T cell epitope-based vaccine design for pandemic novel coronavirus 2019-nCoV across structural and non-structural proteins

Seema Mishra

Department of Biochemistry

School of Life Sciences

University of Hyderabad, India

Email: seema_uoh@yahoo.com

Abstract:

Novel coronavirus 2019 has emerged a pandemic ever since its outbreak in China and is a potent life-threatening disease to mankind across continents. A member of SARS-coronavirus family, its treatment and prevention regimen is till date, non-existent. In unprecedented efforts towards its eradication, various lines of treatment are under way including designing drugs and antibodies. Spreading primarily through human contact, even asymptomatic transfer has been found to occur. Therefore, a quick response to developing an effective immunotherapy regimen is sorely needed in order to prevent further infections. In this study, immunoinformatics approaches have been used to provide putative promiscuous epitopes using genome-wide screening of novel coronavirus genome. Theoretically speaking, the ideal scenario would be to use all the protein targets available to identify potent immunogens as data is scarce on the identity of virulent proteins of the nCoV genome. In this regard, a ranked list of immunogenic epitopes across all of these ten protein targets at various stages of viral life

cycle was obtained. This list includes top ranked helper and cytotoxic T cell epitopes common across MHC alleles, covering all predominant HLA supertypes in population. An interesting observation from this study is that surface (spike) and membrane proteins of nCoV provide with very less number of promiscuous epitopes with high degree of unique epitopes across alleles. This shows that these proteins may be less immunogenic and the vaccination strategy using these proteins may not work at entire population level across continents. Almost all other proteins studied were predicted to harbor a high number of promiscuous epitopes and may prove to be better immunogens. Further, it was necessary to find out globally conserved nonamer epitopes in nCoV genome, in order to help generate a robust immune response. The prominent consensus sequences harboring these nonamer epitopes as clusters were: MGYINVFAFPFTIYSLLLC and KVSIWNLDYIINLI across two proteins and alleles. All of the 57 identified coronaviral epitopes were not present in human proteome. The results from this study are provided to scientific community and may be further utilized to aid in cost- and time-effective vaccine development starting from MHC-binding and T-cell stimulation assays.

Introduction

Novel coronavirus (nCoV) first emerged in population in December 2019 and has rapidly gained foothold across the world resulting in WHO declaring it as pandemic (https://www.who.int/emergencies/diseases/novel-coronavirus-2019). As there is currently no known cure, urgent studies are needed in order to push forward vaccine design and development. Recently, about 77 drugs were identified by world's fastest supercomputer, Summit, against viral spike protein (Smith & Smith, 2020). Immunoinformatics tools have been proved crucial time and again in relation to cancer immunotherapy (Seema Mishra & Subrata Sinha, 2006; 2009). In the absence of effective drugs to date, vaccination is indispensable in order to cure an entire population. More important is the

fact that since this Covid19 disease has affected almost all of the world's population, promiscuous epitopes binding to a variety of HLA alleles for wider dissemination is crucial. In this regard, *in silico* approaches will be significantly useful in helping develop a cure in as fast a manner as possible. Cytotoxic T cell immune responses have been observed in its close relative, SARS and MERS (Oh *et al.*, 2012; Shi *et al*., 2015), and hence, in nCoV case also, cytotoxic T cell-coordinated immune response along with helper T cell response is crucial and needs to be implemented fast. Based on newly available nCoV genome sequence, this study has been embarked upon with the clear objective of providing a ranked list of highly probable and effective promiscuous epitopes with no human crossreactivity.

NCoV genome submitted by CDC, Atlanta (GenBank accession number: MT106054.1) is 29882 bp in length. It harbors multiple structural and non-structural proteins essential at various stages of a viral life cycle. In brief, the sequence of proteins in its RNA genome is as follows: 5′-leader-UTR-replicase-S (Spike)–E (Envelope)-M (Membrane)-ORF6-ORF7a-ORF8-N (Nucleocapsid)-3′UTR-polyA tail (From GenBank; Fehr and Perlman, 2015). While these proteins are key proteins, several proteins such as ORF3a, ORF7a, ORF8 function as accessory proteins playing a role in viral pathogenesis.

While cytotoxic T cell response is the key response to immunodominant antigens in destroying a virus infected cell, helper T cells prime and maintain cytotoxic T cells and so, an effective immunotherapeutic product must contain both kinds of T cell epitopes. These T cell epitopes need to be both high binders to respective HLA alleles as well as be immunogenic. Further analyses using clustering provided us with consensus epitopes eliminating redundant sequences across target proteins and alleles. These epitopes could elicit stronger cellular immune responses to viral proteins. As opposed to common perception that membrane and spike proteins confer better immunogenic ability, an interesting perception is found from this study that it is the opposite case in context of nCoV.

Materials and Methods:

Genome sequence:

The genome sequence of novel coronavirus was retrieved from GenBank accession number MT106054.1 and the corresponding proteins were retrieved. RefSeq sequences of all of the proteins present in this genomic sequence, ORF10 protein (YP_009725255.1), nucleocapsid phosphoprotein (YP_009724397.2), orf8 protein (GenBank: QID21074.1, no RefSeq sequence identified for ORF8), ORF7a protein (YP_009724395.1), ORF6 protein (YP_009724394.1), membrane glycoprotein (YP_009724393.1), envelope protein (YP_009724392.1), ORF3a protein (YP_009724391.1), surface glycoprotein (YP_009724390.1), ORF1ab (YP_009724389.1) were analysed in order to cover the entire genome of nCoV in view of absence of data on its virulent proteins. Fasta sequences of all of these proteins were taken as inputs in several T cell epitope prediction and analysis tools.

Cytotoxic T cell epitopes prediction:

NetCTLpan version 1.1 (http://www.cbs.dtu.dk/services/NetCTLpan/) and PickPocket version 1.1 (http://www.cbs.dtu.dk/services/PickPocket/) were both used to predict and generate a consensus list of top high binders and promiscuous epitopes across several proteins. While NetCTLpan uses neural network algorithm, PickPocket works on the basis of position-specific weight matrices. NetCTLpan, in addition to HLA binding, also predicts TAP-transporter binding and C-terminal proteasome cleavage predictions.The consensus list was chosen to increase prediction accuracy from two different algorithms. Both these tools use representative HLA supertypes and in all, 12 supertypes were present

by default and hence taken. All the parameters used were default parameters. Nonameric peptide epitopes were selected.

Helper T cell epitope prediction:

NetMHCIIpan version 3.2 (http://www.cbs.dtu.dk/services/NetMHCIIpan/) was used to predict helper T cell epitopes across several HLA-DRB1, -DRB3, -DRB4, -DRB5 and HLA-DP as well as HLA-DQ alleles. It works on the basis of quantitative MHC-peptide binding affinity data obtained from the Immune Epitope Database. A consensus list of 15 amino acids long ranked epitopes was generated. For generating top ranked epitopes, these were sorted using descending order of predicted binding affinity.

Immunogenicity prediction:

Immunogenicity is a characteristic property of peptide epitopes that can elicit an immune response. High binding affinity to HLA alleles is not a sufficient criterion for high immunogenicity. Therefore, all the epitopes that were generated as a consensus were checked for their immunogenicity. Immune Epitope database (IEDB) immunogenicity tool (http://tools.iedb.org/immunogenicity/) was used to generate a list of immunogenic epitopes and both the helper and cytotoxic T cell epitopes were scanned for the presence of immunogenic regions. Immunogenicity of a peptide-MHC complex is predicted based on the physicochemical properties of amino acids and their positions in the predicted peptide. Specifically, amino acids with large and aromatic side chains and positions 4-6 are more important to the immunogenicity of the peptide being presented.

Clustering

As globally conserved epitopes are relevant at this time to contain and treat coronavirus infection, clustering approach was used to find patterns among disparate datasets. In order to group epitopes into several clusters, IEDB epitope cluster analysis tool was applied. All the topmost epitopes across

proteins targets were used as inputs with minimum sequence identity threshold as 70%. Cluster-break algorithm was applied for clear representative sequence.

Cross-reactivity analysis:

All the 57 epitopes obtained were used to search against human proteome database from UniProt for any matches to human proteome, thus avoiding cross-reactivity. For this, Multiple Peptide Match tool (https://research.bioinformatics.udel.edu/peptidematch/batchpeptidematch.jsp) of Protein Information Resource was used.

Results & Discussion:

Cytotoxic T lymphocytes (CTL) epitope prediction was done using PickPocket 1.1 and NetCTLpan 1.1 using the same HLA supertypes. A common list of 9 amino acids long, high binders was generated among topmost epitopes in each case. All of the nCoV proteins, including ORF1ab (for ORF1ab, manuscript is under preparation), were used for this study. These common CTL epitopes are enlisted in Table 1 as ranked order. It is found that very few promiscuous epitopes could be seen in the case of surface and membrane proteins common to both the prediction algorithms. These proteins harbour many potential, unique epitopes across the two prediction tools, leading to the surmise that these two proteins will not be a potent immunogen. Nevertheless, a few common promiscuous epitopes across prediction algorithms, though not belonging to top-ranked ones were enlisted for these two proteins. The highest number of common top-ranking epitopes is seen in the case of ORF10 followed by ORF8, ORF6 and Envelope proteins. Immunogenicity prediction of these proteins (table 2) showed that many of these epitopes had a positive score. A clear correlation between HLA binding and immunogenicity is seen in these cases, lending support to the theory that these epitopes may mount a high immune response *in vitro*.

Table 1: Top ranked sequences of CTL epitopes common across HLA supertypes (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*24:02, HLA-A*26:01, HLA-B*07:02, HLA-B*08:01, HLA-B*27:05, HLA-B*39:01, HLA-B*40:01, HLA-B*58:01, HLA-B*15:01) and across the two prediction algorithms used. * not amongst top-ranked ones in NetCTLpan results

Epitope sequences								
Spike (surface)*	**ORF3a**	**Envelope**	**Membrane***	**ORF6**	**ORF7a**	**ORF8**	**Nucleocapsid**	**ORF10**
FVFLVLLPL	FTSDYYQLY	LTALRLCAY	FLFLTWICL	KVSIWNLDY	GTYEGNSPF	IQYIDIGNY	FAPSASAFF	YINVFAFPF
YLQPRTFLL	FLYLYALVY	LIVNSVLLF	YFIASFRLF	NLDYIINLI	FLIVAAIVF	GIIITVAAF	NTASWFTAL	MGYINVFAF
	FVTVYSHLL	FLAFVVFLL		HLVDFQVTI	ITLATCELY	SFYEDFLEY	LQLPQGTTL	VFAFPFTIY
	LLYDANYFL	IVNSVLLFL		MFHLVDFQV	ILFLALITL	SLVVRCSFY		IAQVDVVNF
	FVCNLLLLF	LVKPSFYVY		VTIAEILLI	CVRGTTVLL	QSCTQHQPY		AFPFTIYSL
	YYQLYSTQL	TLAILTALR		ILLIIMRTF	KIILFLALI	YVVDDPCPI		FPFTIYSLL
	WLIVGVALL			WNLDYIINL	ELYSPIFLI	TVSCSPFTI		NSRNYIAQV
	YLYALVYFL			TIAEILLII		EYHDVRVVL		RMNSRNYIA
				SIWNLDYII		HFYSKWYIR		NVFAFPFTI
								QVDVVNFNL
								FTIYSLLLC

Table 2: Immunogenic epitopes across proteins:

Spike (Surface)

Peptide	Length	Score
YLQPRTFLL	9	0.1305
FVFLVLLPL	9	0.04076

ORF3a

Peptide	Length	Score
WLIVGVALL	9	0.18314
YLYALVYFL	9	0.13151
LLYDANYFL	9	0.11841
FLYLYALVY	9	0.03563
FVCNLLLLF	9	-0.06109
FVTVYSHLL	9	-0.08437
FTSDYYQLY	9	-0.1427
YYQLYSTQL	9	-0.24301

Envelope

Peptide	Length	Score
FLAFVVFLL	9	0.30188
TLAILTALR	9	0.1989
LTALRLCAY	9	0.01886
IVNSVLLFL	9	-0.07977
LVKPSFYVY	9	-0.11106
LIVNSVLLF	9	-0.13119

Membrane

Peptide	Length	Score
FLFLTWICL	9	0.35397
YFIASFRLF	9	0.06887

ORF6

Peptide	Length	Score
TIAEILLII	9	0.30101
KVSIWNLDY	9	0.29343
VTIAEILLI	9	0.28951
WNLDYIINL	9	0.24894
NLDYIINLI	9	0.24642
ILLIIMRTF	9	0.16098

Peptide	Length	Score
SIWNLDYII	9	0.15011
HLVDFQVTI	9	0.0982
MFHLVDFQV	9	0.09154

ORF7a

Peptide	Length	Score
FLIVAAIVF	9	0.29611
ILFLALITL	9	0.1895
KIILFLALI	9	0.16214
CVRGTTVLL	9	0.1536
ITLATCELY	9	0.10084
ELYSPIFLI	9	0.03913
GTYEGNSPF	9	-0.01964

ORF8

Peptide	Length	Score
GIIITVAAF	9	0.30966
IQYIDIGNY	9	0.30442
SFYEDFLEY	9	0.28049
EYHDVRVVL	9	0.1807
YVVDDPCPI	9	-0.0051
SLVVRCSFY	9	-0.01663
HFYSKWYIR	9	-0.09452
TVSCSPFTI	9	-0.15801
QSCTQHQPY	9	-0.16503

Nucleocapsid

Peptide	Length	Score
NTASWFTAL	9	0.22775
LQLPQGTTL	9	-0.04022
FAPSASAFF	9	-0.18628

ORF10

Peptide	Length	Score
VFAFPFTIY	9	0.34042
NVFAFPFTI	9	0.30241
MGYINVFAF	9	0.28694
YINVFAFPF	9	0.28259
QVDVVNFNL	9	0.17787
AFPFTIYSL	9	0.1775
NSRNYIAQV	9	0.09731
IAQVDVVNF	9	0.09546
FPFTIYSLL	9	0.05708

Peptide	Length	Score
RMNSRNYIA	9	-0.04962
FTIYSLLLC	9	-0.1479

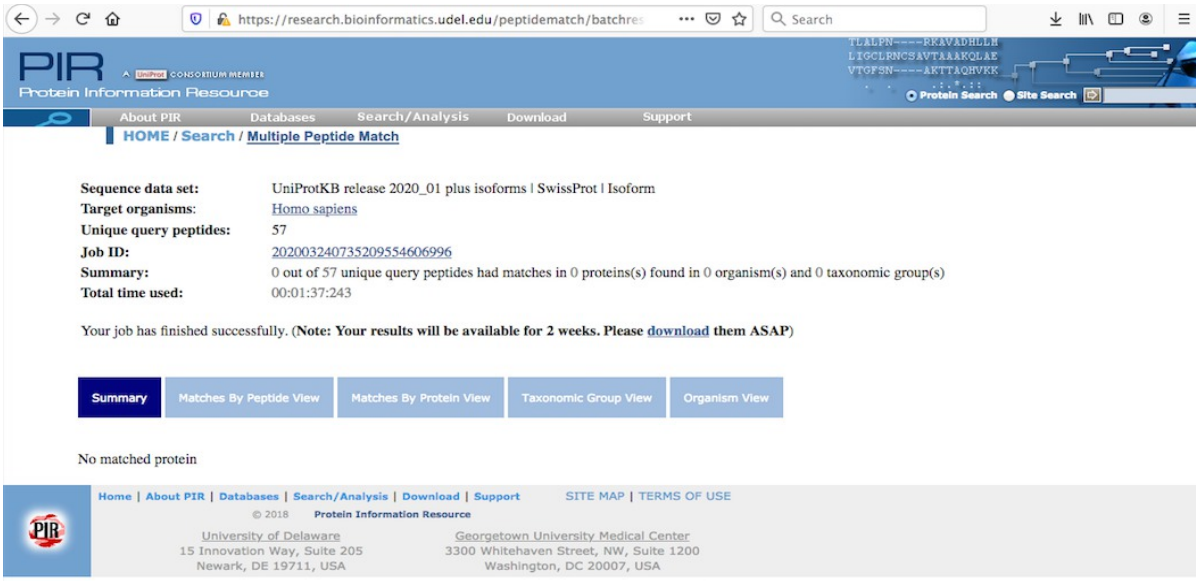
All of these CTL epitopes across the proteins studied were then clustered using IEDB epitope cluster analysis tool (Dhanda *et al.*, 2018) to make further biologically meaningful decisions. Results analyzed suggested that many epitopes were clustered around one consensus sequence, here the number of consensus sequences is two with more than two epitopes (Table 3). The prominent consensus sequences were: MGYINVFAFPFTIYSLLLC and KVSIWNLDYIINLI across two proteins and alleles and epitopes harboring these sequences may be considered immunodominant epitopes and tested first among the ranked list of epitopes. Several consensus sequences had only two peptide epitopes as a cluster. Many singletons (unique epitopes) were also found, lending credence that nCoV is indeed a dangerous pathogen to control, although for effective immunotherapy at a global scale, efforts should already be underway using these ranked list of epitopes.

Table 3: Consensus and singleton sequences generated using IEDB Clustering tool:

Peptide Number	Alignment	Position	Description	Peptide
Consensus	MGYINVFAFPFTIYSLLLC	-0	-0	-0
1	MGYINVFAF----------	1	seq25	MGYINVFAF
2	--YINVFAFPF--------	3	seq24	YINVFAFPF
3	----NVFAFPFTI------	5	seq32	NVFAFPFTI
4	-----VFAFPFTIY-----	6	seq26	VFAFPFTIY
5	-------AFPFTIYSL---	8	seq28	AFPFTIYSL
6	--------FPFTIYSLL--	9	seq29	FPFTIYSLL
7	----------FTIYSLLLC	11	seq34	FTIYSLLLC
Consensus	KVSIWNLDYIINLI	-0	-0	-0
1	KVSIWNLDY-----	1	seq1	KVSIWNLDY
2	--SIWNLDYII---	3	seq9	SIWNLDYII
3	----WNLDYIINL-	5	seq7	WNLDYIINL
4	-----NLDYIINLI	6	seq2	NLDYIINLI
Consensus	LIVNSVLLFL	-0	-0	-0
1	LIVNSVLLF-	1	seq39	LIVNSVLLF
2	-IVNSVLLFL	2	seq41	IVNSVLLFL
Consensus	VTIAEILLII	-0	-0	-0
1	VTIAEILLI-	1	seq5	VTIAEILLI
2	-TIAEILLII	2	seq8	TIAEILLII
Consensus	KIILFLALITL	-0	-0	-0
1	KIILFLALI--	1	seq56	KIILFLALI
2	--ILFLALITL	3	seq54	ILFLALITL
Consensus	RMNSRNYIAQV	-0	-0	-0
1	RMNSRNYIA--	1	seq31	RMNSRNYIA
2	--NSRNYIAQV	3	seq30	NSRNYIAQV
Consensus	MFHLVDFQVTI	-0	-0	-0
1	MFHLVDFQV--	1	seq4	MFHLVDFQV
2	--HLVDFQVTI	3	seq3	HLVDFQVTI
Consensus	FLYLYALVYFL	-0	-0	-0
1	FLYLYALVY--	1	seq44	FLYLYALVY
2	--YLYALVYFL	3	seq50	YLYALVYFL
Consensus	IAQVDVVNFNL	-0	-0	-0
1	IAQVDVVNF--	1	seq27	IAQVDVVNF
2	--QVDVVNFNL	3	seq33	QVDVVNFNL
Singleton	TLAILTALR	-0	seq42	TLAILTALR
Singleton	FLAFVVFLL	-0	seq40	FLAFVVFLL
Singleton	ELYSPIFLI	-0	seq57	ELYSPIFLI
Singleton	NTASWFTAL	-0	seq13	NTASWFTAL
Singleton	QSCTQHQPY	-0	seq19	QSCTQHQPY
Singleton	LTALRLCAY	-0	seq38	LTALRLCAY
Singleton	GIIITVAAF	-0	seq16	GIIITVAAF
Singleton	FVFLVLLPL	-0	seq11	FVFLVLLPL
Singleton	EYHDVRVVL	-0	seq22	EYHDVRVVL
Singleton	FTSDYYQLY	-0	seq43	FTSDYYQLY
Singleton	FLIVAAIVF	-0	seq52	FLIVAAIVF
Singleton	WLIVGVALL	-0	seq49	WLIVGVALL
Singleton	SFYEDFLEY	-0	seq17	SFYEDFLEY
Singleton	FVCNLLLLF	-0	seq47	FVCNLLLLF
Singleton	YFIASFRLF	-0	seq35	YFIASFRLF
Singleton	LLYDANYFL	-0	seq46	LLYDANYFL
Singleton	ITLATCELY	-0	seq53	ITLATCELY
Singleton	ILLIIMRTF	-0	seq6	ILLIIMRTF
Singleton	FVTVYSHLL	-0	seq45	FVTVYSHLL
Singleton	LQLPQGTTL	-0	seq14	LQLPQGTTL
Singleton	GTYEGNSPF	-0	seq51	GTYEGNSPF
Singleton	YYQLYSTQL	-0	seq48	YYQLYSTQL
Singleton	CVRGTTVLL	-0	seq55	CVRGTTVLL
Singleton	YVVDDPCPI	-0	seq20	YVVDDPCPI
Singleton	FLFLTWICL	-0	seq36	FLFLTWICL
Singleton	IQYIDIGNY	-0	seq15	IQYIDIGNY
Singleton	YLQPRTFLL	-0	seq10	YLQPRTFLL
Singleton	LVKPSFYVY	-0	seq37	LVKPSFYVY
Singleton	TVSCSPFTI	-0	seq21	TVSCSPFTI
Singleton	SLVVRCSFY	-0	seq18	SLVVRCSFY
Singleton	HFYSKWYIR	-0	seq23	HFYSKWYIR
Singleton	FAPSASAFF	-0	seq12	FAPSASAFF

Crossreactivity analyses against human proteome based on UniProt data (Fig. 1) showed that all the 57 predicted viral epitopes were not present in human proteome and hence, no crossreactivity to normal human cells may occur. Experimental MHC-peptide binding and T cell assays are now required for *in vitro* testing and development as potent immunogens.

Fig. 1: Multiple Peptide Match of 57 predicted coronaviral epitopes aganist *Homo sapiens* proteome from UniProt.



Helper T cell epitopes:

Helper T lymphocyte epitopes are typically 15 amino acids residues long and were generated against 661 HLA-DRB, 16 HLA-DPalpha, 129 HLA-DPbeta, 29 HLA-DQalpha alleles and 105 HLA-DQ beta alleles. High throughput data for these epitopes is currently being analysed manually to identify common epitopes across alleles and 10 coronaviral proteins.

Conclusions:

A ranked list of CTL epitopes with high HLA binding affinity, high TAP transport efficiency and high C-terminal proteasomal cleavage ranking has been generated. Two different prediction algorithms were implemented in identification of common epitopes for consensus. Immunogenicity scores for these epitopes have also been predicted in order to further narrow down the list to key few epitopes that

can be experimentally tested. Peptide matching with human proteome showed no indication of possible crossreactivity. These epitopes are provided to the scientific community for further *in vitro* and *in vivo* assays and saving their time and costs involved in our urgent bid to tackle nCoV infections and death.

Acknowledgments:

This author acknowledges the tireless help of researchers working towards nCoV control and submitting data to GenBank without which these sequence analyses using Immunoinformatics would not have been possible.

Conflict of interest: This author declares that there is no conflict of interest.

References:

1. Smith, Micholas; Smith, Jeremy C. (2020). Repurposing Therapeutics for COVID-19:Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2Interface. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.11871402.v4

2. Seema Mishra and Subrata Sinha (2009). 'Immunoinformatics and modeling perspective of T cell epitope-based cancer immunotherapy: a holistic picture' J Biomol Struct Dyn. 27(3), pp.293-306.

3. Seema Mishra and Subrata Sinha (2006). 'Prediction and molecular modeling of T cell epitopes derived from placental alkaline phosphatase for use in cancer immunotherapy'. J Biomol Struct Dyn. 24(2), pp.109-121.

4. Anthony R. Fehr and Stanley Perlman (2015). Coronaviruses: An Overview of Their Replication and Pathogenesis

Methods Mol Biol. 1282: 1–23.

5. Hsueh-Ling Janice Oh, Samuel Ken-En Gan, Antonio Bertoletti and Yee-Joo T (2012) Understanding the T cell immune response in SARS coronavirus infection Emerging Microbes and Infections(2012)1,e23; doi:10.1038/emi.2012.26.

6. Shi J, Zhang J, Li S, Sun J, Teng Y, Wu M, et al. (2015) Epitope-Based Vaccine Target Screening against Highly Pathogenic MERS-CoV: An In Silico Approach Applied to Emerging Infectious Diseases. PLoS ONE 10(12): e0144475. doi:10.1371/journal.
pone.0144475

7. Sandeep Kumar Dhanda, Kerrie Vaughan, Veronique Schulten, Alba Grifoni, Daniela Weiskopf, John Sidney, Bjoern Peters, Alessandro Sette: Development of a novel clustering tool for linear peptide sequences . Immunology (2018) doi:https://doi.org/10.1111/imm.12984