337	4,4,5,4,4,4,4,4,4
L	88	4	LGK	367-369	4,4,4
S	137	4	E	397	4
VM	139-140	4,5	K	423	4
TPG	141-143	4	FQN	426-428	4,4,4
NANGVTGWARNCN	146-158	4,4,5,5,4,4,,5,5,4,4,4,4	NQKVNIDELGN	431-441	4,4,4,4,5,5,5,4,4,4,4
SALQ	163-166	4,4,4,4	ANIE	454-457	4,4,4
DDA	176-178	4,4,4	RQEEL	489-493	4,4,4,4,4
FEELGVF	213-219	4,4,4,4,4,4,4	S	497	4
A	222	4	KIE	522-524	4,4,4
ANLV	258-261	4,4,4,4	KYN	527-529	4,4,4
A	278	4	T	535	4
PQDVIDYIND	281-290	4,4,4,4,4,4,5,4,4,4	K	537	4
IS	295-296	4,4	EYWTSRWNL	542-550	4,4,5,4,4,5,4,4,4
T	298	4	L	554	4
IR	300-301	4,4	QL	558-559	4,4
LQTV	312-315	4,4,4,4	TVTIK	563-567	4,5,5,5,4
AIS	318-320	4,4,4	ST	569-570	4

Discussion

Several bacterial pathogens infect their hosts via contact to the cell membrane by their particular attachment ability, and Enterohemorrhagic *E. coli* is not an exception. In this context, a chimeric vaccine which also carried adjuvant sequences was designed. The stated of this paper was developing an immunogen candidate inclusive of various attaching and virulence factors which separated by proper linkers and it's *in silico* study. This approach leads to the best stability also most possible immunogenicity of the chimera. Our sequence fragment contains four supposed antigens which are suitable for expression in *E. coli* system. Carrier piece of Stx toxin (Stx2B) with three adhesion antigens consist of HcpA, EspA and Tir made desire construction. Towards the secretory pathway, proteins are synthesized only in association with the endoplasmic reticulum (ER). The signal peptide will participate in this function, and the N-terminus signal sequence is usually removed in a mature protein. So the existence of these residues is useless in fusion subunits. 110 amino acids from the middle of Tir sequence was too epitope-rich also part B of Stx subunits is nontoxigenic and has a significant role in invading and virulence.

Other apparatus antigens for attaching to hostess moreover that have an important role in EHEC attachment exist in some other pathogenesis strain. Linkers propose several advantages for the generation of multi-domain proteins, as instance improving biological activity, increasing expression efficiency, and achieving desirable pharmacokinetic profiles [36]. (EAAAK)n (n = 1-5) is a rigid empirical linker which secondary structure prediction servers, proved its α-helical forming [37]. The Glu−-Lys+ salt bridge among three Ala can stable its structure [38]. The synchrotron X-ray, small-angle scattering experiments, indicated that short helical linkers (n = 1, 2, 3) lead to multimerization, while the longer linkers (n = 4, 5) solvate monomeric fusion proteins [39]. Some successful experimental result of using four times repeated (EAAAK motif), acknowledged efficient separating between protein domains. In little generated time microorganisms like *E. coli*, it is proved which optimal codons aid to attain better translation speed and accuracy. Overall GC content, codon frequency distribution and codon adaptive index (CAI) were modified as codon bias optimization.

Table 3. Selection of peptides containing T cell epitopes from chimeric protein by MHC Class-I and II binding prediction algorithms. Lowest percentile ranks for each HLA allele type which offered, interpret as a good binder.

MHC Class	Allele	Start	End	Peptide	Percentile rank
I	HLA-A*01:01	351	359	RSDVQSLQY	0.2
I	HLA-A*01:01	72	80	LTDMLQTFV	0.25
I	HLA-A*02:01	224	232	LMFSYMYQA	0.3
I	HLA-A*03:01	256	264	KMANLVDAK	0.35
I	HLA-A*24:02	218	226	VFQAAILMF	0.4
I	HLA-A*03:01	361	369	TISAISLGK	0.45
I	HLA-A*01:01	55	63	LSAIGIPAY	0.45
I	HLA-B*07:02	80	88	VPYRTAVEL	0.5
I	HLA-A*03:01	314	322	TVKAAISAK	0.65
I	HLA-B*07:02	15	23	VPRGSHMAS	0.7
II	HLA-DRB1*09:01	309	323	AGDLQTVKAAISAKA	0.01
II	HLA-DRB3*01:01	274	288	KNAKAKLPQDVIDYI	0.01
II	HLA-DRB1*11:01	46	60	MVVIGIIAILSAIGI	0.14
II	HLA-DRB3*01:01	525	539	FSKYNEDDTFTVKVD	0.19
II	HLA-DRB1*09:01	224	238	LMFSYMYQAQSNLSI	0.25
II	HLA-DRB5*01:01	355	369	QSLQYRTISAISLGK	0.25
II	HLA-DRB3*01:01	170	184	EDVFRFDDANEAAAK	0.31
II	HLA-DRB5*01:01	111	125	TTRYVSAMSVAKGVV	0.31
II	HLA-DRB1*08:021	313	327	QTVKAAISAKANNLT	0.31
II	HLA-DRB4*01:01	41	55	TLIELMVVIGIIAIL	0.7

Briefly, the optimized DNA sequence had a CAI of 0.90 instead 0.75 in prior; that demonstrate the optimized genes can have a good expression in *E. coli* k12. RNA fold server was recruited for prediction of mRNA secondary form insomuch the base pairing formation messenger RNA can change ribosome processivity and its cellular half-life. Moreover its profit was the ability for analyzing extended sequence with two algorithms, according to base pairing probability matrix and minimum free energy (MFE). Predicted RNA structure showed the mRNA had appropriate ΔG and acceptable interior loop for permanent translation in the hostess. The study of protein secondary structure plays a significant role for tertiary structure prediction, with the *ab initio* algorithm or protein fold identification by providing additional constraints. For this step, PSIPRED and GOR method were used, GOR IV calculated secondary structure parameter such as the alpha helix, strand Coil, and PSIPRED will draw a simple scheme and suitable for compartment between native gene and their fusion form, also former and after optimizing. Extinction coefficient index was computed 38640 M−1 cm−1 at 280 nm wavelength, which It is originated from Trp, Tyr and Cystine concentration (W: 0.7%, Y:1.8% and C: .8% that Cystine have a direct relation to cysteine). Theoretical pI value approximately calculated 5.3 on the different server. Hence it shows a few acidic nature of the protein for future experimental process. Aliphatic index of protein (75.45) indicating the stability of protein in the broad range of temperature and calculated according to Ala, Val, Leu, and Ile density. Finally, Instability index (38.46), predicting the acceptable stability of protein in a test tube. For predicting three-dimensional structure of the fusion protein, we can utilize de novo and the comparative method [40]. I-TASSER tool in Zhang Lab server can utilize both of them for this modeling. Swiss model can also do it, but compartment among result indicated more accuracy and reliability of I-TASSER server. For the best outcome, three arrangements submitted to this server. At first only desire gene and their linker among them, then remaining part of vector plus former and in other query, second sequence moreover a linker and His tag at the C-terminal end (whole sequence). In full sequence Submission, the first prediction had significant differences with others. It indicated -0.76 confidence score and 0.62 ± 0.14 TM-score and 9.5 ± 4.5Å RMSD, which a TM-score >0.5 represent a model with proper topology, and the lower RMSD describe a better modeling, in comparison to the native structure. Hence these results interpreting as good accuracy and correct topology of predicted model. The ProSA outcomes (Z-score) demonstrated a passable overall quality for predicted model; it authenticated that our chimeric construction has features close to native structures from the point of stereochemistry. Prediction indicated acceptable protein stability based on Ramachandran plot (5.4% residues were in outlier region). Empirical results are suggesting even until approximately 10% of amino acids be in outlier zone, while ERRAT index and ProSA dot confirms the prediction, so it is allowable to go on with mentioned model. These outlier region residues perhaps are because of the presence of chimeric junctions. Ability to motivate the B-cell and the T-cell response is the final target to design novel vaccines. It is nominated B cells can recognize both of linear and conformational epitopes and for first one BCPred and ABCPred were utilized. Also for discontinues epitope DiscoTope can detect them by 3D structure format. B cell epitope prediction can accomplish in some other servers which they almost use the same algorithm, in addition to more reliability vaxijen software can employ for affirmation.

Immune Epitope Database (IEDB) can analyze sequences in term of T-cell epitopes and binding affinity of MHC molecules [41]. For this purpose, we utilized IEDB instead of other servers, since it has comprehensive HLA alleles which provide most population coverage for MHC-I and MHC-II molecules. Also, most of allergenicity prediction database assess our protein as a non-allergen. However, a variant candidate vaccine against EHEC was introduced. Most of them inclusive of three antigens or lower even in other chimeric protein there are no conclusions about assured success and comprehensive protection, so this study can step to forward this [42–45].

Conclusion

The final state of this study is developing a subunit vaccine inclusive of four proteins and confronting with colonization and invasion factors of Enterohemorrhagic *Escherichia coli*. In this regard bioinformatics tools have many uses in different aspects whether in genome level (sequence optimization for expression efficiently in *E. coli* k12) or transcriptome step (prediction topological and thermodynamical features of mRNA). Other *in silico* tools focused on the proteom phase with prediction and authentication of secondary and tertiary structures moreover associated characterizations like epitopes and allergen prediction. So according to these predictions, assumed protein with cellular and humoral immune motivation may utilize as a candidate for fill the vacuum of a broad vaccine coverage against EHEC.

Acknowledgements

The authors are indebted to Imam Hossein University for their support funding to carry out this research. The authors declare no conflict of interests.

References

1. Nazarian, S., Gargari, S.L.M., Rasooli, I., Amani, J., Bagheri, S., Alerasool, M., An in silico chimeric multi subunit vaccine targeting virulence factors of enterotoxigenic *Escherichia coli* (ETEC) with its bacterial inbuilt adjuvant. *J Microbiol Methods,* 2012, Vol. 90, pp. 36-45.

2. Babiuk, S., Asper, D.J., Rogan, D., Mutwiri, G.K., Potter, A.A., Subcutaneous and intranasal immunization with type III secreted proteins can prevent colonization and shedding of *Escherichia coli* O157: H7 in mice. *Microb Pathog*, 2008, Vol. 45, pp. 7-11.

3. Larrie-Bagha, S.M., Rasooli, I., Mousavi-Gargari, S.L., Rasooli, Z., Nazarian, S., Passive immunization by recombinant ferric enterobactin protein (FepA) from *Escherichia coli* O157. *Iran J Microbiol*, 2013, Vol. 5, p. 113.

4. Ishikawa, S., Kawahara, K., Kagami, Y., Isshiki, Y., Kaneko, A., Matsui, H., Okada, N., Danbara, H., Protection against Shiga toxin 1 challenge by immunization of mice with purified mutant Shiga toxin 1. *Infect Immun*, 2003, Vol. 71, pp. 3235-3239.

5. Konadu, E., Donohue-Rolfe, A., Calderwood, S.B., Pozsgay, V., Shiloach, J., Robbins, J.B., Szu, S.C., Syntheses and Immunologic Properties of *Escherichia coli* O157 O-Specific Polysaccharide and Shiga Toxin 1 B Subunit Conjugates in Mice. *Infect Immun*, 1999, Vol. 67, pp. 6191–6193.
6. Rad, H.S., Mousavi, S.L., Rasooli, I., Amani, J.,, Nadooshan, M.R.J., EspA-Intimin chimeric protein, a candidate vaccine against *Escherichia coli* O157: H7. *Iran J Microbiol*, 2013, Vol. 5, pp. 620–627.
7. Amani, J., Mousavi, S.L., Rafati, S., Salmanian, A.H., Immunogenicity of a plant-derived edible chimeric EspA, Intimin and Tir of *Escherichia coli* O157: H7 in mice. *Plant Sci*, 2011, Vol. 180, pp. 620-627.
8. Gu, J., Liu, Y., Yu, S., Wang, H., Wang, Q., Yi, Y., Zhu, F., Yu, X.J., Zou, Q., Mao, X., Enterohemorrhagic *Escherichia coli* trivalent recombinant vaccine containing EspA, intimin and Stx2 induces strong humoral immune response and confers protection. *Microbes Infect*, 2009, Vol. 11, pp. 835-841.
9. Frankel, G., Phillips, A.D., Rosenshine, I., Dougan, G., Kaper, J.B., Knutton, S., Enteropathogenic and enterohaemorrhagic *Escherichia coli*: more subversive elements. *Mol Microbiol*, 1998, Vol. 30, pp. 911–921.
10. Xicohtencatl-Cortes, J., Monteiro-Neto, V., Ledesma, M.A., Jordan, D.M., Francetic, O., Kaper, J.B., Puente, J.L., Girón, J.A., Intestinal adherence associated with type IV pili of enterohemorrhagic *Escherichia coli* O157: H7. *J Clin Invest*, 2007, Vol. 117, pp. 3519–3529.
11. Doytchinova, I.A., Flower, D.R., VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, 2007, Vol. 8, pp. 4-10.
12. Zuker, M., Stiegler, P., Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 1981, Vol. 91, pp. 33–48.
13. Reuter, J.S., Mathews, D.H., RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 2010, Vol. 11, pp. 129-135.
14. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S.E., Wilkins, M.R., Appel, R.D., Bairoch, A., Protein identification and analysis tools on the ExPASy server. *Methods Mol Biol,* 2005, pp. 571-607.
15. Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., Frishman, D., PROSO II–a new method for protein solubility prediction. *FEBS J*, 2012, Vol. 279, pp. 2192-2200.
16. Rost, B., Sander, C., Schneider, R., PHD-an automatic mail server for protein secondary structure prediction. *Bioinformatics*, 1994, Vol. 10, pp. 53–60.
17. Sen, T.Z., Jernigan, R.L., Garnier, J., Kloczkowski, A. GOR V server for protein secondary structure prediction. *Bioinformatics*, 2005, Vol. 21, pp. 2787-2788.
18. Cheng, J., Randall, A.Z., Sweredoski, M.J., Baldi, P., SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res*, 2005, Vol. 33, pp. 72-76.
19. McGuffin, L.J., Bryson, K., Jones, D.T., The PSIPRED protein structure prediction server. *Bioinformatics*, 2000, Vol. 16, pp. 404–405.
20. Schwede, T., Kopp, J., Guex, N., Peitsch, M.C., SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res*, 2003, Vol. 31, pp. 3381–3385.
21. Roy, A., Kucukural, A., Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*, 2010, Vol. 5, pp. 725–738.
22. Goodsell, D.S., Representing structural information with RasMol. *Curr Protoc Bioinformatics*, 2005, pp.4-5.
24. Wiederstein, M., Sippl, M.J., ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*, 2007, Vol. 35, pp. 407-410.
25. MacArthur, M.W., Laskowski, R.A., Thornton, J.M., Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. *Curr Opin Struct Biol,* 1994, Vol. 4, pp. 731-737.
26. Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C., Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins*, 2003, Vol. 50, pp. 437-450.
27. Doytchinova, I.A., Flower, D.R., VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, 2007, Vol. 8, p. 4.
28. EL-Manzalawy, Y., Dobbs, D., Honavar, V., Predicting linear B-cell epitopes using string kernels. *J. Mol. Recogn*, 2008, Vol. 21, pp. 243–55.
29. Kringelum, J.V., Lundegaard, C., Lund, O., Nielsen, M., Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol*, 2012, p. e1002829.
30. Ansari, H.R., Raghava, G.P., Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome Res*, 2010, Vol. 6, p. 6.
31. Guan, P., Doytchinova, I.A., Zygouri, C., Flower, D.R., MHCPred: A server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res*, 2003, Vol. 31, pp. 3621–3624.
32. Bhasin, M., Raghava, G.P.S., A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J. Biosci*, 2007, Vol. 32, pp. 31–42.
33. Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P.E., Bui, H.H., Buus, S., Frankild, S., Greenbaum, J., Lund, O., Immune epitope database analysis resource (IEDB-AR). *Nucleic Acids Res*, 2008, Vol. 36, pp. 513-518.
34. Saha, S., Raghava, G.P.S., AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res*, 2006, Vol. 34, pp. 202–209.
35. Dimitrov, I., Bangov, I., Flower, D.R., Doytchinova, I., AllerTOP v.2—a server for in silico prediction of allergens. *J Mol Model*, 2014, Vol. 20, p. 2278.
36. Mohammad, N., Karsabet, M.T., Amani, J., Ardjmand, A., Zadeh, M.R., Gholi, M.K., Saffari, M. and Ghasemi, A., *In silico* design of a chimeric protein containing antigenic fragments of Helicobacter pylori; a bioinformatic approach. *Open Bioinforma*, 2016, Vol. 10, p. 97.
37. Amani, J., Salmanian, A.H., Rafati, S., Mousavi, S.L., Immunogenic properties of chimeric protein from espA, eae and tir genes of *Escherichia coli* O157: H7. *Vaccine*, 2010, Vol. 28, pp. 6923-6929.
38. Arai, R., Ueda, H., Kitayama, A., Kamiya, N., Nagamune, T., Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein Eng Des Sel*, 2001, Vol. 14, pp. 529-532.
39. Arai, R., Wriggers, W., Nishikawa, Y., Nagamune, T., Fujisawa, T., Conformations of variably linked chimeric proteins evaluated by synchrotron X-ray small-angle scattering. *Proteins*, 2004, Vol. 57, pp. 829-838.
40. Webb, B., Sali, A., Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics*, 2014, Vol. 47, pp.1-32.
41. Greenbaum, J., Sidney, J., Chung, J., Brander, C., Peters, B., Sette, A., Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, 2011, Vol. 63, pp. 325–335.
42. Bakhshi, M., Ebrahimi, F., Zargan, J., Nazarian, S., Sheikhzade, V., Cloning and Recombinant Expression of EspA as a Virulence Factor of *E. coli* O157: H7. *JMUMS*, 2014, Vol. 24, pp. 12-20.
43. Rabinovitz, B.C., Larzábal, M., Vilte, D.A., Cataldi, A.,

Mercado, E.C., The intranasal vaccination of pregnant dams with Intimin and EspB confers protection in neonatal mice from *Escherichia coli* (EHEC) O157: H7 infection. *Vaccine*, 2016, Vol. 34, pp. 2793-2797.
44. Yazdanparast, A., Mousavi, S.L., Rasooli, I., Amani, J., Jalalinadoushan, M., Immunogenical Study of Chimeric Recombinant Intimin-Tir of *Escherichia coli* O157: H7 in Mice. *Arch Clin*, 2012, Vol. 7, pp. 45-51.
45. Novinrooz, A., Salehi, T.Z., Firouzi, R., Arabshahi, S., Derakhshandeh, A., *In-silico* design, expression, and purification of novel chimeric *Escherichia coli* O157: H7 OmpA fused to LTB protein in *Escherichia coli*. *PLoS One*, 2017, Vol. 12, pp. e 0173761.