Theoretical Biology and Medical Modelling

Research

In silico analysis of chimeric *espA*, *eae* and *tir* fragments of *Escherichia coli* O157:H7 for oral immunogenic applications

Jafar Amani[1,2], S Latif Mousavi[3], Sima Rafati[4] and Ali H Salmanian*[1]

Address: [1]National Institute of Genetic Engineering and Biotechnology (NIGEB), Shahrak-e- Pajoohesh, 15th Km, Tehran -Karaj Highway, Tehran, IR Iran, [2]Baqiyatallah University of Medical Science, Department of Biotechnology, Tehran, IR Iran, [3]Dept of Biology, Faculty of Basic Sciences, Shahed University, Tehran, IR Iran and [4]Molecular Immunology and Vaccine Research Laboratory, Dept. of Immunology, Pasteur Institute of Iran, Tehran, IR Iran

Email: Jafar Amani - jamani@nigeb.ac.ir; S Latif Mousavi - slmousavi@shahed.ac.ir; Sima Rafati - sima-rafatisy@institute.pasteur.ac.ir;
Ali H Salmanian* - salman@nigeb.ac.ir

* Corresponding author

Abstract

Background: *In silico* techniques are highly suited for both the discovery of new and development of existing vaccines. *Enterohemorrhagic Escherichia coli* O157:H7 (EHEC) exhibits a pattern of localized adherence to host cells, with the formation of microcolonies, and induces a specific histopathological lesion (attaching/effacing). The genes encoding the products responsible for this phenotype are clustered on a 35-kb pathogenicity island. Among these proteins, Intimin, Tir, and EspA, which are expressed by attaching-effacing genes, are responsible for the attachment to epithelial cell that leads to lesions.

Results: We designed synthetic genes encoding the carboxy-terminal fragment of Intimin, the middle region of Tir and the carboxy-terminal part of EspA. These multi genes were synthesized with codon optimization for a plant host and were fused together by the application of four repeats of five hydrophobic amino acids as linkers. The structure of the synthetic construct gene, its mRNA and deduced protein and their stabilities were analyzed by bioinformatic software. Furthermore, the immunogenicity of this multimeric recombinant protein consisting of three different domains was predicted.

Conclusion: a structural model for a chimeric gene from LEE antigenic determinants of EHEC is presented. It may define accessibility, solubility and immunogenecity.

Background

Enterohemorrhagic Escherichia coli O157:H7 (EHEC) is an important human pathogen [1], causing diarrhea and in some cases hemolytic-uremic syndrome (HUS), leading to kidney failure and even death [2]. EHEC produces several virulence factors, enabling it to colonize the large bowel and cause disease [3].

Cattle are most frequently identified as the primary source of bacteria, so reduction in *E. coli* O157:H7 prevalence in cattle by vaccination represents an attractive strategy for reducing the incidence of human disease [4]. An experimental vaccine was recently shown to significantly reduce shedding of the organism under natural exposure conditions [5].

These pathogenic bacteria contain a chromosomal island known as the Locus of Enterocyte Effacement (LEE, 35KD), containing genes critical for forming the attachment and effacement (A/E) lesion. This locus can be divided into three functional regions: the first one encoding a type III secretion system; the second containing the genes *eae* and *tir*; and the third consisting of *espD*, *espB*, and *espA* [6,7].

Intimin, a key colonization factor for EHEC O157:H7 acts as an outer membrane adhesion protein which is encoded by the gene *eae*. This protein mediates bacterial attachment through its C-terminal region to enterocytes by binding to Tir (Translocated Intimin Receptor) [8,9].

Tir, a 78-kDa protein, is secreted from EHEC and is efficiently delivered into the host cell [10,11].

The type III secretion system is involved in the secretion of different proteins including EspA, EspB, EspD, and Tir. EspA forms a filamentous structure on the bacterial surface as a bridge to the host cell surface. It delivers EspB, EspD, and Tir directly into the host cell. EspB is delivered primarily into the host cell membrane where it becomes an integral membrane protein and, along with EspD, forms a pore structure through which other bacterial effectors, such as Tir, enter the host cell [6,12]. Additionally, studies on rabbit models indicate that pedestal formation is mediated by the same proteins (Intimin, EspA, EspB, EspD and Tir), and translocated Tir can bind to intimin via amino acids 258 to 361 [3,13].

The Tir-Intimin interaction causes attachment of EHEC to the intestinal cell surface and triggers actin cytoskeletal rearrangements, resulting in pedestal formation. Recent evidence shows that active immunization of mice with recombinant Intimin from *Citrobacter rodentium* as a mouse model pathogen can prevent colonization of bacteria in the digestive tracts of animals [14].

These determinants are potent mucosal immunogens and induce humoral and mucosal responses (IgA instead of IgG) following oral administration [15,16]. Among different systems for oral administration, transgenic plants are becoming more attractive because of their low cost, easy scale-up of production, natural storage organs (tubers and seeds), and established practices for efficient harvesting, storing, and processing [17,18]. Moreover, a number of proteins such as recombinant antibodies and recombinant subunit vaccines have been expressed successfully in transgenic plants [19].

In this study we designed a new structural model containing three putative antigenic determinants of EspA, Intimin and Tir, fused together by hydrophobic linkers. Addition

of the regulatory sequences Kozak and ER-retention signal at the 5' and 3' ends respectively, and codon optimization of this chimeric gene for expression in plants, were used to improve the efficiency of transcription and translation [20-22]. Finally, a novel *in silico* approach was used to analyze the structure of the designed chimeric protein.

Results
Design and construction of chimeric gene
The 282 amino acids from the carboxy terminus of Intimin have been reported to be involved in binding to its receptor Tir [23,24]. The region of Tir involved in the interaction with intimin has also been mapped (residues 258 to 361, designated Tir 103) [25]. For the third fragment, a truncated form of *espA* (lacking 36 amino acids from the N-terminal of the protein, designated EspA 120) was selected. This part of EspA120 is exposed on the bacterial surface [6].

Upon sequence comparison by ClustalW, the C-terminals of intimin (282 amino acids) and EspA (120 amino acids) and the middle part of Tir (103 amino acids) showed high degree of conservation among different strains of *E. coli* O157:H7 (Data not shown).

These three parts were selected for designing a synthetic construct. In order to separate the different domains, linkers consisting of EAAAK repeats and 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 [26]. Four repeated EAAAK sequences were introduced between different domains for more flexibility and efficient separation. The Kozak sequence [27] was added before the start codon in order to ensure high and accurate expression of mRNA in a eukaryotic host. For efficient accumulation of the recombinant protein in Endoplasmic Reticulum (ER), the sequence KDEL was added at the end of the synthetic construct. Arrangements of fragment junctions and linker sites are shown in Figure 1.

Bioinformatic analysis of the wild type and optimized synthetic gene
A synthetic sequence encoding the chimeric gene was designed using plant codon bias. To optimize the synthetic gene, negatively *cis* acting motifs and repeated sequences were avoided. Both the wild type and the synthetic chimera were analyzed for their codon bias (Figure 2A) and GC content (Figure 2B),

The overall GC content was reduced from 41.59 to 40.96%, which should increase the overall stability of mRNA from the synthetic gene. Moreover, there was no sequence stretch within the gene showing an average GC content below 40%.
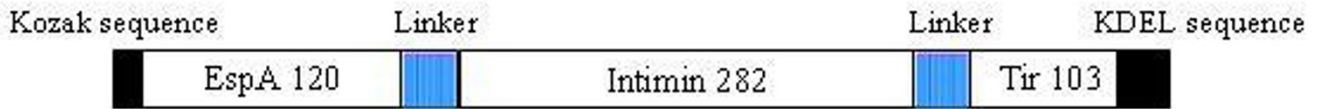
Figure 1
Schematic model which shows the construction of EspA 120, Intimin 282 and Tir 103, bound together by the linkers for expression in plants; these fragments were selected on the basis of the common sequence found in different strains of *E. coli* O157 H7.
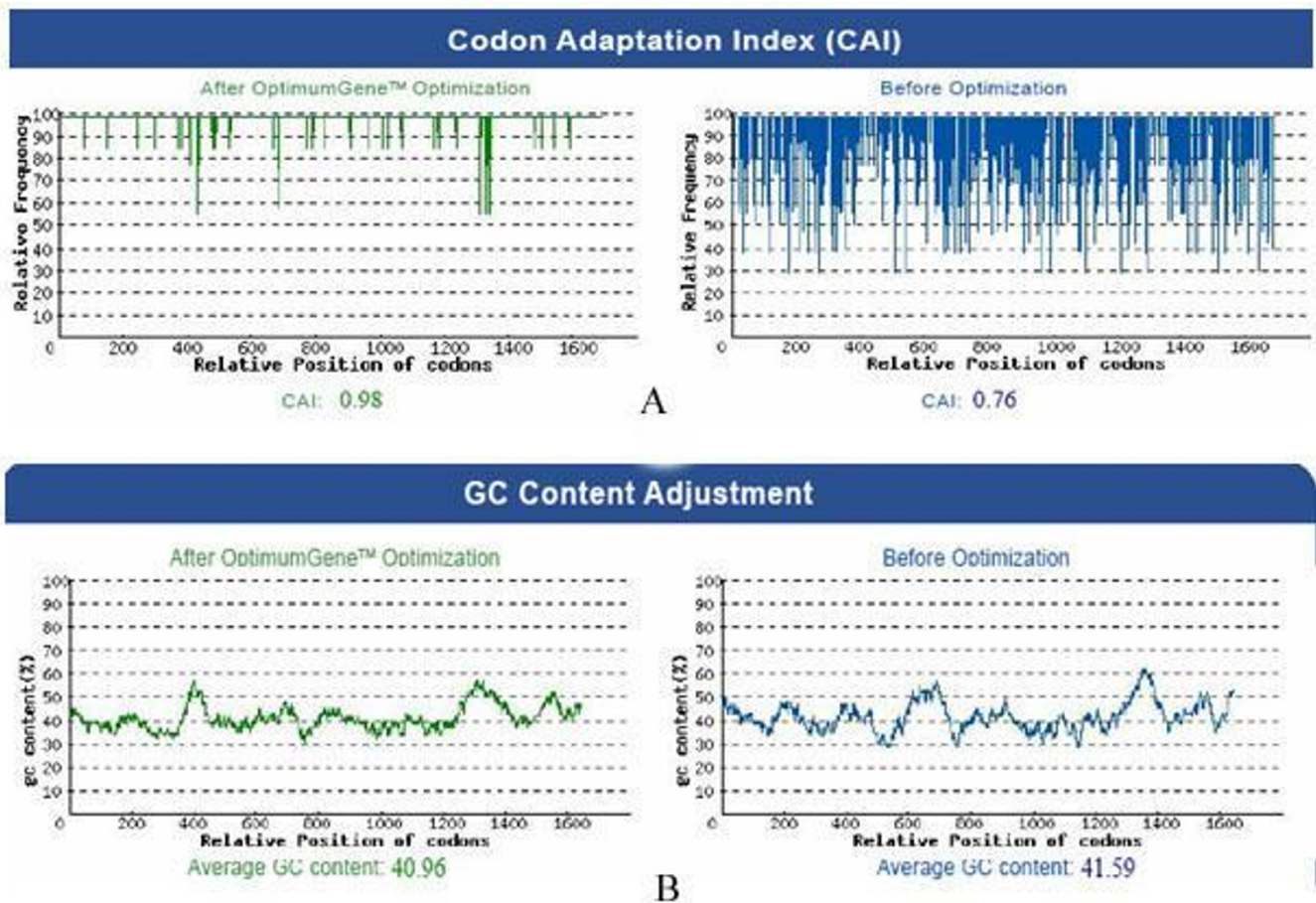

Figure 2
A: Codon usage analysis of wild type and optimized gene for expression in plants. The value of 100 is set for the codon with the highest usage frequency for a given amino acid in the desired expression into plants. This procedure allows us to compare the adaptiveness of different codons relative to each other (*relative adaptiveness*). Plots represent the relative adaptiveness of a given codon at the indicated codon position. B: GC analysis of wild type and optimized chimeric gene. Plots represent the average GC content, before and after optimization.

The optimized gene showed a codon bias for plants and contained no rarely used codon. This is also reflected by the codon adaptation index (CAI), which is a measurement of the relative adaptiveness of the codon usage of a gene compared with the codon usage of highly expressed genes. The chimeric gene showed a CAI of 0.98, compared to that of the wild type gene, which was only 0.76 [28].

Within the synthetic construct, the splice sites, polyadenylation signal, instability elements, and all the *cis*-acting sites that may have a negative influence on the expression rate were removed (Table 1). Furthermore, the necessary restriction enzyme sites (*Xba*I and *Sac*I) were introduced at the ends of the sequence for cloning purpose.

mRNA structure prediction

A genetic algorithm-based RNA secondary structure prediction was combined with comparative sequence analysis to determine the potential folding of the chimeric gene. The 5' terminus of the gene was folded in the way typical of all bacterial gene structures. The minimum free

Table 1: Analysis of cis-acting elements

Splice site	Original	Optimized
GGTAAG	2	0
GGTGAT	4	0
GTAAAA	1	0
GTAAGT	2	0
GTACGT	0	0
Poly A		
AATAAA	1	0
AATGAA	0	0
AATGGA	0	0
TATAAA	2	0
AATAAT	3	0
AAAAAAA	1	0
Poly T		
TTTTTT	1	0
Destabilizing element		
ATTTA	3	0

energy for secondary structures formed by RNA molecules was also predicted. All 34 structural elements obtained in this analysis revealed folding of the RNA construct. The data showed the mRNA was stable enough for efficient translation in the new host (Data not shown) [29].

Protein secondary structure prediction

The secondary structure of the chimeric protein was predicted by online software. Three prediction methods were compared for evaluating the structure of this protein. The results showed that helix structures lie in the regions of aa 129 to 148 and aa 431 to 450, which are related to the hydrophobic amino acids inserted between different domains (Figure 3) [30,31].

Tertiary structural prediction for the chimeric protein

Comparative and *ab initio* modeling of the synthetic sequence was exploited to produce 3D models of the chimeric protein. Two hundred thirty three-dimensional models were generated for this chimeric protein. The models were uploaded to the server to draw the tertiary structural illustrations with Swiss-PdbViewer and Rasmol software in order to determine the final structure of the protein. Furthermore, SCRATCH servers http://www.igb.uci.edu/ developed by California University were used for protein structure prediction by PSI-BLAST and neural networks. There were two α-helices and several β-turns, which were consistent with the results of secondary structure analyses. The results of tertiary structure prediction showed the formation of three separate domains of the chimeric protein (Figure 4) [32,33].

Evaluation of model stability

The profile of energy minimization was calculated by spdbv (Swiss-PdbViewer) (-1391.230 Kcal/mol) indicating that the recombinant protein had acceptable stability compared to that of original structure of each domain. Additionally, the data generated by a Ramachandran plot confirmed the structural stability of the protein (Figure 5).

Solvent accessibility prediction

The solvent accessibility distributions were characterized using the major hydrophobic and polarity properties of residual patterns. These patterns showed that the mean residue accessible surface area (ASA) gave a high solvent accessibility value, approximately fifty percent (Data not shown) [34].

Prediction of B-cell epitopes

Different factors such as hydrophilicity, plasticity, exterior accessibility, antigenicity and secondary structure were used to predict the chimeric protein epitopes. The epitopes located on the surface of the protein could interact easily with antibodies, and they were generally flexible. Bcepred software was used to determine the
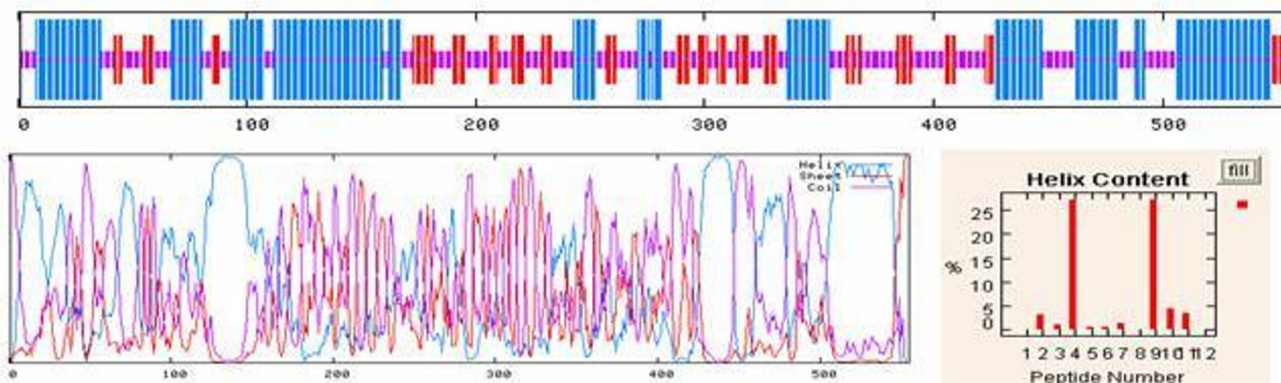
Figure 3
Analysis of chimeric EspA-Intimin-Tir protein secondary structure.

continuous B cell epitope based on single characters including hydrophilicity, antigenicity, flexibility, accessibility, polarity and exposed surface (Table 2). As shown in Table 2, linkers between different domains (aa 129 to 148 and aa 431 to 450) contained no epitope sites [35-37]. Furthermore, the conformational epitopes for B cells were predicted by the Discotope server (Table 3) [38].

Discussion
Many bacterial pathogens infect or invade their hosts via mucosal surfaces. This process is initiated by the attachment of the bacteria to the cell membrane via specific receptors. Enterohemorrhagic *E. coli* is a good model and has been well studied in this context. In this bacterium, the antigens Intimin, EspA, and Tir are required for attachment to the intestinal mucosa [39]. If the function of these receptors was impaired, the bacterium could not attach to the host cell surface and the disease would be suppressed.
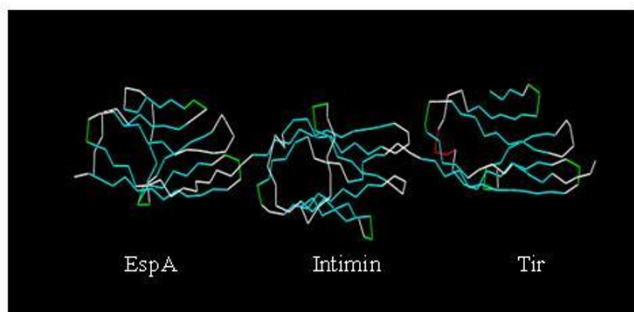

Figure 4
Ab initio **and comparative modeling was used to predict the tertiary structure of the chimeric protein, EspA-Intimin-Tir**. The result was viewed by Rasmol software.

This impairment is related to the production of immunoglobulin class A (IgA), which is the dominant antibody on the mucosal surface [2].

Therefore, mucosal immunization especially via the oral route is an attractive strategy for inducing protective immunity against mucosal pathogens [40]. Several vehicles (Polymers, Alginate, Polyphosphazenes and other biodegradable polymers, Immunostimulating complexes (ISCOM), Liposomes) [41] have been used for delivering antigen to the target tissue. The capacity of plants for producing vaccines which could induce mucosal immunity is a great advantage. Plant cells act as a natural microencapsulation system to protect the vaccine antigens from being degraded in the upper digestive tract before they can reach the gut-associated lymphoid tissue (GALT) [18]. Studies on B subunit labile toxin (LTB) suggest that plant-based oral vaccines can significantly boost mucosal immune responses that have been primed by parenteralinjection [42].

One the most important problems in transgenic plants is low level production of recombinant immunogenic protein. To solve this problem, different strategies such as strong promoter, organelle targeting and organelle transformation have been used [17]. Furthermore, synthetic genes with plant codon optimization have been used to mimic highly expressed plant genes. The effective applications of synthetic genes in plants have been proven by other researchers [16].

Two types of vaccines are available against *E. coli* O157:H7: one is a genetically engineered vaccine tested on a small group of adult volunteers. It appears safe and stimulates the production of antibodies against the potentially fatal pathogen [43]. The other is *Econiche* (made
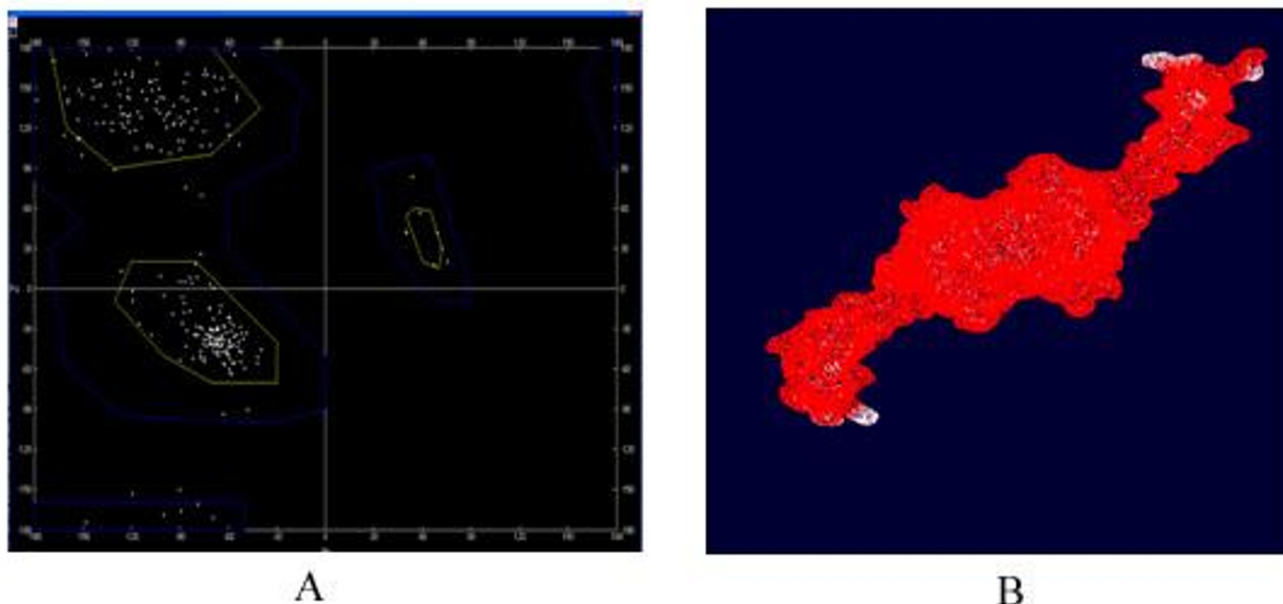
Figure 5
(A) Evaluation of model stability based on a Ramachandran plot and (B) energy minimization.

from an extract of lysed bacteria containing type III secretion proteins) for vaccination of healthy cattle as an aid in reducing shedding of *Escherichia coli* O157: H [44]. Both of these vaccines are high risk and are insufficiently safe and for this reason we attempted to design multi component antigens which can create protection and prevent colonization. This construct should contain essential antigenic factors of *E. coli* O157:H7 that are exposed completely.

On the basis of knowledge of molecular modeling and immuno-informatics, a novel approach was employed to identify a set of peptides that could be used as a vaccine

either in natural or in synthetic form. This approach has been extended to the entire proteomes of other microorganisms such as T-cell epitopes of secretory proteins of *Mycobacterium tuberculosis* [45,46], Tertiary Structure of *Mycobacterium leprae* Hsp65 Protein [47], T-cell antigen of *Chlamydia* [48], tandem repeat antigens from *Leishmania donovani* [49], and Envelope Glycoprotein of *Japanese Encephalitis Virus* (JEV) [50] to identify new sets of potentially antigenic proteins.

Here we designed new constructs of EHEC antigens including EspA, Intimin and Tir that contained essential determinants for bacterial attachment and effacement.

Table 2: Epitopes predicted in chimeric protein by different parameters based on Bcepred software

Prediction parameters	Epitope positions*
Hydrophilicity	1-14, 25-38, 47-55, 108-115, 128-144, 160-166, 202-219, 222-230, 232-242, 262-268, 283-291, 301-309, 319-329, 392-404, 448-475, 482-490, 512-526, 528-547, 430-446.
Flexibility	4-10, 25-35, 43-51, 104-113, 199-214, 217-226, 279-287, 307-314, 316-325, 389-403, 447-453, 480-485, 539-545.
Accessibility	2-18, 27-42, 45-55, 81-87, 95-101, 113-120, 128-144, 147-155, 157-166, 169-177, 179-191, 201-217, 250-259, 276-282, 289-298, 319-331, 340-349, 374-384, 391-401, 430-463, 467-493, 510-551.
Exposed surface	28-42, 251-259, 340-346, 392-398, 450-457, 472-479, 482-490, 520-527, 530-550.
Polarity	32-39, 128-144, 157-164, 249-259, 430-446, 473-480, 510-526, 533-552.
Antigenic propensity	38-44, 112-119, 174-180, 312-319, 352-360, 363-370, 413-419, 498-508.

* Number shows position of amino acids.

Table 3: One hundred and eighteen discontinuous B-Cell epitopes of chimeric protein predicted by the Discotope server

Start & End position	Start & End position	Start & End position	Start & End position	Start & End position	Start & End position	Start & End position	Start & End position	Start & End position
49-6	174-19	202-19	252-12	312-10	351-10	377-11	389-8	412-14
50-13	175-21	204-16	253-17	324-14	359-11	378-10	390-5	425-5
51-13	176-12	229-8	254-4	325-13	361-13	379-12	391-8	426-5
52-17	177-11	231-6	255-8	326-9	364-14	380-13	392-11	427-7
54-15	178-13	232-7	256-8	327-13	367-12	381-11	402-12	433-11
55-16	179-21	242-12	257-9	328-12	368-8	382-10	403-7	439-11
69-9	189-10	243-12	258-9	329-12	369-9	383-8	404-9	
70-8	193-12	244-16	259-9	330-12	370-9	384-5	405-10	
72-9	194-12	245-13	260-11	331-14	371-12	385-6	406-8	
89-15	195-17	246-12	276-10	345-7	372-10	386-7	407-8	
90-14	196-14	247-11	277-12	346-7	373-11	387-6	408-11	
136-8	197-13	249-18	281-12	347-7	374-12	388-10	409-10	
137-7	200-19	250-17	286-12	349-7	375-12	389-8	410-15	
138-9	201-21	251-13	287-8	350-11	376-12	388-10	411-14	

Theoretically, the DNA fragment consisted of these three putative antigens and could be synthesized as a unique construct optimally suited for expression in a plant system. Several factors which can affect the expression of foreign genes in plant systems such as messenger RNA instability [51], premature polyadenylation [52], abnormal splicing [53], and improper codon usage have been reported [54]. In order to increase the mRNA stability, DNA motifs that might contribute to mRNA instability in plants, such as the ATTTA sequence and the potential polyadenylation signal sequence AATAAA, were eliminated from the synthetic gene (for detail see Table 1). The synthetic DNA fragment which encoded the mature chimeric gene was constructed based on the codon usage of highly expressed nuclear-encoded genes of tobacco (*Nicotiana tobaccum L.)* as a model, and canola (*Brassica napus L.*) as the final target plant [55].

The efficiency of heterologous protein production can be diminished by biased codon usage. Approaches normally used to overcome this problem include targeted mutagenesis to remove rare codons or the addition of rare codon tRNAs in specific cell lines. Recently, improvements in the

technology have enabled synthetic genes to be produced cost-effectively, making this a feasible alternative [56]. In addition, as each step in the process of gene expression, from the transcription of DNA into mRNA to the folding and posttranslational modification of proteins, is regulated by complex cellular mechanisms, a relationship is expected to exist between mRNA expression levels and protein solubility in the cell. By formulating a relation between the mRNA expression level and the recombinant protein, production can be reasonably predicted [57].

In eukaryotic mRNA, the consensus sequence surrounding the start codon (Kozak seq. 5'GCC ACCATGGC) can increase the correctness and efficiency of translation up to 10 fold. In the synthetic construct, the 5'GCCACC sequence was added before the ATG codon. The second codon following the initial methionine was Ala, encoded by the codon GCT, and the necessary GC was provided; therefore there was no need to replace the other nucleotides or amino acids [27]. Codons that are rarely used in plants, such as XCG and XUA (X denotes U, C, A, or G), were avoided in the construction of the synthetic gene (Figure 2B). It has been reported that rare codons in

mRNA tend to form higher-order secondary structures, which might require additional time for ribosomal movement through the critical region [58].

An ideally biased gene would show a codon adaptation index (CAI) of 1.0. Even though no natural plant gene reaches this theoretical value, this index was increased from 76% in the wild type chimeric sequence to 98% in this synthetic gene. Furthermore, the G/C ratio and distribution were balanced from 41.59 to 40.96 percent with no significant changes, and this has been reported to be associated with low mRNA stability and expression in higher plants [55]. The nucleotide that encodes the ER retention signal (KDEL) which helps to accumulate the recombinant protein inside the endoplasmic reticulum was fused in-frame at the 3' end of the chimeric gene [15,16]. Finally, the required restriction enzyme sites (*Xba*I and *Sac*I) were introduced at the ends of the synthetic gene for future cloning into plant expression vectors.

Graphical depiction of the predicted minimum free energy for the synthetic gene showed that the average energy minimization was near - 400 Kcal/mol.

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.

In the protein structure prediction, the chimeric protein formed three domains that were separated by two main α-helix moieties which could help the protein to form a final structure. These α-helix structures are related to the designation of special amino acid sequences, residues 129-148 and 431-450, which are inserted between domains. With these results we could speculate that these parts could support the stable structure of a protein which contained three domains.

B-cell epitopes for the chimeric protein could be predicted on the basis of the structural prediction and solvent accessibility. Hopp and Woods in the 1980s developed a method for predicting B-cell epitopes with hydrophilicity parameters. Since then, several distinct methods such as Hydrophilicity method, Accessibility method, Antigenicity method, Flexibility method and secondary structure analysis have been developed [36,39]. Applying just one of these methods is not enough for obtaining results good enough to predict the B-cell epitope. In this study, we combined all the data obtained by these analyses and predicted the B-cell epitopes.

The integrated results showed that the most likely B-cell epitopes of this chimeric protein, as shown in Table 2, were located in three distinct parts, selected as the EspA, Intimin, and Tir domains.

For eliciting an immune response against *E. coli* O157:H7, studies have shown that production of the carboxy terminal part of Intimin in a transgenic plant cell line and its application via the oral route is more effective than injection [16]. In this study, we designed a multi domain antigen which was selected on the basis of three immunogenic parts of attaching/effacing loci from *E. coli* O157:H7, which were then optimized upon plant codon preference for analyzing mucosal and systematic immunity.

Conclusion

Bioinformatics tools for predicting epitopes are now a standard methodology. In silico epitope mapping, combined with in vitro and in vivo verification, accelerates the discovery process by approximately 10-20-fold. Development of sophisticated bioinformatics tools will provide a platform for more in-depth analysis of immunological data and facilitate the construction of new hypotheses to explain the complex immune system function [59].

In this study, we have combined several techniques and profiles to improve the state-of-the-art prediction of 3D structure and relative solvent accessibility. Building a homology model for this chimeric protein has been used to understand the antigenic sites and structural conformation domains which were used to predict continuous and discontinuous epitopes. Also, for the antibody-antigen interaction, it is important to know how much area of surface is exposed; accordingly we defined the exposed areas and surface accessibility.

Considering the multi colonization factor of this bacterium, multi antigenic parts should be used for repressing this pathogen. For this reason, more research should focus on designing multi antigenic proteins from *E. coli* O157:H7. This study and a few others [60,61] indicate that epitope construction and prediction will be useful not only in vaccine development but also in the prospective engineering and re-engineering of protein therapeutics, reducing the risk of undesired immunogenecity and improving the likelihood of success in clinical use.

Finally, the conclusions drawn for *E. coli* O157:H7 proteins could be combined with expression profiling to identify genes whose expression changes under shifting environmental conditions [62].

In conclusion, we believe that all of these findings will intensify efforts to develop a vaccine candidate against *E. coli* O157:H7.

Methods

Sequence analysis

Related sequences for *espA* (40 sequences), *eae* (32 sequences) and *tir* (50 sequences) were obtained from Genbank (accession no. not shown). Multiple sequence alignments were performed using ClustalW software (EBI, UK) http://www.ebi.ac.uk/Tools/clustalw2/ in order to identify a fragment common to all the sequences.

Construct design

An antigenic sequence was constructed by fusing the C-terminal of *espA*, C-terminal of *eae* and middle fragment of *tir* using hydrophobic amino acid linkers (accession no. GQ205376).

The *in silico* gene analysis and multi parameter gene optimization of the synthetic chimera gene was performed using Stand-alone softwares such as Leto (Entelechon, Germany), DNA 2.0 http://www.dnatwopointo.com, DNAsis MAX (Hitachi Software), and online data bases and softwares such as the codon database http://www.kazusa.or.jp/codon, Gene bank codon data base and Swissprot reverse translation online tool http://www.bioinformatics.org/sms2/rev_trans.html. The desired properties were verified by Gen-Script (NJ, USA). The multimeric gene was synthesized by ShineGene Molecular Biotech, Inc (Shanghai, China).

Bioinformatic analysis of chimeric recombinant protein

The messenger RNA secondary structure of the chimeric gene was analyzed by the program mfold http://www.bioinfo.rpi.edu/applications/mfold. Recombinant protein Secondary-structure predictions were performed by the neural-network-based algorithm program (PHD), and for 3D structure, online *ab initio* software was used http://www.igb.uci.edu/[63]. 3D structural stability of the synthetic protein was further analyzed by Swiss-PdbViewer for energy minimization [64]. Solvent accessibility of different residues was evaluated by DSSP and other online programs (VADAR) http://redpoll.pharmacy.ualberta.ca/vadar/. The predictive value of the hyper glycosylation code which may act in plants is well established based on online software http://www.cbs.dtu.dk/services/[65].

Prediction of B-cell epitopes

The amino acid sequence was analyzed using three web-based B-cell epitope prediction algorithms; Bcepred http://www.imtech.res.in/raghava/bcepred/, Continuous B cell epitopes prediction methods based on physico-chemical properties on a non-redundant dataset, and the Discotope http://www.cbs.dtu.dk/services/DiscoTope/ Server for predicting discontinuous B cell epitopes from three-dimensional protein structures. Briefly, chimeric proteins were analyzed first for continuous B-cell epitopes using

Bcepred and then using the Discotope server to predict discontinuous B cell epitopes. Finally, we used the VaxiJen server to predict the immunogenecity of the whole antigen and its subunit vaccine [48,66,67].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All four authors (JA, SLM, SR, AHS) contributed equally to this manuscript. All authors read and approved the final manuscript.

Acknowledgements

References

1. Van Diemen PM, Dziva F, Abu-Median A, Wallis TS, Bosch H Van den, Dougan G, *et al.*: **Subunit vaccines based on intimin and Efa-1 polypeptides induce humoral immunity in cattle but do not protect against intestinal colonisation by enterohaemorrhagic Escherichia coli O157:H7 or O26:H.** *Vet Immunol Immunopathol* 2007, **116(1-2):**47-58.
2. Babiuk S, Asper DJ, Rogan D, Mutwiri GK, Potter AA: **Subcutaneous and intranasal immunization with type III secreted proteins can prevent colonization and shedding of Escherichia coli O157:H7 in mice.** *Microb Pathog* 2008, **45(1):**7-11.
3. Li Y, Frey E, Mackenzie AM, Finlay BB: **Human response to Escherichia coli O157:H7 infection: antibodies to secreted virulence factors.** *Infect Immun* 2000, **68(9):**5090-5095.
4. McNeilly TN, Naylor SW, Mahajan A, Mitchell MC, McAteer S, Deane D, *et al.*: **Escherichia coli O157:H7 colonization in cattle following systemic and mucosal immunization with purified H7 flagellin.** *Infect Immun* 2008, **76(6):**2594-2602.
5. Van Donkersgoed J, Hancock D, Rogan D, Potter AA: **Escherichia coli O157:H7 vaccine field trial in 9 feedlots in Alberta and Saskatchewan.** *Can Vet J* 2005, **46(8):**724-28.
6. Kühne SA, Hawes WS, La Ragione RM, Woodward MJ, Whitelam GC, Gough KC: **Isolation of recombinant antibodies against EspA and intimin of Escherichia coli O157:H7.** *J Clin Microbiol* 2004, **42(7):**2966-76.
7. Garrido P, Blanco M, Moreno-Paz M, Briones C, Dahbi G, Blanco J, Blanco J, Parro V: **STEC-EPEC Oligonucleotide Microarray: A New Tool for Typing Genetic Variants of the LEE Pathogenicity Island of Human and Animal Shiga Toxin-Producing Escherichia coli (STEC) and Enteropathogenic E. coli (EPEC) Strains.** *Clin Chem* 2006, **52(2):**192-201.
8. China B, Jacquemin E, Devrin AC, Pirson V, Mainil J: **Heterogeneity of the eae genes in attaching/effacing Escherichia coli from cattle: comparison with human strains.** *Res Microbiol* 1999, **150(5):**323-32.
9. La Ragione RM, Patel S, Maddison B, Woodward MJ, Best A, Whitelam GC, Gough KC: **Recombinant anti-EspA antibodies block Escherichia coli O157:H7-induced attaching and effacing lesions in vitro.** *Microbes Infect* 2006, **8(2):**426-33.
10. Paton AW, Manning PA, Woodrow MC, Paton JC: **Translocated intimin receptors (Tir) of Shiga-toxigenic Escherichia coli isolates belonging to serogroups O26, O111, and O157 react with sera from patients with hemolytic-uremic syndrome and exhibit marked sequence heterogeneity.** *Infect Immun* 1998, **66(11):**5580-6.
11. Goffaux F, China B, Dams L, Clinquart A, Daube G: **Development of a genetic traceability test in pig based on single nucleotide polymorphism detection.** *Forensic Sci Int* 2005, **151(2-3):**239-47.

12. Yuste M, Orden JA, De La Fuente R, Ruiz-Santa-Quiteria JA, Cid D, Martínez-Pulgarín S, Domínguez-Bernal G: **Polymerase chain reaction typing of genes of the locus of enterocyte effacement of ruminant attaching and effacing Escherichia coli.** *Can J Vet Res* 2008, **72(5):**444-48.

13. Cleary J, Lai LC, Shaw RK, Straatman-Iwanowska A, Donnenberg MS, Frankel G, Knutton S: **Enteropathogenic Escherichia coli (EPEC) adhesion to intestinal epithelial cells: role of bundle-forming pili (BFP), EspA filaments and intimin.** *Microbiology* 2004, **150(3):**527-38.

14. Dean-Nystrom EA, Gansheroff LJ, Mills M, Moon HW, O'Brien AD: **Vaccination of pregnant dams with intimin(O157) protects suckling piglets from Escherichia coli O157:H7 infection.** *Infect Immun* 2002, **70(5):**2414-8.

15. Kang TJ, Han SC, Jang MO, Kang KH, Jang YS, Yang MS: **Enhanced expression of B-subunit of Escherichia coli heat-labile enterotoxin in tobacco by optimization of coding sequence.** *Appl Biochem Biotechnol* 2004, **117(3):**175-87.

16. Judge NA, Mason HS, O'Brien AD: **Plant cell-based intimin vaccine given orally to mice primed with intimin reduces time of Escherichia coli O157:H7 shedding in feces.** *Infect Immun* 2004, **72(1):**168-75.

17. Schillberg S, Twyman RM, Fischer R: **Opportunities for recombinant antigen and antibody expression in transgenic plants-technology assessment.** *Vaccine* 2005, **23(15):**1764-9.

18. Lal P, Ramachandran VG, Goyal R, Sharma R: **Edible vaccines: current status and future.** *Indian J Med Microbiol* 2007, **25(2):**93-102.

19. Suo G, Chen B, Zhang J, Duan Z, He Z, Yao W, Yue C, Dai J: **Effects of codon modification on human BMP2 gene expression in tobacco plants.** *Plant Cell Rep* 2006, **25(7):**689-97.

20. Mechold U, Gilbert C, Ogryzko V: **Codon optimization of the BirA enzyme gene leads to higher expression and an improved efficiency of biotinylation of target proteins in mammalian cells.** *J Biotechnol* 2005, **116(3):**245-49.

21. Lim LH, Li HY, Cheong N, Lee BW, Chua KY: **High-level expression of a codon optimized recombinant dustmite allergen, Blot5, in Chinese hamster ovary cells.** *Biochem Biophys Res Commun* 2004, **316(4):**991-96.

22. Gustafsson C, Govindarajan S, Minshull J: **Codon bias and heterologous protein expression.** *Trends Biotechnol* 2004, **22(7):**346-353.

23. Batchelor M, Prasannan S, Daniell S, Reece S, Connerton I, Bloomberg G, et al.: **Structural basis for recognition of the translocated intimin receptor (Tir) by intimin from enteropathogenic Escherichia coli.** *EMBO J* 2000, **19(11):**2452-64.

24. Frankel G, Candy DC, Everest P, Dougan G: **Characterization of the C-terminal domains of intimin-like proteins of enteropathogenic and enterohemorrhagic Escherichia coli, Citrobacter freundii, and Hafnia alvei.** *Infect Immun* 1994, **62(5):**1835-42.

25. Hartland EL, Batchelor M, Delahay RM, Hale C, Matthews S, Dougan G, et al.: **Binding of intimin from enteropathogenic Escherichia coli to Tir and to host cells.** *Mol Microbiol* 1999, **32(1):**151-8.

26. 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* 2001, **14(8):**529-32.

27. Kozak M: **The scanning model for translation: an update.** *J Cell Biol* 1989, **108(2):**229-41.

28. Graf M, Deml L, Wagner R: **Codon-optimized genes that enable increased heterologous expression in mammalian cells and elicit efficient immune responses in mice after vaccination of naked DNA.** *Methods Mol Med* 2004, **94:**197-210.

29. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31(13):**3406-15.

30. Garnier J, Gibrat JF, Robson B: **GOR method for predicting protein secondary structure from amino acid sequence.** *Methods Enzymol* 1996, **266:**540-53.

31. Rost B, Sander C, Schneider R: **PHDsec an automatic mail server for protein secondary structure prediction.** *Comput Appl Biosci* 1994, **10(1):**53-60.

32. Yang S, Onuchic JN, García AE, Levine H: **Folding time predictions from all-atom replica exchange simulations.** *J Mol Biol* 2007, **372(3):**756-63.

33. Ginalski K: **Comparative modeling for protein structure prediction.** *Curr Opin Struct Biol* 2006, **16(2):**172-7.

34. Rost B, Sander C: **Conservation and prediction of solvent accessibility in protein families.** *Proteins* 1994, **20(3):**216-26.

35. Parker JM, Guo D, Hodges RS: **New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites.** *Biochemistry* 1986, **25(19):**5425-32.

36. Kolaskar AS, Tongaonkar PC: **A semi-empirical method for prediction of antigenic determinants on protein antigens.** *FEBS Lett* 1990, **276(1-2):**172-4.

37. Ponnuswamy PK, Prabhakaran M, Manavalan P: **Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins.** *Biochim Biophys Acta* 1980, **623(2):**301-16.

38. Saha S, Bhasin M, Raghava GP: **Bcipep: a database of B-cell epitopes.** *BMC Genomics* 2005, **6(1):**79.

39. Karpman D, Békássy ZD, Sjögren AC, Dubois MS, Karmali MA, Mascarenhas M, et al.: **Antibodies to intimin and Escherichia coli secreted proteins A and B in patients with enterohemorrhagic Escherichia coli infections.** *Pediatr Nephrol* 2002, **17(3):**201-11.

40. Julia Scerbo M, Bibolini MJ, Barra JL, Roth GA, Monferran CG: **Expression of a bioactive fusion protein of Escherichia coli heat-labile toxin B subunit to a synapsin peptide.** *Protein Expr Purif* 2008, **59(2):**320-6.

41. Gerdts V, et al.: **Mucosal delivery of vaccines in domestic animals.** *Vet Res* 2006, **487(37):**487-510.

42. Lauterslager TG, Florack DE, Wal TJ van der, Molthoff JW, Langeveld JP, Bosch D, et al.: **Oral immunisation of naive and primed animals with transgenic potato tubers expressing LT-B.** *Vaccine* 2001, **19(17-19):**2749-55.

43. Stephenson J: **E coli O157 Vaccine.** *JAMA* 1998, **279(11):**818-b.

44. Potter AA, Klashinsky S, Li Y, Frey E, Townsend H, Rogan D, Erickson G, Hinkley S, Klopfenstein T, Moxley RA, Smith DR, Finlay BB: **Decreased shedding of Escherichia coli O157:H7 by cattle following vaccination with type III secreted proteins.** *Vaccine* 2004, **22:**362-369.

45. Mustafa AS: **Recombinant and synthetic peptides to identify Mycobacterium tuberculosis antigens and epitopes of diagnostic and vaccine relevance.** *Tuberculosis* 2005, **85(5-6):**367-76.

46. Vani J, Shaila MS, Chandra NR, Nayak R: **A combined immuno-informatics and structure-based modeling approach for prediction of T cell epitopes of secretory proteins of Mycobacterium tuberculosis.** *Microbes Infect* 2006, **8(3):**738-46.

47. Rossetti RAM, Lorenzi JCC, Giuliatti S, Silva CL, Coelho CAAM: **In Silico Prediction of the Tertiary Structure of M. leprae Hsp65 Protein Shows an Unusual Structure in Carboxy-terminal Region.** *J Comp Sci Syst Biol* 2008, **1:**126-131.

48. Barker CJ, Beagley KW, Hafner LM, Timms P: **In silico identification and in vivo analysis of a novel T-cell antigen from Chlamydia, NrdB.** *Vaccine* 2008, **26(10):**1285-96.

49. Goto Y, Coler RN, Reed SG: **Bioinformatic identification of tandem repeat antigens of the Leishmania donovani complex.** *Infect Immun* 2007, **75(2):**846-51.

50. Kolaskar AS, Kulkarni-Kale U: **Prediction of three-dimensional structure and mapping of conformational epitopes of envelope glycoprotein of Japanese encephalitis virus.** *Virology* 1999, **261(1):**31-42.

51. Murray EE, Rocheleau T, Eberle M, Stock C, Sekar V, Adang M: **Analysis of unstable RNA transcripts of insecticidal crystal protein genes of Bacillus thuringiensis in transgenic plants and electroporated protoplasts.** *Plant Mol Biol* 1991, **16(6):**1035-50.

52. Jarvis P, Belzile F, Dean C: **Inefficient and incorrect processing of the Ac transposase transcript in iae1 and wild-type Arabidopsis thaliana.** *Plant J* 1997, **11(5):**921-31.

53. Haseloff J, Siemering KR, Prasher DC, Hodge S: **Removal of a cryptic intron and subcellular localization of green fluorescent protein are required to mark transgenic Arabidopsis plants brightly.** *Proc Natl Acad Sci USA* 1997, **94(6):**2122-7.

54. Perlak FJ, Fuchs RL, Dean DA, McPherson SL, Fischhoff DA: **Modification of the coding sequence enhances plant expression of insect control protein genes.** *Proc Natl Acad Sci USA* 1991, **88(8):**3324-8.

55. Campbell WH, Gowri G: **Codon Usage in Higher Plants, Green Algae, and Cyanobacteria.** *Plant Physiol* 1990, **92(1):**1-11.

56. Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O: **Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study.** *Protein Expr Purif* 2008, **59(1):**94-102.

57. Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M: **A Relationship between mRNA Expression Levels and Protein Solubility in E. coli.** *JMol Biol* 2009, **388(2):**381-89.

58. Thanaraj TA, Argos P: **Ribosome-mediated translational pause and protein domain organization.** *Protein Sci* 1996, **5(8):**1594-612.

59. ChuanTong J, Tan TW, Ranganathan S: **Methods and protocols for prediction of immunogenic epitopes.** *BRIEF BIOINFORM* 2006, **8(2):**96-108.

60. Barbosa MD, Vielmetter J, Chu S, Smith DD, Jacinto J: **Clinical link between MHC class II haplotype and interferon-beta (IFN-beta) immunogenicity.** *Clin Immunol* 2006, **118(1):**42-50.

61. Koren E, De Groot AS, Jawa V, Beck KD, Boone T, Rivera D, Li L, Mytych D, *et al.*: **Clinical validation of the "in silico" prediction of immunogenicity of a human recombinant therapeutic protein.** *Clin Immunol* 2007, **124(1):**26-32.

62. Allen TE, Herrgard MJ, Liu M, Qiu Y, Glasner JD, Blattner FR, Palsson B: **Genome-Scale Analysis of the Uses of the Escherichia coli Genome: Model-Driven Analysis of Heterogeneous Data Sets.** *J BACTERIOL* 2003, **185(21):**6392-6399.

63. Langedijk JP, Daus FJ, van Oirschot JT: **Sequence and structure alignment of Paramyxoviridae attachment proteins and discovery of enzymatic activity for a morbillivirus hemagglutinin.** *J Virol* 1997, **8:**6155-67.

64. Edwards YJ, Cottage A: **Bioinformatics methods to predict protein structure and function. A practical approach.** *Mol Biotechnol* 2003, **23(2):**139-66.

65. Xu J, Tan L, Lamport DT, Showalter AM, Kieliszewski MJ: **The O-Hyp glycosylation code in tobacco and Arabidopsis and a proposed role of Hyp-glycans in secretion.** *Phytochemistry* 2008, **69(8):**1631-40.

66. Davies MN, Flower DR: **Harnessing bioinformatics to discover new vaccines.** *Drug Discov Today* 2007, **12(9-10):**389-95.

67. Korber B, LaBute M, Yusim K: **Immunoinformatics comes of age.** *PLoS Comput Biol* 2006, **2(6):**e71.