

Zika virus phosphoproteome through the computational looking-glass and what we found there?

Zika virus (ZIKV), a mosquito borne disease, has been linked to congenital anomalies in humans. Recent outbreaks of ZIKV have caused the scientific community to focus on this previously neglected pathogen. Information regarding phosphorylation potential of ZIKV proteins has yet to be elucidated. Deciphering phosphorylation potential of viral proteins is important because it assists in viral replication and its association with host receptors. Here, we present a vignette of the phosphorylation potential of ZIKV proteins using computational approach.

Zika virus phosphoproteome through the computational looking-glass and what we found there?

Zarrin Basharat*, Tabeer Khan, Azra Yasmin

Microbiology & Biotechnology Research Lab, Department of Environmental Sciences,
Fatima Jinnah Women University, 46000, Pakistan.

*Corresponding author

Email: zarrin.iiui@gmail.com

In wake of the recent outbreaks of ZIKV, which resulted in massive increase in microcephaly in the affected areas, it has recently been declared as a global health emergency by WHO (Heymann *et al.*, 2016). Rapid analysis of mechanisms underlying the pathogenesis of this virus is required for speedy development of therapeutics. Computational screening using webservers trained on experimental phosphoprotein datasets is a handy method for quick scrutinization of phosphosites in pathogenic proteins.

Virus protein phosphorylation aids in replication and disrupts the normal host-cell functions, leading to pathogenesis (Schwartz and Church, 2010). Phosphorylation and dephosphorylation of protein residues may also be involved in cell regulation/ signalling processes (Basharat and Yasmin, 2015). Phosphorylation information is thus, critical for understanding virus protein functionality and in drug design. Prediction of this post translational modification for viral proteins has been attempted previously for Ebola virus nucleoprotein (Yasri and Wiwanitkit, 2015), envelope 2 protein of Hepatitis C virus genotype 1a (Afzal *et al.*, 2011), envelope proteins of Hepatitis C genotype 5a (Gedezha *et al.*, 2014), Hepatitis G virus envelope glycoprotein E2 (Ranjbar *et al.*, 2013), whole proteome of rotavirus (Chattopadhyay *et al.*, 2010), using computational webservers.

We have used support vector machine based computational method, ViralPhos (<http://csb.cse.yzu.edu.tw/ViralPhos/>) (Huang *et al.*, 2013), for phosphorylation potential prediction of the ZIKV proteome. ViralPhos is specifically tailored for phosphorylation analysis of viral proteins and probes virus substrate site motifs for identification of potential phosphorylation sites on the viral proteins.

We identified a plethora of phosphosites in the whole proteome of ZIKV (Table 1). It is anticipated that this information will be handy in further experimental studies focused on phosphorylation of ZIKV proteins. This computational analysis provides rapid baseline information and paves way for further elucidation of the phosphorylation in ZIKV pathogenesis, until the availability of experimentally determined phosphoproteome of this virus.

References

Afzal S, Idrees M, Ali M, Ilyas M, Hussain A, Akram M, Butt S, Saleem S, Rehman I, Ali L, Shahid M. Envelope 2 protein phosphorylation sites S75 & 277 of hepatitis C virus genotype 1a and interferon resistance: A sequence alignment approach. *Viol J* 2011; **8**(1):1.

Basharat Z, Yasmin, A. (2016). *In silico* assessment of phosphorylation and O- β -GlcNAcylation in human NPC1 protein critical for Ebola virus entry. *Infect Genet Evol.* **34**: 326-338.

Chattopadhyay S, Bagchi P, Dutta D, Mukherjee A, Kobayashi N, Chawlasarkar M. Computational identification of posttranslational modification sites and functional families reveal possible moonlighting role of rotaviral proteins. *Bioinformatics* 2010; **4**(10):448.

Gededzha MP, Mphahlele MJ, Selabe SG. Characterization of HCV Genotype 5a Envelope Proteins: Implications for Vaccine Development and Therapeutic Entry Target. *Hepat Mon* 2014; **14**(11).

Heymann DL, Hodgson A, Sall AA, Freedman DO, Staples JE, Althabe F, Baruah K, Mahmud G, Kandun N, Vasconcelos PF, Bino S. Zika virus and microcephaly: why is this situation a PHEIC?. *Lancet* 2016; **387**(10020):719.

Huang K, Lu C, Bretaña NA, Lee T, Chang T. ViralPhos: incorporating a recursively statistical method to predict phosphorylation sites on virus proteins. *BMC Bioinformatics* 2013; **14**(Suppl 16):S10.

Ranjbar MM, Ghorban K, Alavian SM, Keyvani H, Dadmanesh M, Ardakany AR, Motedayen MH, Sazmand A. GB virus C/hepatitis G virus envelope glycoprotein E2: computational molecular features and immunoinformatics study. *Hepat Mon* 2013; **13**(12).

Schwartz D, Church GM. Collection and motif-based prediction of phosphorylation sites in human viruses. *Sci Signal* 2010; **3**(137) :rs2.

Yasri S, Wiwanitkit V. Phosphorylation sites within Ebola virus nucleoprotein. *Asian Pac J Trop Biomed* 2015; **5**(7): 598-9.

Table 1. Identification of phosphorylated sites on ZIKV proteome consisting of proteins.

Reference genome sequence with Accession no: YP_009227198.1 was used as an information source of ZIKV proteins. Phosphorylated residue in the protein fragment is underlined.

Protein	Location	Length (amino acids)	Phosphorylated residue	Fragment containing phosphorylated residue
Glycoprotein E	291-790	500	322	HGGCV <u>T</u> VMAQD
			337	DIELV <u>T</u> TTVSN
			338	IELV <u>T</u> TTVSNM
			339	ELV <u>T</u> TTVSNMA
			366	DSRC <u>P</u> TQGEAY
			371	TQGEA <u>Y</u> LDKQS
			405	KGSLV <u>T</u> CAKFT
			416	CSKKM <u>T</u> GKSIQ
			436	LSVHG <u>S</u> QHSGM
			446	MIGYE <u>T</u> DEDRA
			456	AKVEV <u>T</u> PNSPR
			459	EVT <u>P</u> NSPRAEA
			465	PRAEA <u>T</u> LGGFG
			471	LGGFG <u>S</u> LGLDC
			480	DCEP <u>R</u> TGLDFS
			517	HAGAD <u>T</u> GTPHW
			540	HAKRQ <u>T</u> VVVVLG
553	EGAVH <u>T</u> ALAGA			

			590	RLKGV <u>S</u> YSLCT
			595	SYSLCT <u>A</u> AFTF
			601	AAFTF <u>T</u> KVPAE
			611	ETLHG <u>T</u> VTVEV
			613	LHGTV <u>T</u> VEVQY
			618	TVEVQ <u>Y</u> AGTDG
			639	DMQTL <u>T</u> VPVGR
			646	VGRLI <u>T</u> ANPVI
			654	PVITE <u>S</u> TENSK
			655	VITE <u>S</u> TENSKM
			672	PFGDS <u>Y</u> IVIGV
			683	GDKKI <u>T</u> HHWHR
			691	WHRSG <u>S</u> TIGKA
			700	KAFEAT <u>V</u> RGAK
			713	AVLGD <u>T</u> AWDFG
			719	AWDFG <u>S</u> VGGVF
			769	NTKNG <u>S</u> ISLTC
			771	KNGSIS <u>L</u> TCLA
			773	KNGSIS <u>L</u> TCLA
Capsid protein C	12-122	111	71	INRWG <u>S</u> VGKKE
DEXDc	1679-1824	146	1688	KKKQL <u>T</u> VLDLH
			1699	PGAGK <u>T</u> RRVLP
			1717	KKRLR <u>T</u> VILAP
			1723	VILAP <u>T</u> RVVAA
			1743	PVRYM <u>T</u> TAVNV
			1749	TAVNV <u>T</u> HSGTE

			1753	VTHSGTEIVDL
			1763	LMCHATFTSRL
			1788	DEAHFTDPSSI
			1791	HFTDPSSIAAR
			1801	RGYISTRVEMG
			1816	IFMTATPPGTR
NS2A	1154- 1368	215	1171	LKKRM ¹ TKIIM
			1172	KKRMTTKIIMS
			1178	KIIMSTISMAVL
			1212	FAEMNTGGDVA
			1242	FRANWTPRESM
			1246	WTPRESMLLAL
			1258	SCLLQTAISAL
			1307	PLARGTLLVAW
			1318	RAGLATCGGIM
			1332	LKGKGSVKKNL
			1363	GLLLLTRSGKR
NS2B	1372- 1498	127	1377	PSEVLTAVGLI
			1413	VSYVVSGKSVD
			1420	KSVDMYIERAG
			1436	KDAEVTGNSPR
			1439	EVTGN ¹ PRLDV
			1449	VALDESGDFSL
			1453	ESGDFSLVEED
NS3	1499- 2115	617	1517	KKGETTDGVYR
			1521	TTDGVY ¹ VMTR

			1531	RRLLGSTQVG <u>V</u>
			1532	RLLG <u>S</u> TQVG <u>V</u> G
			1551	TMWHVTKGAAL
			1576	KQDLV <u>S</u> YCGPW
			1616	PGIFKTKDGDI
			1635	AGTSG <u>S</u> PILDK
			1658	VIKNG <u>S</u> YVSAI
			1672	KREEE <u>T</u> PVECF
			1688	KKKQL <u>T</u> VLDLH
			1699	PGAGK <u>T</u> RRVLP
			1717	KKRLR <u>T</u> VILAP
			1723	VILAP <u>T</u> RVVAA
			1743	PVRYM <u>T</u> TAVNV
			1749	TAVNV <u>T</u> HSGTE
			1753	VTHSG <u>T</u> EIVDL
			1763	LMCHA <u>T</u> FTSRL
			1788	DEAH <u>F</u> TDPSSI
			1791	H <u>F</u> TDP <u>S</u> SIAAR
			1801	RGYIS <u>T</u> RVEMG
			1816	IFMTA <u>T</u> PPGTR
			1820	A <u>T</u> PPG <u>T</u> RD <u>A</u> FP
			1829	FPDSN <u>S</u> PIMDT
			1834	SPIMD <u>T</u> EVEVP
			1844	PERAW <u>S</u> SGFDW
			1863	VWFV <u>P</u> SVRNGN
			1875	IAAC <u>L</u> TKAGKR

			1888	QLSRK <u>T</u> FETEF
			1947	GPMPV <u>T</u> HASAA
			1978	GGCAE <u>T</u> DEGHA
			2024	EFKLR <u>T</u> EQRKT
			2029	TEQRK <u>T</u> FVELM
			2068	GTTNN <u>T</u> IMEDS
			2073	TIMED <u>S</u> VPAEV
			2080	PAEVW <u>T</u> KYGEK
			2106	HAALK <u>S</u> FKEFA
NS1	792- 1144	353	803	FSKKE <u>T</u> RCGTG
			807	ETRCG <u>T</u> GVFIY
			812	TGVFI <u>Y</u> NDVEA
			828	KYHPD <u>S</u> PRRLA
			877	NGVQL <u>T</u> VVVGS
			919	VRAAK <u>T</u> NNSFV
			922	AKTNN <u>S</u> FVVDG
			942	HRAWN <u>S</u> FLVED
			955	FGVFH <u>T</u> SVWLK
			966	VREDY <u>S</u> LECDP
			976	PAVIG <u>T</u> AVKGR
			1012	LIEMK <u>T</u> CEWPK
			1020	WPKSH <u>T</u> LWTDG
			1023	SHTLW <u>T</u> DGVEE
			1046	LSHHN <u>T</u> REGYR
			1052	REGYR <u>T</u> QVKGP
			1073	EPCPG <u>T</u> KVYVE

			1080	VYVEE <u>T</u> CGTRG
			1083	EETCG <u>T</u> RGPSL
			1087	GTRG <u>P</u> SLRSTT
			1090	G <u>P</u> SLR <u>S</u> TTASG
			1091	PSLR <u>S</u> TTASGR
			1092	SLR <u>S</u> TTASGRV
			1107	CCRE <u>C</u> TMPPLS
NS4B	2266- 2516	251	2290	REEGAT <u>M</u> MGFSM
			2308	SAWAI <u>Y</u> AALTT
			2312	IYAAL <u>T</u> TLITP
			2316	LTTL <u>I</u> TPAVQH
			2324	VQHAV <u>T</u> TSYNN
			2325	QHAV <u>T</u> TSYNNY
			2326	HAV <u>T</u> TSYNNYS
			2327	AV <u>T</u> TSYNNYSL
			2404	AAQR <u>T</u> AAGIM
			2424	VT <u>D</u> IDTMTIDP
			2465	AGAL <u>I</u> TAATST
			2475	TLWEG <u>S</u> PNKYW
			2483	KYWNS <u>S</u> TATSL
			2505	ASLI <u>Y</u> TVTRNA
NS4A	2119- 2263	145	2127	MEALG <u>T</u> LPGHM
			2133	LPGHM <u>T</u> ERFQE
			2153	RAETG <u>S</u> RPYKA
			2156	TGSR <u>P</u> YKAAAA
			2166	AQLPET <u>L</u> ETIM

			2243	PEKQRSPQDNQ
FtsJ	2571- 2742	172	2586	LEERGYLQPYG
			2605	RGGWSY ^Y AATI
			2620	EVRGY ^T TKGGPG
			2667	DIGESS ^S SPEV
			2668	IGESS ^S SPEVE
			2669	GESS ^S SPEVEE
			2675	PEVEE ^T RTRLRV
			2704	VLCPY ^T STMME
			2706	CPYT ^T STMETM
			2710	STMME ^T MERLQ
			2731	PLCRN ^S THEMY
			2732	LCRN ^S THEMYW
			2736	STHEMY ^W VSGA
HELICc	1855- 1971	117	1863	VWFP ^S VRNGN
			1875	IAACL ^T KAGKR
			1888	QLSRK ^T FETEF
			1947	GPMPV ^T HASAA
NS2B	1369- 1498	130	1377	PSEVL ^T AVGLI
			1413	VSYV ^S GKSVD
			1420	KSVD ^M YIERAG
			1436	KDAEV ^T GNSPR
			1439	EVTGN ^S PRLDV
			1449	VALDE ^S GDFSL
			1453	ESGDF ^S LVEED
S7	1516-	151	1521	TTDGV ^Y RMTR

	1666		1531	RRLLGSTQVGV
			1532	RLLGSTQVGVG
			1551	TMWHVTKGAAL
			1576	KQDLVSYCGPW
			1616	PGIFKTKDGDI
			1635	AGTSGSPILDK
			1658	VIKNGSYVSAI
E stem	694-790	97	700	KAFEATVRGAK
			713	AVLGDTAWDFG
			719	AWDFGSVGGVF
			769	NTKNGSISLTC
			771	KNGSISLTCLA
			773	GSISLTCLALG
Glycoprotein M	216-290	75	224	LPSHSTRKLQT
			233	QTRSQTWLESR
			267	AWLLGSSTSQK
			271	WLLGSSTSQKV
			272	LLGSSTSQKVI
			273	LGSSTSQKVIY