African Journal of Biotechnology Vol. 11(8), pp. 2058-2066, 26 January, 2012
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.1843
ISSN 1684–5315 © 2012 Academic Journals

Full Length Research Paper

Bioinformatic analysis to discover putative drug targets against diarrheal causative agents

Muhammad Shakil Khan, Sana Zahra and Hamid Rashid*

Department of Bioinformatics, Mohammad Ali Jinnah University, Islamabad, Pakistan.

Accepted 23 September, 2011

Availability of genome sequences of pathogens has offered a tremendous amount of information that can be useful in drug target identification. Complete genome sequences of several pathogenic bacteria including *Shigella* spp., *Escherichia coli*, *Vibrio cholerae*, *Salmonella typhimurium* and *Yersinia enterocolitica* mostly involved in causing diarrheal diseases have been determined. Detection of bacterial genes that are common in all of them, non-homologous to human genes and essential for the survival of the pathogen represents a promising means of identifying novel drug targets. Results recommended that among these common proteins, surface associated proteins might be most useful. Our approach has identified seven essential proteins that may be considered as potential drug targets; motifs were identified for these proteins to determine their functions and antigenicity and profiling was also done to identify probable epitopes among the candidate antigens. This approach enables rapid potential drug target identification, thereby greatly facilitating the search for new drug targets against the causative agents of diarrheal diseases. These results highlight the significance of *in silico* systematic drug target identification in the post-genomic era.

Key words: Diarrheal diseases, motifs, antigenicity profiling.

INTRODUCTION

Diarrhea remains the second leading cause of death around the world for children under five years of age, accounting for approximately 15% of under-five child deaths worldwide, or almost two million deaths annually (WHO, 2008). Diarrhea is the passage of loose or watery stools occurring three or more times in a 24 h period. The three types of diarrhea are: acute diarrhea, which lasts less than 14 days, persistent diarrhea, which lasts 14 days or more, and dysentery Diarrhea with blood in the stool with or without mucus. The new approach of the genomic era, to develop vaccines starting from the genomic information rather than growing the causative microorganism, was named reverse vaccinology (Rappuoli, 2000) and can be used for the development of vaccines against pathogens for which the applications of Pasteur's principles have failed.

In the post-genomic era, pathogen genome sequencing efforts have expanded in order to include multi-representatives of the same species and this pan-genome approach has shown tremendous potential for making vaccines that once might have been impossible to design. Comparative genomics allows antigen candidates to be identified on the basis of sequence conservation in different serotypes and strains of a given pathogen. Identification of shared virulence factors and protective antigens could support a concept of combined vaccination approaches (Zagursky et al., 2003). Although one of the most important features of effective vaccine antigens is their ability to induce antibodies and or to activate immune cells, it is anticipated that with the wealth of knowledge currently being generated, it would be possible to develop predictive algorithms to pinpoint proteins that are likely to be immunogenic and/or protective (Mayers et al., 2003).

More advanced however, is the strategy to mine genomic sequence databases of intracellular pathogens for predicted T-cell epitopes and validate them experimentally, on the basis of immune recognition (Flower, 2003). Detection of bacterial genes that are common in all of them, non-homologous to human genes and are essential for the survival of the pathogen,

*Corresponding author. E-mail: drhamid@jinnah.edu.pk. Tel: 111 87 87 87 ext: 136.

represents a promising means of identifying novel drug targets. Results recommend that among these common proteins, surface associated proteins might be most useful. Our approach has identified seven essential proteins that may be considered as potential drug targets. Motifs were identified for these proteins to determine their functions and antigenicity profiling was done to identify probable epitopes among the candidate antigens. This approach enables rapid potential drug target identification, thereby greatly facilitating the search for new drug targets against the causative agents of diarrheal diseases. These results also highlight the significance of *in silico* systematic drug target identification in the post-genomic era.

MATERIALS AND METHODS

Retrieval of DNA sequences

Complete DNA sequences were mostly retrieved from www.ncbi.nlm.nih.gov. In the case of the five bacterial species, section related to core nucleotides was selected for sequence analysis.

Retrieving coding sequences from genomic databases

DNA and protein encoding DNA sequences were taken and analyzed for the outer membrane, lipopolysaccharide, adhesins, and siderophores (iron uptake). Separate datasets were made in which these proteins were stored locally. For selecting consensus sequences, open reading frames (ORF) of these surface associated proteins were retrieved from http://spock.genes.nig.ac.jp/~genome/org.html.

Protein-protein BLAST

Protein-protein BLAST (Altschul et al., 1990) was performed by using alignment tools. Dataset of all the surface associated proteins (ORFs of a particular bacterial species) were BLAST against all other selected bacterial species in order to obtain similar sequences. This procedure was repeated for selected proteins for example, outer membrane proteins of *Escherichia coli* were analyzed by BLAST one by one against the outer membrane proteins (obtained from database) of the rest of the bacterial species (http://blast.ncbi.nlm.nih.gov/Blast. cgi?PROGRAM=blastp &BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW _DEFAULTS=on&LINK_LOC=blasthome).

Selection of proteins from alignment files

Sequences were selected by observing the alignment criteria; bit score, expected value (E value) and % identities. However, as the E value varies from sequence to sequence depending upon their length, % identity was taken as the selection criteria whose range falls between 99 to 100% using both the BioEdit and CLC protein workbench software packages. More refinement was made to improve the alignment so that only those sequences whose % identity was almost 100%, that is either 99% or 100%, were selected.

Expression profile data comparison

In silico, two-dimensional gel electrophoresis (2-DE) patterns were observed in order to study the expression profile of each species. For this, two separate websites were used; GELBANK and JVIRTUAL GEL. GELBANK was available from the NCBI FTP server. This website incorporates only completed genomes and information pertinent to 2-DE. Link is available at www.gelbank.anl.gov. JVirGel is a software for the simulation and analysis of proteomics data (http://www.jvirgel.de/). The Java TM based software determines the theoretical isoelectric points (pI), calculates molecular weights (MW) of proteins and visualizes these as a virtual two-dimensional (2D) protein map.

Selection of common proteins

Common proteins were selected from the data; the proteins that were common in all bacterial species, were selected by using BLAST (Altschul et al., 1990) and the proteins that showed 99 to 100% identity were included in the list.

Prediction of protein sub-cellular localization

A protein's subcellular localization is influenced by several features present within the protein's primary structure, such as the presence of membrane-spanning alpha-helices (http://psort.hgc.jp/). PSORT uses information obtained from each analysis to generate an overall prediction of localization site. Developed by Kenta Nakai in 1991, PSORT is an algorithm which assigns a probable localization site to a protein given an amino acid sequence alone. Originally developed for prediction of protein localization in Gram-negative bacteria, PSORT was expanded into a suite of programs PSORT (Nakai and Kanehisa, 1991), PSORT II (Horton and Nakai, 1997) and iPSORT (Bannai et al., 2002) capable of handling proteins from all classes of organisms.

Identification of protein domains and function

Domain in a protein is important for the functional ability of a protein. SBASE is a collection of protein domain sequences collected from the literature, from protein sequence databases and from genomic databases. In addition, function of a particular protein can also be identified (http://hydra.icgeb.trieste.it/sbase/).

Transmembrane helices by hidden Markov model

For prediction of transmembrane helices, TMHMM version 2.0 (Krogh et al., 2001) was used (http://www.cbs.dtu.dk/services/ TMHMM/).

Discrimination and prediction of transmembrane beta strands in outer membrane proteins

A statistical method was devised by scientists based on amino acid composition for discriminating outer membrane proteins from other globular and membrane proteins (Gromiha et al., 2004). Furthermore, the method was able to correctly exclude alpha-helical membrane proteins up to an accuracy of 80% (http://psfs.cbrc.jp/ tmbeta-net/).

Motif identification

MEME (Bailey et al., 2010) I s basically a software package to

Table 1. Genome size of bacterial species causing diarrheal diseases.

Organism	Species	Genome size (bp)	DNA	Database	Accession
E. coli	O157:H7 EDL933	5528445	Circular	www.ncbi.nlm.nih.gov	NC_002655
S. typhimurium	LT2	4857432	Circular	www.ncbi.nlm.nih.gov	NC_003197
V. cholera	O1 Biovar eltor str. N16961	2961149	Circular	www.ncbi.nlm.nih.gov	NC_002505,06
Y. enterocolitica	Subspecies 8081	4615899	Circular	www.ncbi.nlm.nih.gov	NC_008800
S. flexneri	2a str. 2457T	4599354	Circular	www.ncbi.nlm.nih.gov	NC_004741
S. dysenteriae	Sd197	4369232	Circular	www.ncbi.nlm.nih.gov	NC_007606
S. sonnei	Ss046	4825265	Circular	www.ncbi.nlm.nih.gov	NC_007384
S. boydii	Sb227	4519823	Circular	www.ncbi.nlm.nih.gov	NC_007613

discover motifs (highly conserved regions) in groups of related DNA or protein sequences and search sequence databases using known and well defined motifs (http://meme.nbcr.net/meme/intro.html).

BioEdit

BioEdit is a biological sequence editor that runs in Windows 95/98/NT/2000/XP and is intended to provide basic functions for protein and nucleic acid sequence editing, alignment, manipulation and analysis. In this study it was mainly used for alignments that is, local alignment (Altschul et al., 1990), multiple sequence alignment (Thompson et al., 1994) and identification of common proteins. This software has the inbuilt functionality that it creates its own local database (Hall, 1999).

BLAST

Local BLAST (Altschul et al., 1990) was performed in order to obtain common proteins on the basis of % identities and secondly it was performed for the reverse BLAST (protein-protein BLAST against each other) to improve the results.

CLC combine workbench

In combine workbench we mostly used the protein workbench as our project requirements. Protein workbench was used for antigenicity, hydrophobicity and secondary structure prediction.

RESULTS AND DISCUSSION

Table 1 shows organism used during this study, their species name, genome size, type of DNA, database used and their accession numbers. The surface associated proteins that were focused in this project belong to the following category: outer membrane, lipopolysaccharide, siderophores (iron uptake), secreted and adhesins proteins. These proteins were BLAST against each other using different bioinformatics tools in order to identify those proteins which were common. For example, protein-protein BLAST of outer membrane protein of *E. coli* was performed against all the other bacterial species. Those alignments were selected whose identity values were between 99 to 100%. Significant results were obtained for most of the species except *V. cholera,* and *Y. enterocolitica* in case of surface associated proteins. Expression profile of various proteins is important due to the fact that only data obtained from ORF's that is not sufficient. There is a possibility that ORFs that are under consideration are not expressed in that organism as through scientific data, it is clear if all ORFs are expressed at a particular time (www.gelbank.anl.gov).

More also, in order to study the *in silico* expression of surface associated proteins, two different computational tools were used which included GELBANK and JVIR GEL. Protein expression is the key step for *in silico* identification of vaccine targets because proteins that are not expressed will be insignificant. For protein expression, JVIR GEL was tested to be a more reliable tool due to the fact that it is easier to use than GELBANK, and users can also obtain a list of all the expressed proteins divided into secretory proteins, membrane and other proteins, hence no need for the user to click on every expressed spot one by one to check which protein it is.

List of common proteins

From the alignment files that were generated previously, a list was obtained containing the common proteins with respect to three species (*E. coli*, *Salmonella typhimurium* and *Shigella* spp.). Since the alignment regarding *V. cholera* and *Y. enterocolitis* generated no significant results, the list therefore does not satisfy the required criteria for these two organisms. This list was also generated manually by observing separate lists of common proteins along with their respective gene names.

PSORT and *SBASE prediction of common proteins*

Sub-cellular localization is important due to the fact that the function of a particular protein can be identified by it. The prediction programs PSORTB and SBASE can be used to predict protein localization in Gram-negative organisms. PSORTB and SBASE were used to predict all

Table 2. Identification of proteins causing diarrheal diseases using SBASE tool.

Gene name	Protein	PSORT prediction	Description
ompA	Outer membrane protein 3a	Outer Membrane Protein	Outer membrane protein
lspA	Prolipoprotein signal peptidase	Cytoplasmic Membrane Protein	Lipopolysaccharide
yaeC	putative lipoprotein	Unknown	Lipopolysaccharide
Pal	Peptidoglycan-associated lipoprotein	Outer Membrane Protein	Lipopolysaccharide
Lpp	Murein lipoprotein	Unknown	Lipopolysaccharide
vacJ	Lipoprotein precursor	Unknown	Lipopolysaccharide
feoB	Ferrous iron transport protein B	Cytoplasmic Membrane Protein	Siderophores

Table 3. SBASE prediction of common proteins causing diarrheal diseases.

Gene name	SBASE domain	SBASE function
ompA		
lspA		
yaeC	NLPA lipoprotein	Surface Antigen
pal		
Lpp	No domain predicted	Murein lipoprotein
vacJ	VacJ like lipoprotein	Surface lipoprotein
feoB		

Table 4. TM beta segments and number of transmembrane helices of diarrheal agents by Hidden Markov Model (HMM).

Gene name	Predicted tm-beta segment	TMHHMM
ompA	14	0
lspA	8	0
yaeC	10	4
pal	7	0
Lpp	2	0
vacJ	10	0
feoB	30	10

outer membrane, secreted, and lipoproteins. PSORTB and SBASE use different methods for prediction of localizations. PSORTB uses primary sequence analysis algorithms and gives a high precision, but low sensitivity prediction, while SBASE on the other hand utilizes analysis of text annotations for the closest homologue of any given query sequence and has medium-high precision and sensitivity; it thus makes a higher number of predictions. SBASE tool was therefore used in this project in order to obtain knowledge about domains and functions of common proteins. PSORTB predicted some of the common proteins to be unknown. SBASE tool was applied (regarding their domains and function) and the results are shown in Table 2. In the case of SBASE, the results obtained are shown in Table 3.

Prediction of membrane proteins

Membrane proteins play key roles in biological systems as pores, ion channels and receptors. Hidden Markov Model was used to simulate topological states of regions of a protein. These states include transmembrane helices, cytoplasmic l and non-cytoplasmic loops. The predicted transmembrane domains are shown in Table 4.

Number of motifs of common proteins

Motif is a highly conserved region in a protein and is responsible for determining the role and function of that protein. In this case, motifs were retrieved from MEME database and are shown in Table 5.

Table 5. Motif prediction for proteins causing diarrheal diseases.

Protein	Gene name	Number of motif
Prolipoprotein signal peptidase	*Isp A*	11
Putative lipoprotein	*yae C*	16
Peptidoglycan-associated lipoprotein	*pal*	11
Murein lipoprotein	*lpp*	5
Lipoprotein precursor	*vacJ*	15
Ferrous iron transport protein B	*feoB*	41
Outer membrane protein 3a	*ompA*	21

Antigenicity profile of common genes

Antigenicity profile was taken in order to determine the antigenic determinants, that recognizes the antibodies or epitopes that show affinity for antibodies. The common proteins were analyzed for antigenic determinants by using CLC combine workbench. Different antigenic plots were obtained against all the common proteins. The proteins which are more hydrophilic are more antigenic because they are localized inside the cell; towards the cytoplasm and away from the membrane (phospholipid

Table 6. Antigenic propensity and determinants for proteins causing diarrheal diseases.

Gene name	Average antigenic propensity	Antigenic determinant
ompA	1.0199	12
lspA	1.058	7
yaeC	1.0456	13
Pal	1.0085	5
Lpp	1.0215	2
vacJ	1.0182	8

Table 7. T Cell epitope prediction for proteins causing diarrheal diseases (http://www.imtech.res.in/raghava/mhcbn/map.html).

Gene name	Sequence	Position	Score	T cell epitope	Host organism MHC
ompA	YRFGQGEAAPVVAPAPAPAP	189	1	No matching peptide	Human
	GAGAFGGYQVNPYVGFEMGY	57	0.986		
	GWSQYHDTGFINNNGPTHEN	35	0.975		
	ARGMGESNPVTGNTCDNVKQ	297	0.965		
lspA	QSDEVYEAANKVFNGGAVKG	251	0.972	No matching peptide	Human
pal	GAGTGMDANGGNGNMSSEEQ	38	1	No matching peptide	Human
	DERGTPEYNISLGERRANAV	110	0.973		
yaeC	QSDEVYEAANKVFNGGAVKG	251	0.972	No matching peptide	Human
Lpp	NWSKRDRNTAIGAGAGALGG	26	0.983	No matching peptide	Human
vacJ	ANGGELKPQENPNAQAIQDD	225	0.993	No matching peptide	Human
	AWRDYVPQPARNGLSNFTGN	58	0.99		
	VMVNYFLQGDPYQGMVHFTR	83	0.973		
	LVGCASSGTDQQGRSDPLEG	15	0.969		
feoB	PIEASKGDGEMGTGAMGVMD	636	0.996	No matching peptide	Human

bilayer). In contrast, proteins which are hydrophobic in nature are away from the cytoplasm; towards the membrane (phospholipid bilayer), and are less or not antigenic. Moreover, B and T cell epitopes of common candidates were also studied and results are shown in Tables 6 and 7. Antigenic plot of each common protein are also shown in Figure 1a to g. The X plot indicates that the amino acids below the X-axis are hydrophobic, while that above the X-axis are hydrophilic.

Multiple sequence alignment of common proteins

Multiple sequence alignment was performed in different species to identify conservation among the common surface associated proteins (CLC protein workbench) and are shown in Figure 2. Common proteins are the ones that appeared in one species and were also common in other species for example, in the case of *E. coli*, list of proteins similar in *E. coli* and other four species was stated. Secondly, in the case of *Salmonella*, list was made regarding other three species excluding *E. coli* from the list because *E. coli*'s alignment with *Salmonella* was previously done in making the *E. coli* common proteins file. Same method was repeated with other species as well, but no significant results were obtained in the case of *V. cholerae* and *Y. enterocolitis* in all the alignment files. In general, the overall degree in conservation between the protein sequences of close homologs was about 93.56% for the sequenced representatives of *Escherichia* and *Shigella* species, 69.21% for *Shigella* and *Salmonella* and 53.31% for *Salmonella* and *Yersinia*. This shows the close relationship between the *Salmonella, E. coli* and *Shigella* species.

Furthermore, common proteins were then analyzed for their subcellular localization and their domains and functions. ASE was considered to be more accurate than PSORTB because PSORTB confirmed certain common proteins as unknown, while SBASE showed the domains and function of those proteins. In addition, using bioinformatics tools motifs of common proteins were identified. Motifs are sequence patterns that show the conservation among different protein sequences. Conserved regions are important in proteins that show the properties of protein folding. Antigenic profiling of common proteins was done by using CLC combine workbench. The antigenicity plots were generated to

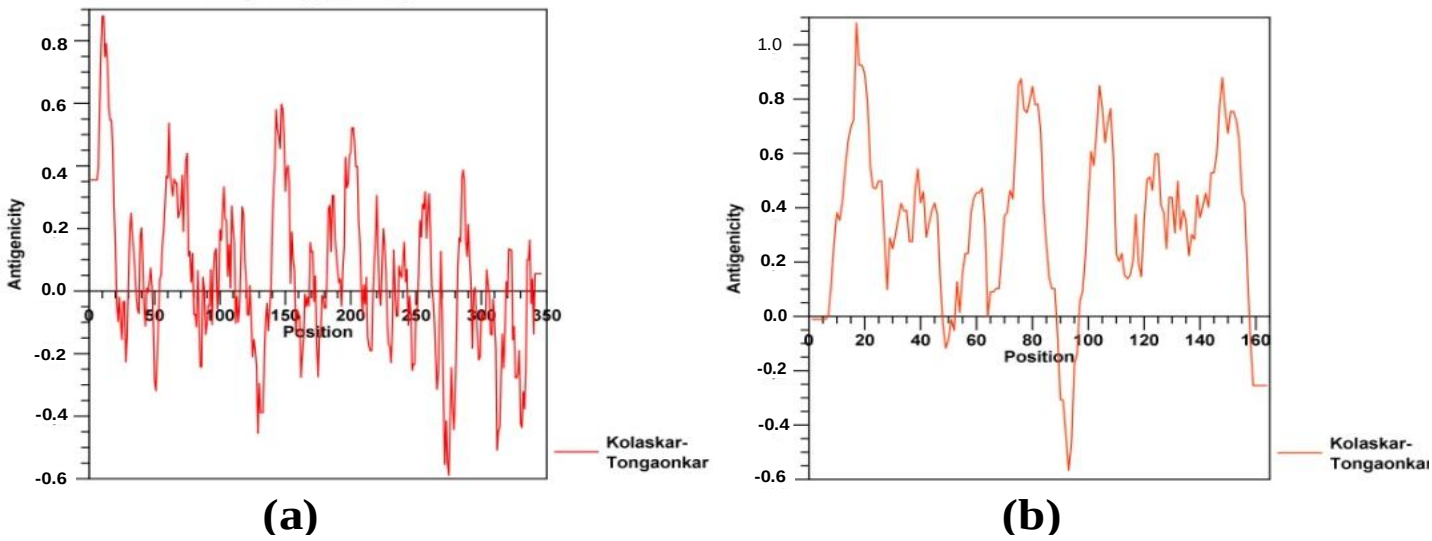

Figure 1. Antigenicity plots of common proteins causing diarrheal diseases. (a) ompA; (b) IspA; (c) Lpp; (d) Pal; (e) VacJ; (f) YaeC; (g) feoB.

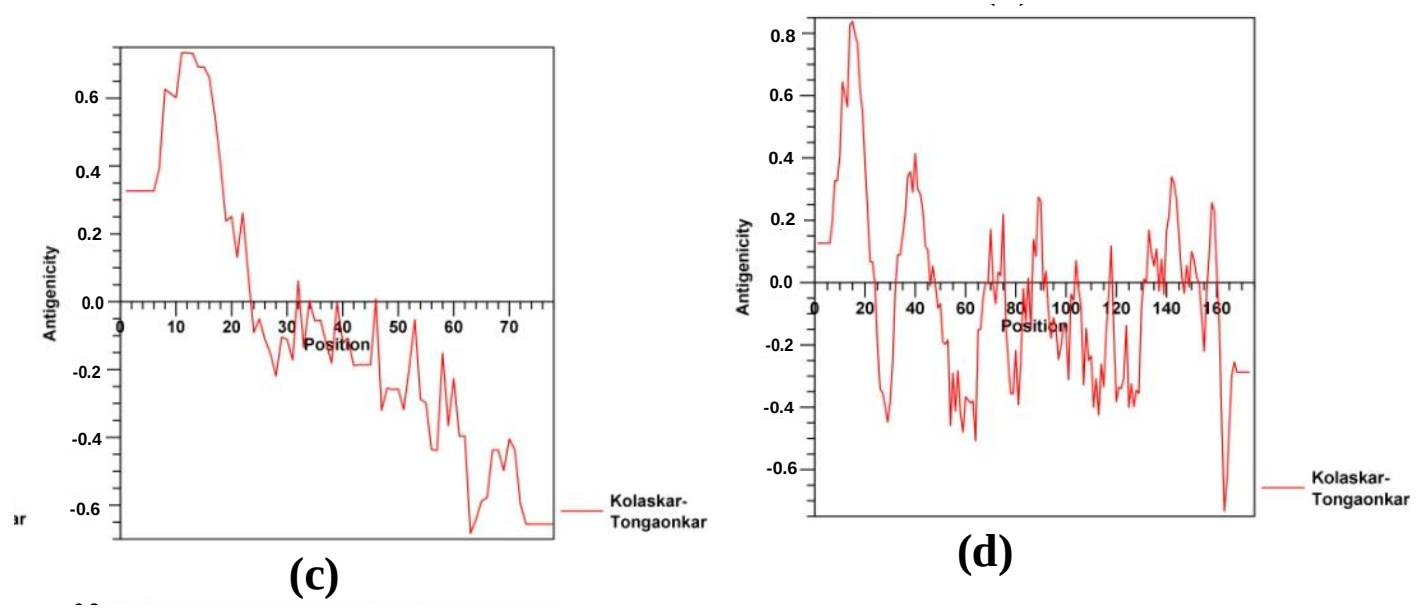

Figure 1. Contd.

identify those regions that are more antigenic as compared to others. Average antigenic propensity and antigenic determinants were also identified which indicates probable epitopes of common proteins. The common proteins among the species, *E. coli*, *Salmonella*, and *Shigella* sp., were antigenic. The antigenicity report therefore shows that they contain antigenic determinants (epitopes). Among these common proteins, only seven proteins were selected for the study depending upon their antigenic propensity and antigenic determinants. Table 6 shows the list of proteins with their antigenic propensity and determinants. Among them, FeoB, YaeC and OmpA were found to be more antigenic as compared to other eight proteins because they had 25, 13 and 12 antigenic determinants, respectively.

This study was an *in silico* approach to predict possible genes that can be used as vaccines. Docking procedures can be used for further verification of these targets, as it will reveal the active sites of target genes. Further work can be done on antigenic profiling to investigate antigenic determinants (epitopes) with greater accuracy. Moreover, *in vivo* experiments should be carried out in various animal models in order to confirm that the genes under consideration are possible vaccine targets or not.

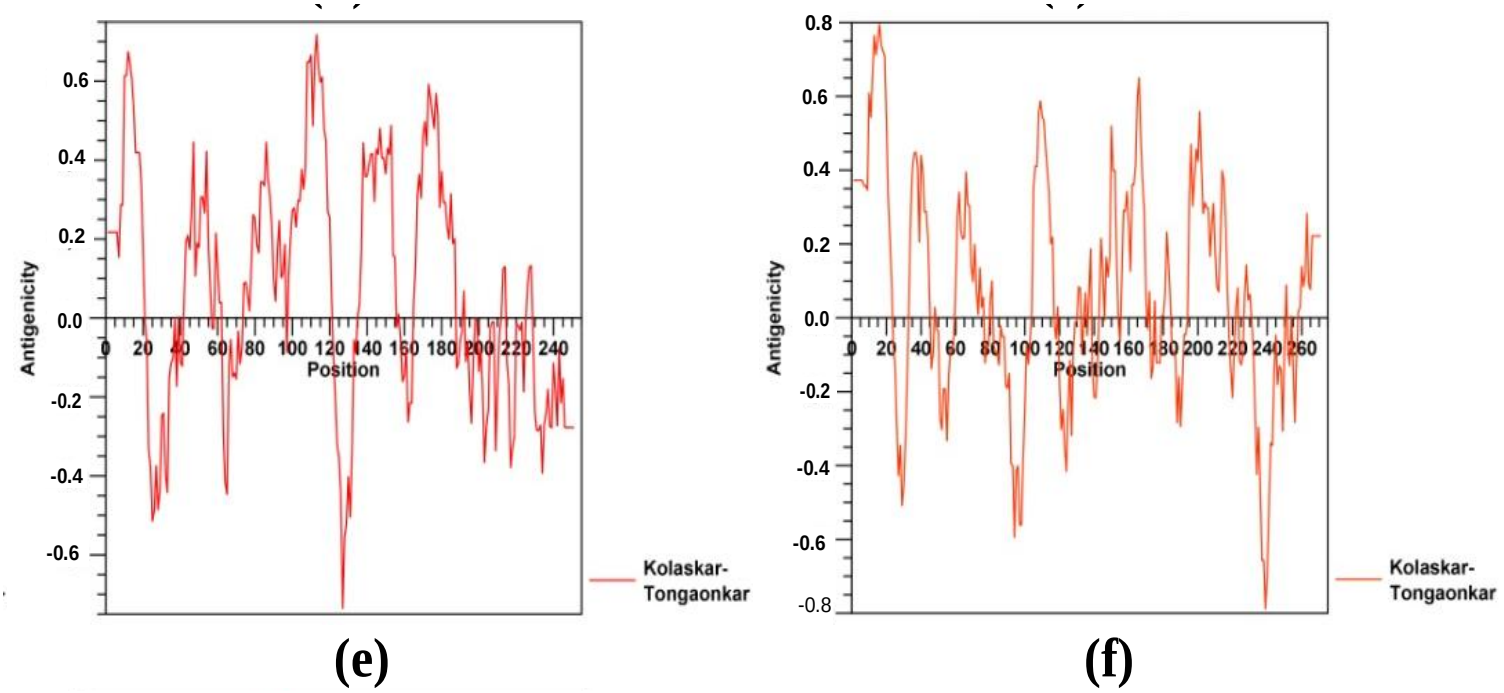

Figure 1. Contd.

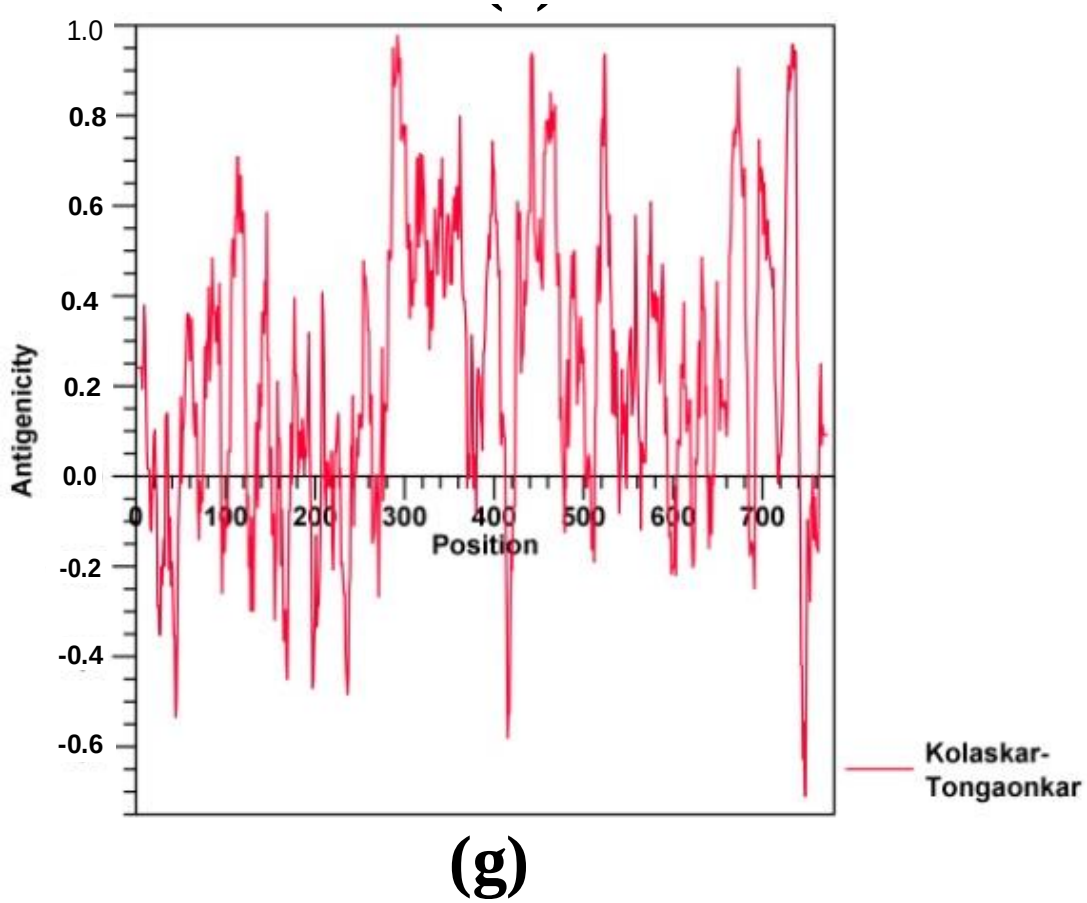

Figure 1. Contd.

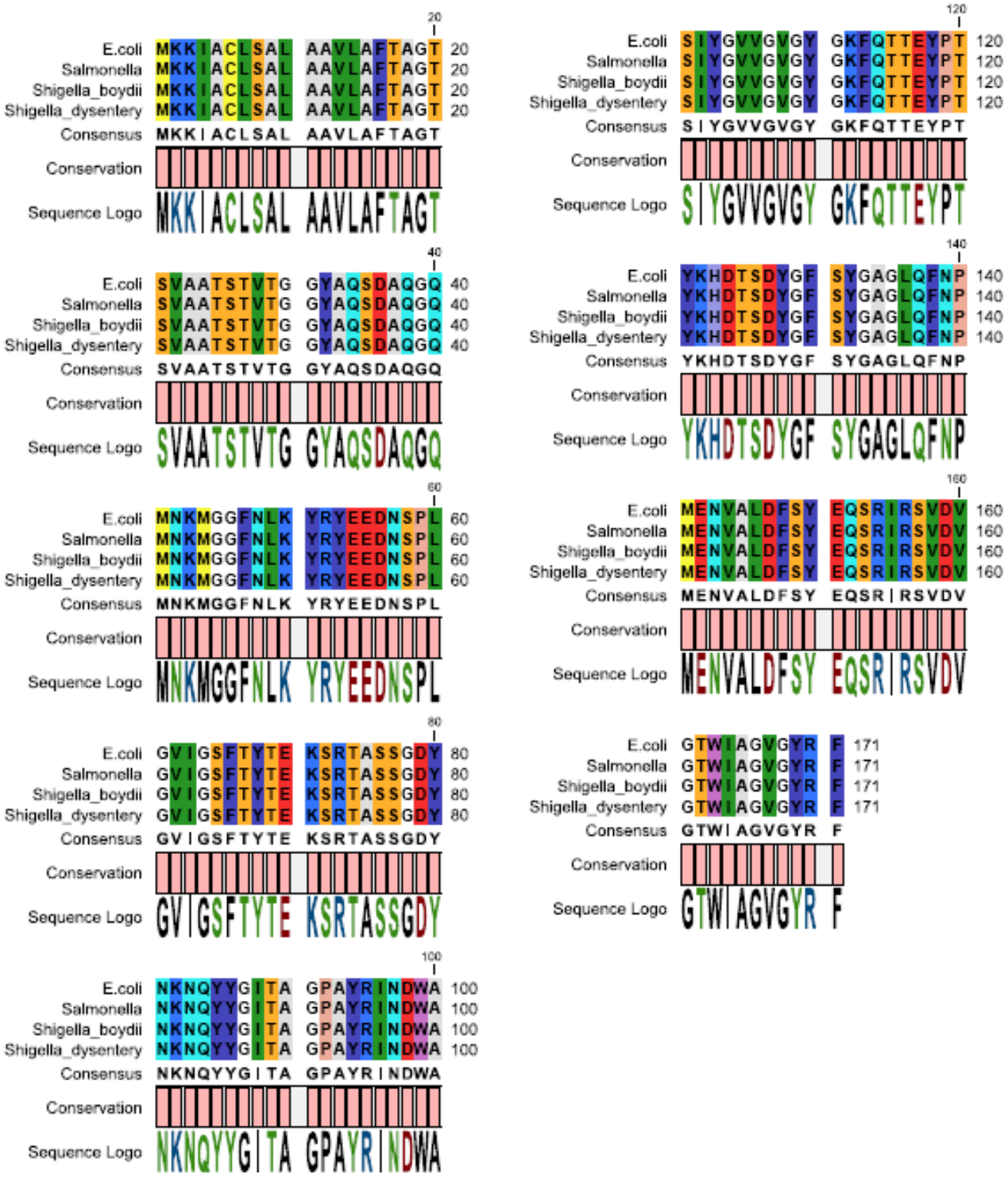

Figure 2. Multiple sequence alignment of common proteins causing diarrheal diseases.

Conclusion

We demonstrated that our *in silico* approach predicted vaccine candidates for diarrheal pathogens. Although false positives will still be present among these predicted targets, the method offers an interesting approach for further elucidation of genetic networks involved in the expression of all five species virulent genes. We believe that the identified genes retained in the respective genomes through selective advantage may play a key

role in pathogenesis. We have shown that following the simple strategy applied to bacterial enteric pathogens in this study may be helpful for the development of new vaccines, by exploiting surface proteins as putative vaccine candidates. The selected seven common proteins were also investigated for their potential as biological markers for identification of closely related common diarrhea-causing bacteria species. We therefore hope that future vaccines will be multivalent and will induce specific immunity against the protective antigens. It will also protect against both disease and possibly infection of newborns regardless of passively acquired antibodies, and will have a long duration of immunity, not requiring booster doses and will also be free of adverse reactions.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J. Mol. Biol. *215*: 403-410.

Bailey TL, Mikael B, Tom W, Philip M (2010). The value of position-specific priors in motif discovery using MEME. BMC Bioinformatics. 11(1):179. 2010.

Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002). Extensive feature detection of N – terminal protein sorting signals. Bioinformatics. 18(2): 298 -305.

Flower DR (2003). Towards in silico prediction of immunogenic epitopes. Trend Immunol. 24: 667-674.

Gromiha MM, Ahmad S, Suwa M (2004). Neural Network Based Prediction of Transmembrane Beta Strand Segments in Outer Membrane Proteins. J. Comp. Chem. 25: 762-767**.**

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

Horton P, Nakai K (1997). Better prediction of protein cellular localization sites with the k nearest neighbour classifier. Proc. Int. Conf. Intell. Syst. Mol. Biol. 5: 147-152.

Krogh A, Larsson B, Heijne GV, Sonnhammer ELL (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. 305(3): 567-580.

Mayers C, Duffield M, Rowe S, Miller J, Lingard B, Hayward S, Titball RW (2003). Analysis of known bacterial protein vaccine antigens reveals biased physical properties and amino acid composition. Comp. Func. Genom. 4: 468-478.

Nakai K, Kanehisa M (1991). Expert system for predicting protein localization sites in gram negative bacteria. Proteins. 11(2): 95 -110.

Rappuoli R (2000). Pushing the limits of cellular microbiology: microarrays to study bacteria-host cell intimate contacts. Proc. Natl. Acad. Sci. 97: 13467-13469.

Suzuki T, Murai T, Fukuda I, Tobe T, Yoshikawa M, Sasakawa C (1994). Identification and characterization of a chromosomal virulence gene, vacJ, required for intercellular spreading of Shigella flexneri. Mol Microbiol. 11(1): 31-41.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22): 4673-4680

World Health Organization, 'Safe Water, Better Health', WHO, Geneva, (2008). [http://www.who.int/quantifying_ehimpacts/publications/saferwater/en/index.html].

Zagursky RJ, Russell D (2003). Bioinformatics: Use in Bacterial Vaccine Discovery. In Bio Computing: Computer Tools for Biologists. Edited by Brown SM. Eaton Pub. 261-285.