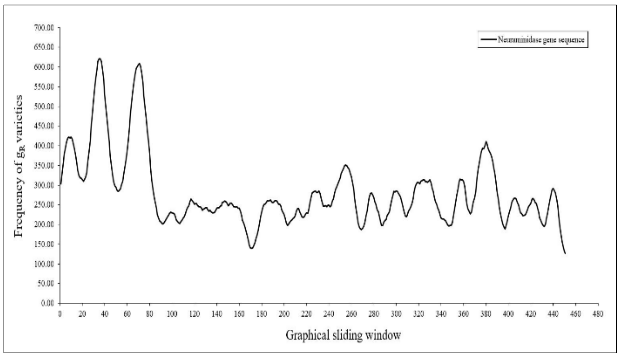
MOL2NET, International Conference Series on Multidisciplinary Sciences

H1N1 – First 100 Years: Regions of Least Variability in Neuraminidase Gene Sequence

Proyasha Roy[a], Tathagata Dutta[a] and Ashesh Nandy[a]

[a] *Centre for Interdisciplinary Research and Education, 404B Jodhpur Park, Kolkata 700058, India*

Graphical Abstract



Abstract.

A century has passed since the first global pandemic was brought about by the influenza A (H1N1) virus in 1918. Using the graphical representation and numerical characterization method, a survey of all the H1N1 neuraminidase (NA) gene sequences over the period between 1918 and 2018 shows that certain regions have remained well conserved while others continue to exhibit high variability. Regions of low variability in the NA protein and associated high solvent accessibility identified by our analysis provides an indication of vaccines targeting for longer lifetime applicability. In light of the resistance developed by influenza to antivirals and vaccines, consideration of this analytical study will aid in improving the efficacy of viral treatment.

Introduction

100 years ago, during the fading years of World War I in 1918, influenza A (H1N1) virus swept across the globe infecting an estimated half a billion people. Although secondary pneumonia infection was a major cause of the extremely high mortality rate, a reported 25 - 50 million people succumbed to the influenza A virus [1]. This was followed by two more influenza pandemics in 1957-58 and 1968-69, and an outbreak in Hong Kong in the year of 1997, albeit of different influenza types [2,3]. The H1N1 "Swine flu" in 2009 is currently the last known influenza pandemic. Treatment involves vaccination and antivirals that include neuraminidase inhibitors like oseltamivir, zanamivir, laninamivir and peramivir [4], and M2 proton channel blockers like amantadine and rimantadine [5]. The influenza virus being an RNA virus with no self-repairing mechanisms, high rates of mutation of the genome are seen during

replication due to which seasonal vaccines ("flu shots") with varying compositions need to be developed and manufactured annually as is the practice in the United States.

The influenza virus genome is a negative-sense single-stranded RNA that has 8 segments encoding 10-14 proteins based on the type of influenza virus. The structural genes comprise of PB2, PB1, PA, HA, NP, NA, M1 and NS1 in sequential order. The M1 and NS1 are post-translationally spliced to form M2 and NEP, respectively [6].

In our study, all H1N1 neuraminidase (NA) sequences available in the Virus Variation Resource were analyzed and the regions that have remained conserved in the last 100 years were identified as potential candidates for peptide vaccines. An alignment-free graphical sliding window method (GSWM) based on 2D graphical analysis [7,8] permitted fast and highly sensitive computation which the widespread multiple sequence alignment algorithms cannot yet deliver. A similar study was carried out earlier by Ghosh *et al.* 2010 [9].

Materials and Methods

10,920 full-length coding sequences (1410 nt) of H1N1 NA gene were retrieved from the Virus Variation Resource database on 23rd October, 2018 for all host types, regions and time periods. There were 2 sequences with unknown year of collection but were included in the dataset for analysis nonetheless. There was one sequence found for each of the periods in 1910 – 1929 and 1950 – 1969. Considering the NA protein, 6,136 sequences of length 469 amino acids were also similarly obtained; there were no protein sequences available for the period 1920 - 1929.

The graphical method proposed by Nandy [7] involves representation of a gene sequence in a 2D Cartesian plane using the ACGT axes system. In order to quantify the composition and distribution of nucleotide bases in a gene sequence, the numerical descriptor, graph radius (g_R) was calculated. Identical base distribution in any two or more sequences yields identical graph radii. Alternatively, dissimilar g_R's imply dissimilar sequences. To analyze the conserved regions in the gene sequences, the graphical sliding window method (GSWM) was employed. A window of 30 nucleotides was used to scan the entire length of each sequence, sliding ahead at every third nucleotide, thus, moving forward at every codon in the coding sequences. The 30 nt sliding window corresponds to the length of 10 amino acid peptides that HLA II can recognize. From our understanding of the g_R value, across the 10,920 sequences, identical base distributions in any window region will display identical g_R values; a window region in any two or more sequences displaying the same g_R value implies that the base composition and distribution in the window regions in the two sequences are identical. Regions with high variability will result in multiple g_R values for each window of the sequences due to dissimilar base compositions and distributions among them. Therefore, one window may have at least one variety of g_R or more. Table 1 displays a portion of the g_R calculated for the first 4 windows of NA gene accession numbers CY089804, HM189306, HQ695939 and JF820279 as example.

Table 1. A portion of the g_R calculated for the first 4 windows of four NA gene sequences. The number of varieties seen in these four windows are given below.

Accession No.	**Year**	**Country**	**Host**	**g_R (1-30)**	**g_R (4-33)**	**g_R (7-36)**	**g_R (10-39)**
CY089804	2010	Thailand	Swine	6.106917753	7.034597043	5.973366815	5.845130927
HM189306	2010	China	Swine	6.106917753	7.034597043	5.907528154	5.764354064
HQ695939	2010	China	Swine	6.106917753	7.034597043	5.907528154	5.570258322
JF820279	2011	China	Swine	6.106917753	7.034597043	5.907528154	5.764354064
			Variety	**1**	**1**	**2**	**3**

At the protein level, the sequences were projected onto a 20-dimensional space with each axis corresponding to an amino acid [10]. Similar to the graph radius g_R, the numerical descriptor for the protein graphs are defined by p_R. A GSW of 10 amino acids was utilized to scan the protein sequences for p_R varieties for each window. The window progressed forward at each residue position. Solvent accessible surface areas (ASA) were also analyzed for the NA protein using the SABLE server.

Results and Discussion

Based on the length of the sequence and the GSW size, 461 windows were determined for all the NA 10,920 gene sequences. The windows were sequentially numbered. Window number 455 of nucleotide positions 1363 to 1392 contained 86 varieties, the lowest number (See Table 2), while window number 39 of nucleotide positions 115 to 144 displayed 686 varieties, the highest. Figure 1 displays a visual representation of the number of g_R varieties for each window in which the vertical axis corresponds to the number or frequency of g_R varieties while the horizontal axis denotes the window number. A moving average of size 10, i.e., an average of 10 g_R values, was calculated for successive windows in order to generate a smoother graph.

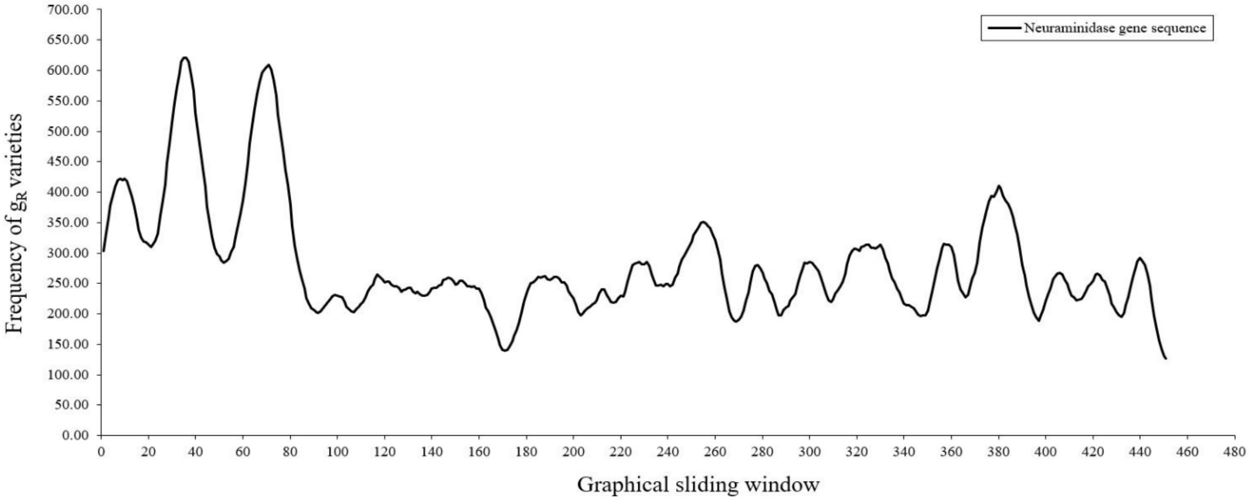


Figure 1. A visual representation of the number of g_R varieties for each window in which the y-axis denotes the frequency of g_R varieties and the x-axis denotes the window number of the sequences. A moving average of size 10, i.e., an average of 10 g_R values, was calculated for successive windows in order to generate a smooth graph.

Table 2. The 10 lowest varieties of the g_R values of all 10,920 NA gene sequences.

Nucleotide position	Window number	Frequency of variability
1363-1392	455	86
1360-1389	454	102
1366-1395	456	104
532-561	178	114
529-558	177	116
1369-1398	457	118
1357-1386	453	120
823-852	275	127
526-555	176	132
1111-1140	371	134
1375-1404	459	134

In an identical manner, the p_R varieties generated showed that the protein region between 292 and 303 had the lowest variability with 18 descriptor varieties while the region between 73 and 84 had the highest number of varieties, i.e., 624. Table 3 summarizes the least variable regions in the top 3 percentile.

Table 3. The 10 lowest varieties of the p_R values of all 6,136 NA protein sequences.

Amino acid position	Window number	Frequency of variability
292-303	292	18
291-302	291	23
290-301	290	24
293-304	293	25
176-187	176	26
400-411	400	30
175-186	175	30
294-305	294	31
295-306	295	31
171-182	171	32
401-412	401	32
172-183	172	33
170-181	170	34
174-185	174	34
272-283	272	34

Figure 2 graphically depicts the varieties along the NA protein with average solvent accessibility (ASA). The ASA analysis was carried out in order to determine whether the least variable windows were surface exposed. This exercise is based on the fact that such surface exposed regions with high amino acid conservation (least variability in p_R values) prove to be effective candidates for peptide vaccines that have the potential to replace the need for annually developing and manufacturing expensive flu vaccines. The ASA indices were also treated with the moving average method of size 10. With a threshold of more than 50% of the ASA range (5.6 to 44.5), Table 4 summarizes the least variable regions in the top 25 percentile range with high solvent accessibility surface exposure ascertained from the graph. Interestingly, similar features around the 176, 275, 410 and 430 windows reflect our determination of such suitable targets in H5N1 NA [9] with only 514 sequences, implying good conservation among closely related influenza subtypes over the 100year period.

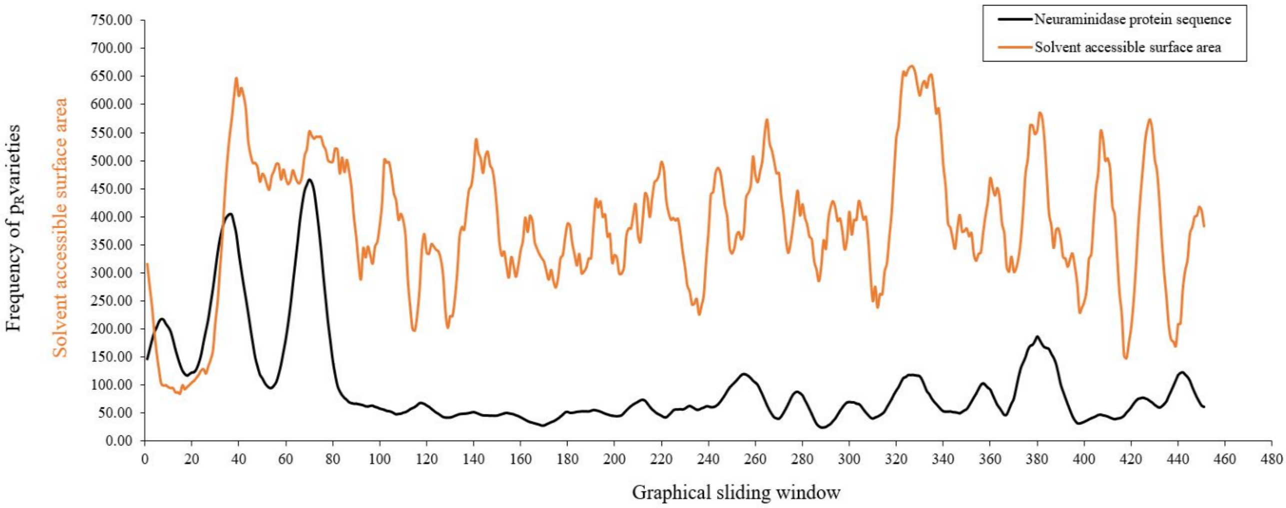


Figure 2. ASA (orange) and protein variety profiles (black) of H1N1 neuraminidase protein sequences of our sample database. Low protein variety and high ASA regions mark best candidates for long lasting peptide vaccines.

Table 4. Amino acid window regions and protein varieties determined as best candidates for peptide vaccines

Amino acid window	Frequency of variability
125	44.3
159	43.0
176	48.0
194	46.2
205	44.3
230	44.9
231	45.1
241	44.8
264	46.3

Conclusions

Extensive research has been done in understanding the conserved regions and the hotspots of H1N1 genomes and proteins. Our study looked at sequences of neuraminidase NA gene and protein over a 100-year period from 1918 to 2018. Graphical characterization and numerical representation of the sequences showed the nucleotide regions between 1363 and 1392 to be least variable and between 115 and 144 to be most variable over the period. There were 10 conserved regions in the NA gene which were found in the top 3 percentile range. In the protein, the minimum variability was expressed by the region between amino acids 292 and 303, while the region between 73 and 84 had the maximum variability. 11 peptide regions with the lowest variability that corresponded to high solvent accessibility regions over the 100-year period were determined which could serve as effective peptide vaccines over many cycles of mutations. That some of these regions were strongly conserved even among some influenza subtypes is borne out by comparison with our earlier study [9] of H5N1 NA. It is indeed surprising how strongly these features continue to exist in the influenza neuraminidase structure.

References

1. Taubenberger JK; Morens DM. 1918 Influenza: the mother of all pandemics. *Emerg Infect Dis.* 2006, 12(1), 15-22.

2. Influenza pandemics of the 20th century. *Emerg Infect Dis.* 2006, 12(1), 9-14.

3. Snacken R; Kendal AP; Haaheim LR; Wood JM. The next influenza pandemic: lessons from Hong Kong, 1997. *Emerg Infect Dis.* 1999, 5(2), 195-203.

4. Heneghan CJ; Onakpoya I; Jones MA; et al. Neuraminidase inhibitors for influenza: a systematic review and meta-analysis of regulatory and mortality data. *Health Technol Assess.* 2016, 20(42), 1-242.

5. Hay AJ; Wolstenholme AJ; Skehel JJ; Smith MH. The molecular basis of the specific anti-influenza action of amantadine. *EMBO J.* 1985, 4(11), 3021-4.

6. Bouvier NM; Palese P. The biology of influenza viruses. *Vaccine* 2008, 26(Suppl 4), D49-53.

7. Nandy A. A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Application to Globin Genes. *Curr Sci.* 1994, 66, 309-314.

8. Raychaudhury C; Nandy A. Indexing scheme and similarity measures for macromolecular sequences. *J Chem Info and Comput Sci* 1999, 39, 243-247.

9. Ghosh A; et al. Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase. *BMC Structural Biology* 2010, 10, 6.

10. Nandy A; Ghosh A; Nandy P. Numerical Characterization of Protein Sequences and Application to Voltage-Gated Sodium Channel Alpha Subunit Phylogeny. *In Silico Biology* 2009, 9, 77-87.