



ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

ANTIGENIC SITE DETECTION OF THE SPIKE PROTEIN AND DESIGNING OF A VACCINE CANDIDATE AGAINST SARS CORONAVIRUS USING REVERSE VACCINOLOGY: AN *IN SILICO* APPROACH

***Pompi Sharma, R.L.Bezbaruah**

Biotechnology Division, CSIR-North East Institute of Science and Technology, Jorhat-785006, Assam.

Email: pompi.sharma86@gmail.com

Received on 12-10-2011

Accepted on 27-10-2011

Abstract

“Severe acute respiratory syndrome (SARS)” is the recent pandemic caused by a noble corona virus, SARS-CoV. SARS-CoV is a member of the Coronaviridae family which is enveloped; positive stranded RNA virus with the largest single-stranded RNA genome (approximately 27-31 Kb in length) among the known RNA viruses. Studies have shown that infection of the SARS-CoV is initiated by binding with RBD in the viral ‘S’ or the ‘Spike’ protein. Hence vaccination with S-protein would be ideal for disease prevention. The second major feature of coronavirus S-protein is its capacity to induce neutralizing antibodies and protective immunity, and it is thereby considered as a potential target for vaccine development. In this present study, an effort was taken to design a candidate vaccine applying the reverse vaccinology approach. The method includes identification of potential candidate vaccine, epitope prediction, peptide designing, and energy evaluation of the candidate vaccine followed by validation study. The high scoring vaccine candidate was selected. The result revealed that the designed candidate vaccine has a high binding affinity with T-cell receptor. Energy evaluation study showed that the vaccine has energy of -4820127112.693kcal/mol and most of the region of vaccine lies under the favourable portion of protein and none of the residues lies in the unfavourable regions. The peptide of the Spike protein can act as a very stable and a potential vaccine candidate against SARS.

Keywords: Reverse vaccinology, Spike protein, T-cell epitope, vaccine design.

Introduction

The SARS virus is a Corona virus which is a large, enveloped, RNA virus with club-like projections. The infection causes in the epithelium of the respiratory and digestive tracts. In the digestive tract it reduces absorptive capacity and causes diarrhoea, acute gastroenteritis, dehydration, and electrolyte imbalances. And in the respiratory tract it causes cold infections, acute respiratory distress, sore throat, malaise, fever, headache and cough, mild pneumonia.

[1]. The genome of the virus encodes basically four structural proteins namely: spike (S), membrane (M), envelope (E), and nucleocapsid (N), as well as some other non-structural proteins. Among all, the S protein plays an essential role in receptor-binding, virus entry and membrane fusion. It contains 1255 amino acids and 23 potential N-linked glycosylation sites.

The amino terminus contains a short sequence composed of hydrophobic amino acids. The carboxyl terminus consists of a transmembrane domain and a cytoplasmic tail rich in cysteine residues. The majority of protein (residues 12–1195) is outside the virus particle, which can be divided into amino-terminal S1 and carboxyl-terminal S2 domain. The S1 domain residues from 13-667 that binds to the host cell receptor, angiotensin-converting enzyme 2 (ACE2), while the S2 domain residues from 668-1255 are responsible for membrane fusion between virus and target cells. Monoclonal antibodies against S1 domain can block the receptor binding and contain potent neutralization activity against SARS-CoV. However, S2 domain can also inhibit SARS-CoV infection [2].

Studies shows that the SARS-CoV found in animals does not cause SARS like disease in the natural host neither transmits from animals to human but it has been seen that under certain conditions, the virus may have evolved into the early human SARS-CoV, gaining the ability to be transmitted from animals to humans or even from humans to humans. The SARS-CoV infection starts when the RBD on the viral S protein (S1 subunit) binds to the ACE2 on target cells forming the fusogenic core between the HR1 and HR2 regions in the S2 domain which brings the viral and target cell membranes into close proximity resulting virus fusion and entry. A salient feature is the prominent immune response to the virus after infection. In previous studies of

coronavirus, the S protein was identified as a potent immunogen, which induced neutralizing antibodies and elicited cytotoxic T-cell responses [3][4].

Materials and Methods

The objective of this study was to identify immunogenic S-protein peptides that can serve as cytotoxic T-cell epitopes in SARS vaccines.

Reverse Vaccinology Approach

The “reverse vaccinology” approach is not based on growing microorganisms in the wet laboratory but rather working in computer using algorithms to extract information from the microorganisms. It starts by searching for immunogenic antigens through *in silico* analyses of the pathogen's genome. This allows saving time and money while facing pathogens for which cell culture is difficult or impossible. The concept of reverse vaccinology was applied for the first time in the attempt to develop a vaccine against serogroup B Neisseria meningitides (MenB), the major cause of sepsis and meningitis in children and young adults [4]. This approach permits the researchers to select any protein encoded by the genome of a pathogen.

It is found to be more effective with prokaryotes (extra- as well as intracellular). Reverse Vaccinology is considered as a turning stone in vaccinology. It shows how powerful and useful Bioinformatics can be in the post-genomic era [5].

Protein Sequence Analysis

The protein sequences for SARS Coronavirus Tor2 strain were obtained from the JCVI-CMR database. The Comprehensive (CMR) is a freely available website which is used to display publicly available information on complete prokaryotic genomes. It offers a wide variety of tools and resources that allows the researcher to access all of the bacterial genome sequences completed to date (<http://cmr.jcvi.org>).

The FASTA formats for the sequences were obtained from NCBI (www.ncbi.nlm.nih.gov).

They were then checked for similarity with *Homo sapiens*, coding DNA sequences (CDS) using TFASTY. TFASTY compares a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations. For further analysis, the sequences which showed no similarity with the human CDS were taken. Those sequences were then taken to find any similarity with the human proteome using

the BLASTP. Again the sequences that showed no similarity were further selected for the immunogenic analysis.

Antigenic Site Determination

Antigenic sites were determined by EMBOSS antigenic and Kolaskar and Tongaonkar antigenicity scale. Both finds out the potential antigenic regions of a protein sequence developed by Kolaskar and Tongaonkar based on a semi-empirical method making use of physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes. These methods have been applied to large number of proteins and it was found that these can predict antigenic determinants with about 75% accuracy which is better than most of the known methods (<http://emboss.bioinformatics.nl>) [6]. Kolaskar and Tongaonkar analysis further revealed that hydrophobic residues cystein, leucine and valine, if they occur on the surface of a protein are more likely to be a part of antigenic sites [7].

Immuno - Epitope Database Resource (IEDB Analysis)

The IEDB contains data related to antibody and T cell epitopes for humans, non-human primates, rodents, and other animal species. The database also contains MHC binding data from a variety of different antigenic sources and immune epitope data from the FIMM (Brusic), HLA Ligand (Hildebrand), TopBank (Sette), and MHC binding (Buus) databases. These databases and their investigators are hereby acknowledged as major contributors to the IEDB. In addition to the database, the IEDB website hosts an Analysis Resource, which contains a collection of tools to predict and analyze epitopes. Predictions done by IEDB with low percentile score shows good binders (www.immuneepitope.org) [8].

Solvent Accessible Regions

NetSurfP was used to predict the surface accessibility as well as the secondary structure of amino acids in an amino acid sequence. There are mainly two symbols “B” for “buried” and “E” for “exposed”. The exposed regions of the protein obtained from solvent accessibility analysis are the most probable antigenic sites. If exposed, would be able to interact with the environment [9][10].

Designing and Optimisation of the vaccine

The candidate vaccine was designed using ArgusLab4.0.1 [11]. And then optimisation was done using ChemBio3D Ultra 12.0 software [12].

Results

Selection of the candidate vaccine

Because the S protein of SARS-CoV is involved in receptor recognition as well as virus attachment and entry, it represents one of the most important targets for the development of SARS vaccines and therapeutics [18][19]. Out of the 14 sequences for SARS-Cov Tor2 strain obtained from Comprehensive Microbial Resource, Spike protein (orf2) showed no similarity with the known *Homo sapiens* CDS in TFASTY. TFASTY searches a nucleic acid database, translated in all six frames, using a protein query sequence. Further in BLAST, the protein showed no significant similarity or less than 35% (if showed) with the proteome of *Homo sapiens*. Hence the protein was selected as the potential vaccine candidate against SARS CoV Tor2 strain.

Antigenic peptide prediction

Results from Emboss: Antigenic gives the idea of the antigenic score. Within the Spike protein sequence (orf2), 7th (GLTVLPPLTD), 14th (EPVLKGVKL) and 23rd (MGCVLAW) number antigenic sites were found to be having the highest SAR (in percent) scores (Table No.1). Further confirmation for the selection of antigenic sites was done with the Kolaskar and Tongaonkar prediction method. Table No. 2(i, ii, iii).

Table No-1: Antigenic sites obtained from EMBOSS-Antigenic and its percent SAR scores.

Positions	Antigenic sites (EMBOSS: Antigenic)	SAR(Solvent Accessible Region) Scores in %
7	GLTVLPPLTD	0.45
14	EPVLKGVKL	0.44
23	MGCVLAW	0.42

Table No. 2: Antibody-Epitope Prediction (Kolaskar and Tongaonkar antigenicity)

Table No. 2.i. 7th antigenic site of orf2

Position	Residue	Peptide start position	Peptide End Position	Peptide	Score
4	V	1	7	GLTVLPP	1.113
5	L	2	8	LTVLPPL	1.167 (maximum)
8	L	5	11	LPPLLTD	1.093 (minimum)

Table No. 2.ii: 14th antigenic site of orf2.

Serial No	Residue	Peptide Start Position	Peptide end Position	Peptide	Score
4	L	1	7	EPVLKGV	1.105 (minimum)
5	K	2	8	PVLKGVK	1.116
6	G	3	9	VLKGVKL	1.143 (maximum)

Table No.2.iii: 23rd antigenic site of orf2.

Sl. No	Residue	Peptide Start Position	Peptide end Position	Peptide	Score
4	V	1	7	MGCVLAW	1.100 (average) (minimum) (maximum)

Solvent Accessible Region:

Solvent accessibility prediction of the spike protein was done using NetSurfP. Tables No. 3(i ,ii, iii).

Table No. 3: Solvent Analysis Results using NetSurfP.

Table No. 3.i: Solvent Analysis Results for the 7th antigenic site of orf2 (all amino acids are exposed).

Sl.No	Amino Acid	Class Assignment	Relative Surface Accessibility (RSA)	Absolute Surface Accessibility
1	G	E	0.920	72.428
2	L	E	0.392	71.812
3	T	E	0.525	72.859
4	V	E	0.360	55.286
5	L	E	0.290	53.099
6	P	E	0.327	46.415
7	P	E	0.421	59.797
8	L	E	0.309	56.560
9	L	E	0.327	59.855
10	T	E	0.581	80.626
11	D	E	0.921	132.731

Table No.3.ii: Solvent Analysis Results of the 14th antigenic site of orf2 (all amino acids are exposed)

Sl. No	Amino Acid	Class Assignment	Relative Surface Accessibility (RSA)	Absolute Surface Accessibility
1	E	E	0.872	152.269
2.	P	E	0.528	74.923
3	V	E	0.496	76.266
4	L	E	0.864	43.029
5.	K	E	0.589	121.096
6	G	E	0.538	42.364
7	V	E	0.421	64.723
8	K	E	0.652	134.034
9	L	E	0.682	124.819

Table No.3.iii: Solvent Analysis Results for the 23rd antigenic site of orf2(all amino acids are exposed except 'C' at the third position).

Sl.No	Amino Acid	Class Assignment	Relative Surface Accessibility(RSA)	Absolute Surface Accessibility
1	M	E	0.686	137.329
2.	G	E	0.444	34.959
3	C	B	0.244	34.187
4	V	E	0.381	58.498
5	L	E	0.309	56.541
6	A	E	0.403	44.367
7	W	E	0.597	143.506

IEDB Analysis

IEDB analysis results shows that the 7th antigenic site has lowest percentile of 0.80 as compared to the 14th antigenic site which showed a percentile score of 60.20. Table No. 4(i, ii).

Table No. 4: IEDB analysis results for checking the binding affinity towards the T-cell receptor

Table No. 4.i: IEDB analysis (MHC-I binding prediction results-Prediction Results) for the 7th antigenic site of the orf2 (sequence: GLTVLPPLLTD) where lowest percentile score is 0.80.

Sequence	Method Used	Peptide length	Percentile Rank
VLPLLTD	SMM	8	0.80
GLTVLPPL	SMM	8	1.60
GLTVLPPLL	Consensus(ANN, SMM,CombLib_Sidney2008)	9	2.70
GLTVLPPLLTD	Consensus(ANN, SMM)	10	8.50
GLTVLPPLLTD	SMM	11	11.60
LTVLPPLLTD	Consensus(ANN, SMM,CombLib_Sidney2008)	9	30.10
LTVLPPLLTD	Consensus(ANN, SMM)	10	31.20
TVPPLLTD	Consensus(ANN, SMM,CombLib_Sidney2008)	9	40.20
TVPPLLTD	SMM	8	45.40
LTVLPPLL	SMM	8	50.80

Table No. 4.ii: IEDB analysis results (MHC-I binding prediction results-Prediction Results for 14th antigenic site of the orf2 (Sequence: EPVLKGVKL) where lowest percentile score is 60.20 which is very high compared to the 7th antigenic site.

Sequence	Method Used	Peptide length	Percentile Rank
PVLKGVKL	SMM	8	60.20
EPVLKGVKL	Consensus(ANN, SMM,CombLib_Sidney2008)	9	86.60
EPVLKGVK	SMM	8	95.30

Designing and Energy Optimisation

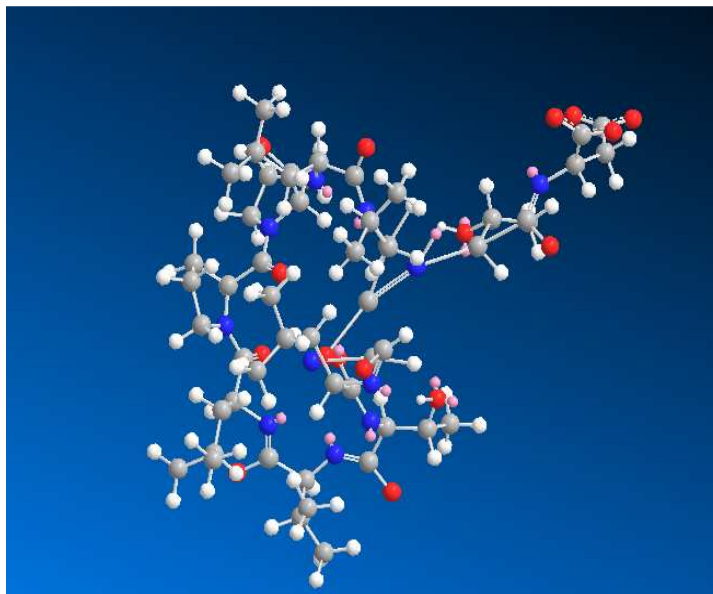
The peptides, i.e., 7th (GLTVLPPLLD), 14th (EPVLKGVKL) and 23rd (MGCVLAW) of the Spike protein (orf2) were built using ArgusLab4.0.1. After building the peptides, their geometry was cleaned or optimised with steepest decent as the line search method. To make the structures stable, energy minimisation was done using ChemBio3D Ultra 12.0. MM2 Force field was applied to minimise the energy. Table No. 5, Figure No. 1.

Table No. 5: Energy Minimisation using ChemBio3D Ultra 12.0 (7th antigenic site had lowest among the three).

Sl. No.	Antigenic site	Result	
1	GLTVLPPLLD (7 th antigenic site)	Stretch	2829.7415
		Bend	1174.8965
		Stretch Bend	80.5984
		Torsion	63.4886
		Non-1,4 VDW	2038.1091
		1,4 VDW	124.0642
		Charge/charge	1129.0351
		Charge/Dipole	-4820134552.627
		Total Energy	-4820127112.693 kcal/mol
2	EPVLKGVKL (14 th antigenic site)	Stretch	23.1426
		Bend	56.9218
		Stretch Bend	4.8625

		Torsion	12.9576
		Non-1,4 VDW	1.3146
		1,4 VDW	38.2010
		Charge/charge	1219.2261
		Charge/dipole	-136.5343
		Dipole/dipole	-0.1026
		Total Energy	1219.9892 kcal/mol
3	MGCVLAW (23 rd antigenic site)	Stretch	17.7723
		Bend	58.8043
		Stretch Bend	2.7805
		Torsion	-0.7507
		Non-1,4 VDW	52.9009
		1,4 VDW	30.6068
		Charge/charge	675.5572
		Charge/dipole	-225.0901
		Total Energy	612.5811 kcal/mol

Figure No. 1: Designed peptide (7th antigenic site) after energy minimisation using ChemBio3D Ultra 12.0 .



Ramachandran Plot

As the 7th antigenic site showed lowest percentile score and is considered to have a good binding affinity with the T-cell receptor as well as after energy minimisation, this site had the lowest energy of -4820127112.693kcal/mol; it is considered for further validation by the Ramachandran Plot analysis. Figure 2 and Figure 3.

Figure 2: Ramachandran Plot Statistics.

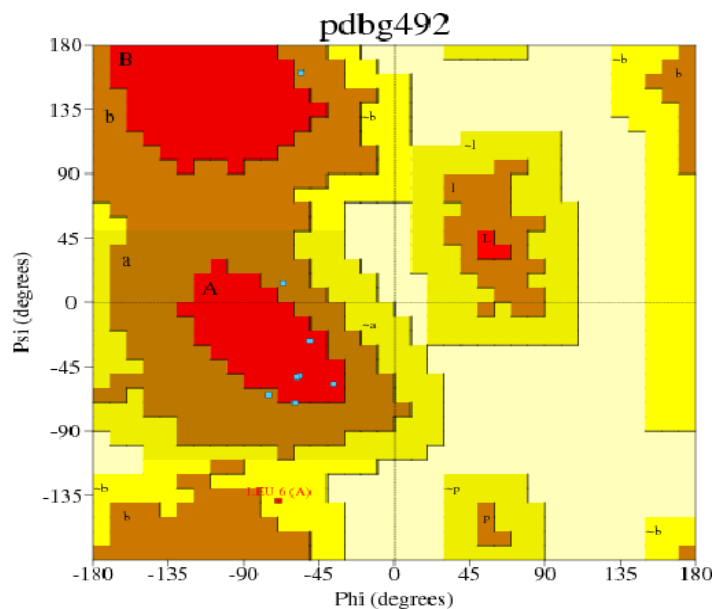
	No. of residues	%-tage
Most favoured regions [A, B, L]	4	57.1%
Additional allowed regions [a, b, l, p]	2	28.6%
Generously allowed regions [~a, ~b, ~l, ~p]	1	14.3%
Disallowed regions [XX]	0	0.0%

Non-glycine and non-proline residues	7	100.0%

End-residues (excl. Gly and Pro)	1	

Glycine residues	1	
Proline residues	2	

Figure 3: Ramachandran Plot.



Discussions

Antigenic site determination

Analysis with the Kolaskar and Tongaonkar analysis revealed that 7th antigenic site had the highest antigenic score of 1.167 on its 5th position which is the position of 'Leucine'. On the 6th position of the 14th antigenic site, 'Leucine' residue had the minimum score of 1.105. Though lysine and glycine were found to have a high score as compared to leucine, they are not considered in the study as it did not show the hydrophobicity and also they were not the part of the antigenic sites. The 23rd antigenic site gave a high score of 1.100 at the 'V' residue of its 4th position

Solvent accessibility using NetSurfP

As seen in the Table No. 3 (i,ii,iii), it has been noticed that every amino acid residue in the 7th and the 14th antigenic site were found to be 'exposed' or 'E' indicating that they have the immense ability to interact with the environment. In the 23rd antigenic site only one 'buried' or 'B' was found at the third position which also indicated that it had also the ability of interaction with the environment.

IEDB analysis

The results in the Table No. 4(i, ii) indicate that the 7th antigenic site which has a very low percentile score of 0.80 has got the highest binding affinity with the T-cell receptor as compared to the 14th sequence whose has the lowest percentile score of 60.20.

Energy minimisation

After minimisation (Table No. 5), the energy of the 7th antigenic site was found to be -4820127112.693kcal/mol, for the 14th antigenic site, energy was 1219.9892kcal/mol and that for 23rd antigenic site it was found to be 612.581kcal/mol. From all the results it could be inferred that the 7th antigenic site was having the lowest energy (Figure No. 1) as compared to the other two peptides and the lowest energy indicates a better and a stable structure.

Ramachandran Plot Analysis: The analysis (Figure No. 2) as well as the figure (Figure No.3) of the 7th antigenic site indicates that none of the amino acids were present in the disallowed region. Majority of the residues (57.1%) were present in the most favoured region.

Conclusion

Through this work, a small attempt has been done to design an *in silico* potential vaccine candidates, found through the screening of the proteome of SARS coronavirus TOR2 strain. Severe acute respiratory syndrome (SARS) is a newly emerged infectious disease caused by SARS-associated coronavirus (SARS-CoV). SARS-CoV, S or Spike protein has vital roles in viral infection and pathogenesis. The functions of the S protein are, first of all, to bind to species-specific host cell receptors. Binding of the S protein to receptor triggers a fusion event between the viral envelope and a cellular membrane, in some cases the plasma membrane, in other cases the endosomal membrane, and this results in internalization of the viral nucleocapsid into the cytoplasm. In

some cases, but not all, the S protein also traverses all the way to the plasma membrane of infected cells, and so it can induce the fusion of adjacent cells to form syncytia. Finally the S protein is the principal viral antigen eliciting neutralizing antibody on the part of the host. The identification of effective vaccine depends on the identification of the composition of protective antigens. Hence, in this study, major efforts were directed towards identifying and testing protein antigen. After finding out the main antigenic sites having highest scores in all analysis, further study were done to find its T-cell binding affinity where it was found that 7th antigenic site had the highest binding affinity among other antigenic sites, 14th and 23rd. After designing them in ArgusLab, their energies were minimised when it was found that the 7th antigenic site had the lowest energy and was the most stable among the three. For further validation of the 7th antigenic site, Ramachandran Plot analysis was done where it was found that none of its residues were present in the disallowed region and majority of the residues were in the most favoured region.

Acknowledgement

The authors would like to thank Department of Biotechnology, Government of India, for providing the financial assistance at the BIF Centre and Biotech Hub. Heart full thanks also goes to Dr. P.G.Rao, Director, CSIR-NEIST, Jorhat, for being an all time inspiration to carry out the research work.

References

1. Understanding the SARS Genome/Proteome Using DS Gene, DS GeneAtlas, and DS AtlasStore Publication by Lisa Yan, Mikhail Velikanov, Paul Flook, Wenjin Zheng, Sándor Szalma, and Scott Kahn, "Assessment of Putative Protein Targets Derived from the SARS Genome," FEBS Letters 2003, 554, 257-263.
2. Peptide Mimicing between SARS Coronavirus Spike Protein and Human Proteins Reacts with SARS Patient Serum K.Y. Hwa,W.M. Lin, Y.-I. Hou and T.-M. Ye.
3. T-Cell Epitopes in Severe Acute Respiratory Syndrome (SARS) Coronavirus Spike Protein Elicit a Species T-Cell Immune Response in Patients Who Recover from SARS. Yue-Dan Wang,Wan-Yee Fion Sin, Guo-Bing Xu,Huang-Hua Yang,Tin-yau Wong,Xue-Wen Pang,Xiao-Yan He,Hua-Gang Zhang,Joice Na Lee Ng, Chak-Sum Samuel Cheng,Jing Ju,Li Meng,Rui-Feng Yang,Sik-To Lai Zhi-Hong Guo,Yong Xie, and Wei-Feng Chen1.

4. SARS Vaccine Development Shibo Jiang, Yuxian He, and Shuwen Liu Balmer, P. and Miller, E. (2002)
Meningococcal disease: how to prevent and how to manage. *Curr. Opin. Infect. Dis.* 15, 275–281.
5. NERVE: New Enhanced Reverse Vaccinology Environment Sandro Vivona, Filippo Bernante and Francesco Filippini.
6. A semi-empirical method for prediction of antigenic determinants on protein antigens A. S. Kolaskar and Prasad C. Tongaonkar.
7. Kolaskar AS, Tongaonkar PC. *FEBS Letters* 1990, 276:172-174
8. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B. The immune epitope database 2.0. *Nucleic Acids Res.* 2010 Jan; 38(Database issue):D854-62. Epub 2009 Nov 11.
9. Peterson B, Peterson TN, Anderson P et al. *BMC Struct Biol.* 2009.
10. Antigenic epitopes and MHC binders in OMP A of fish pathogen by (A bioinformatics study) Mohammed Neema, Iddya Karunasagar, Indrani Karunasagar, UNESCO-MIRCEN Centre of Fisheries, Karnataka Veterinary, Animal and Fisheries Sciences University, Mangalore, 575002, Karnataka.
11. Thompson, M.A. 2004 ArgusLab 4.0.1 Seattle, WA Planaria Software LLC.
12. ChemBioOffice Ultra 2010-A Great Benefit for Academia/CambridgeSoft Solution featuring E-Notebook (www.cambridgesoft.com).
13. Shigeta S, Yamase T, Current status of anti-SARS agents. *Antivir Chemother* 2005; 16; 23-32.
14. Lit, et al. Long-term persistence of robust antibody and cytotoxic T cell responses in recovered patients infected with SARS coronavirus. *PLoS ONE.* 2006; 1:e24.

Corresponding Author:

Pompi Sharma*,

Email: pompi.sharma86@gmail.com