Journal of
Medical Microbiology & Diagnosis

Weltman, J Med Microb Diagn 2019, 8:2
DOI: 10.4172/2161-0703.1000301

Research Article Open Access

Identical Subsequences of Contiguous Amino Acids in Influenza Virus Hemagglutinin and in Human Proteins

Joel K Weltman*

Faculty of Medicine, Alpert Medical School, Brown University, USA

Abstract

Background: Influenza virus is a significant public health problem throughout the world. Increased insight into the basic biology of the virus may enable the development of more effective anti-influenza preventives and therapeutics.

Methodology: The occurrence of specific amino acid subsequences in H1N1 and H3N2 influenza virus hemagglutinins was used for joint detection of those subsequences in human proteins. Only subsequences consisting of at least 5 contiguous amino acids were considered for further study.

Results: Ten H1N1 hemagglutinin amino acid subsequences and nine H3N2 hemagglutinin amino subsequences were identified as also occurring in proteins of human origin. The length of the subsequences selected for further study, ranged from 5 contiguous amino acids to 8 contiguous amino acids.

Conclusion: The joint occurrence of amino acid subsequences in influenza hemagglutinins and in human proteins may help explain the relatively low efficacy of current anti-influenza vaccines. It is proposed that the identification of the joint subsequences may be useful for the improved design of anti-influenza therapeutics and especially anti-influenza vaccines.

Keywords: Influenza virus; H1N1; H3N2 influenza virus; Anti-influenza vaccines

Introduction

Influenza remains a highly significant global public health problem for which vaccination is an extremely important preventive intervention [1,2]. However, currently used anti-influenza vaccines are only 43-62% effective [3]. Therefore, it is extremely urgent and important for us to increase our understanding of influenza virus biology so that more effective vaccines and therapeutic agents can be designed.

This is a report of the occurrence of amino acid subsequences within sequences of hemagglutinin proteins of influenza virus subtypes H1N1, H3N2 and also within human proteins. Understanding the biology of protein subsequences which occur both in the influenza virus and in the human host may facilitate the design of anti-influenza vaccines and drugs.

This research is a continuation of the previous report of the occurrence of identical tetrapeptide subsequences in influenza hemagglutinins and in human proteins [4]. The present report addresses identical subsequences of at least pentapeptide length.

Materials and Methods

Entire sets of full-length (566 amino acids) hemagglutinin (HA) proteins of influenza H1N1 virus (16,679 sequences) and influenza H3N2 virus (19,641 sequences) were downloaded from the NCBI Influenza Virus Database (https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database) on 08 Feb 2019. A consensus sequence of each of the two datasets was determined with JalView 2.10.5 [5].

Computing was performed with 64-bit Anaconda3 Python 3.7.1 (https://www.anaconda.com/distribution/). Information entropy (H) (in bits) was computed by the method of Shannon [6].

Viral HA protein 3-state secondary structure (h=alpha helix, e=extended strand and c=random coil) was computed on the Raptorx website [7]. Numpy arrays of the multiplication products of the secondary protein structures were obtained by multiplying each H1N1 HA secondary structure array (length=566 aa positions) by the corresponding H3N2 HA secondary structure array (length=566 aa positions) according to equations 1, 2 and 3, where the symbol × represents element-wise multiplication:

$$\text{helix product} = h(\text{H1N1}) \times h(\text{H3N2}) \quad \text{(equation 1)}$$

$$\text{extended strand product} = e(\text{H1N1}) \times e(\text{H3N2}) \quad \text{(equation 2)}$$

$$\text{random coil product} = c(\text{H1N1}) \times c(\text{H3N2}) \quad \text{(equation 3)}$$

Because of a sequence length limitation, the sequences from viral HA helical region aa 385-427 were split into two subsequences, 20 and 23 amino acids in length, prior to BLASTP analysis. The consensus amino acids at positions with multiplication product values ≥ 0.75 in the influenza HA product array were selected as probes for matching subsequences in human proteins on the NCBI BLASTP website.

The complete set of non-redundant (nr) sequences of human (*Homo sapiens* (tax id: 9606)) proteins on the National Center for Biotechnology Information database (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) was screened with BLASTP for occurrence of the subsequences that had been identified within the secondary structures of the H1N1 and H3N2 influenza virus HA proteins [8]. In each case,

***Corresponding author:** Joel K Weltman, Clinical Professor Emeritus of Medicine, Alpert Medical School, Brown University, Providence, RI 02912, USA, Tel: 4012457588; E-mail: joel_weltman@brown.edu

Received March 26, 2019; **Accepted** May 04, 2019; **Published** May 10, 2019

Citation: Weltman JK (2019) Identical Subsequences of Contiguous Amino Acids in Influenza Virus Hemagglutinin and in Human Proteins. J Med Microb Diagn 8: 301. doi:10.4172/2161-0703.1000301

the 100 most frequent influenza subsequences within the human protein sequences were downloaded. Each subsequence within the downloaded BLASTP results was considered for further analysis if that subsequence contained at least 5 contiguous amino acids.

Results

The downloaded H1N1 and H3N2 HA sequence sets yielded the following full-length consensus amino acid sequences:

H1N1_HA=MKAILVVLLYTFATANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDKHNGKLCKLRGVAPLHLGKCNIAGWILGNPECESLSTASSWSYIVETSSSDNGTCYPGDFIDYEELREQLSSVSSFERFEIFPKTSSWPNHDSNKGVTAACPHAGAKSFYKNLIWLVKKGNSYPKLSKSYINDKGKEVLVLWGIHHPSTSADQQSLYQNADAYVFVGTSRYSKKFKPEIAIRPKVRDQEGRMNYYWTLVEPGDKITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPITIGKCPKYVKSTKLRLATGLRNVPSIQSRGLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAADLKSTQNAIDKITNKVNSVIEKMNTQFTAVGKEFNHLEKRIENLNKKVDDGFLDIWTYNAELLVLLENERTLDYHDSNVKNLYEKVRSQLKNNAKEIGNGCFEFYHKCDNTCMESVKNGTYDYPKYSEEAKLNREEIDGVKLESTRIYQILAIYSTVASSLVLVVSLGAISFWMCSNGSLQCRICI

and

H3N2_HA=MKTIIALSYILCLVFAQKLPGNDNSTATLCLGHHAVPNGTIVKTITNDRIEVTNATELVQNSSIGEICDSPHQILDGENCTLIDALLGDPQCDGFQNKKWDLFVERSKAYSNCYPYDVPDYASLRSLVASSGTLEFNNESFNWTGVTQNGTSSACIRRSNSSFFSRLNWLTHLNYKYPALNVTMPNNEQFDKLYIWGVHHPGTDKDQIFLYAQSSGRITVSTKRSQQAVIPNIGSRPRIRDIPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRSGKSSIMRSDAPIGKCKSECITPNGSIPNDKPFQNVNRITYGACPRYVKQSTLKLATGMRNVPEKQTRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGRGQAADLKSTQAAIDQINGKLNRLIGKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDLWSYNAELLVALENQHTIDLTDSEMNKLFEKTKKQLRENAEDMGNGCFKIYHKCDNACIGSIRNGTYDHNVYRDEALNNRFQIKGVELKSGYKDWILWISFAISCFLLCVALLGFIMWACQKGNIRCNICI

The secondary structures of these consensus amino acid sequences are shown in Figure 1. In Figure 1, there is significant correlation between the H1N1 and H3N2 distributions of helices (Spearman rho=0.8062, $p=1.1389 \times 10^{-130}$), beta strands (Spearman rho=0.6761, $p=7.8722 \times 10^{-77}$) and random coils (Spearman rho = 0.5764, $p = 2.016 \times 10^{-51}$).

The secondary structure distributions shown in Figure 1 were used to calculate a helix multiplication product array (equation 1), an extended strand multiplication product array (equation 2) and a random coil multiplication product array (equation 3) by multiplying corresponding secondary structure values at each of the H1N1 and H3N2 amino acid positions. These resulting element-wise multiplication product arrays are shown in Figure 2.

The distributions of H1N1 × H3N2 influenza hemagglutinin secondary structure multiplication product domains shown in Figure 2 are much simpler than the pre-multiplication distributions. Each simplified domain that has a multiplication product equal to or greater than 0.75 and that also contains five or more contiguous amino acids is marked in red. The large multiplication product and high contiguity cut off values were both invoked to minimize the effects of randomization on the outcomes of this study. There are five of these red-marked domains with alpha helical structure, two marked domains with extended strand structure and three marked domains with random coil structure. The amino acid composition of each of these 10 regions is next given in Table 1.

The minimum sequence length in Table 1 is 5 contiguous amino acids and the maximum is 43 contiguous amino acids. Each of the influenza virus amino acid sequences indicated in Table 1 was used as a probe in a BLASTP search of human protein sequences. Because of their length of 43 amino acids, the H1N1 and H3N2 subsequences from viral HA helical region 385-427 were each fractionated into subse-

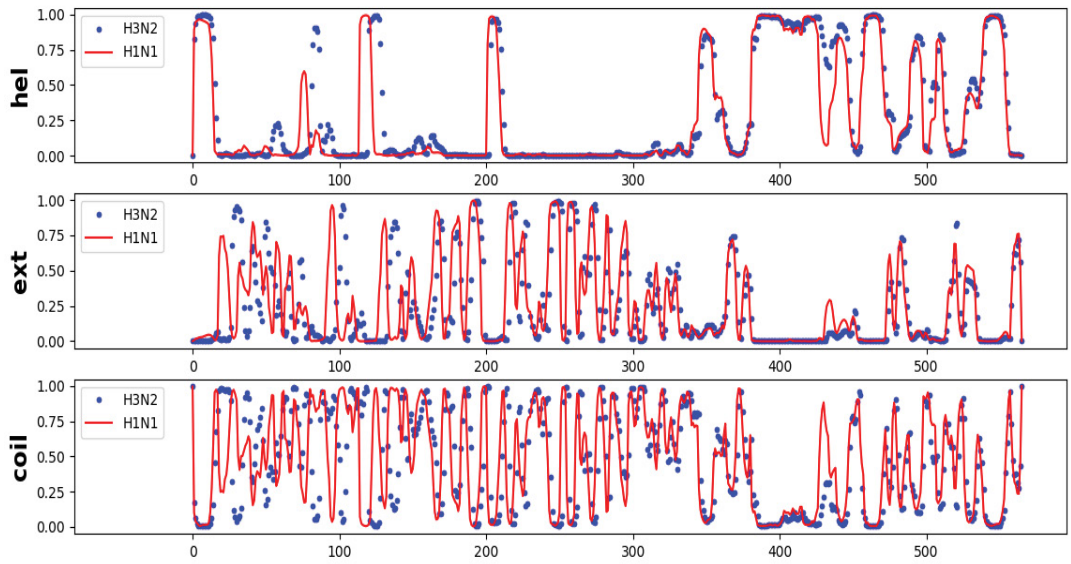

Figure 1: 3-State Secondary Structure of H1N1 and H3N2 Hemagglutinins. hel=alpha helix, ext=extended strand and coil=random coil.

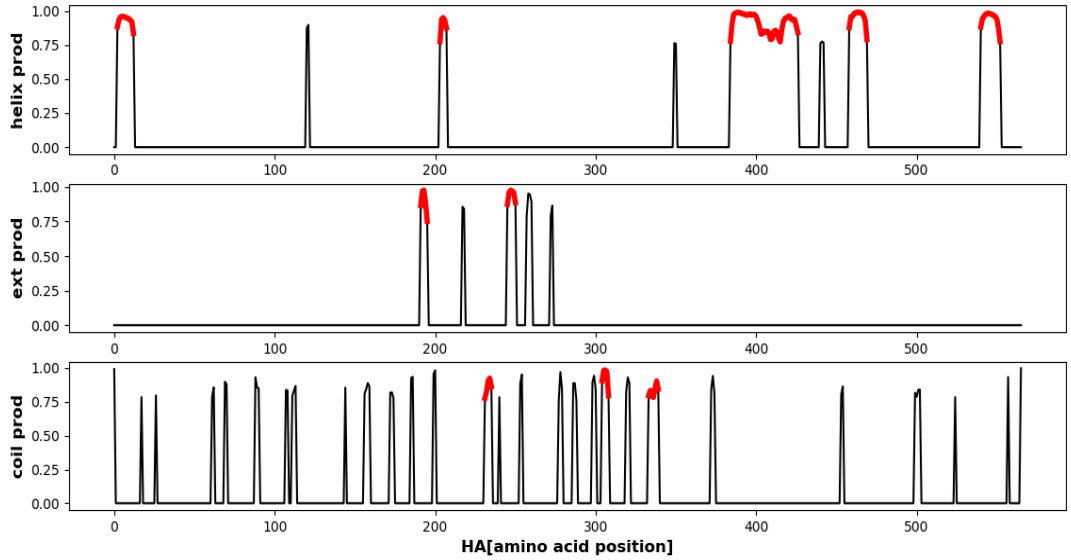

Figure 2: Distribution of H1N1*H3N2 Influenza Hemagglutinin Secondary Structure Multiplication Product Arrays. Element-wise multiplication product arrays are represented by solid black lines. Multiplication product arrays with values >= 0.75 and which consist of at least five contiguous amino acids are colored red.

Alpha helix (aa:3-13) len=11 AILVVLLYTFA TIIALSYILCL
Alpha helix (aa:204-208) len=5 DQQSL DKDQI
Alpha helix (aa:385-427) len=43 TQNAIDKITNKVNSVIEKMNTQFTAVGKEFNHLEKRIENLNKK STQAAIDQINGKLNRLIGKTNEKFHQIEKEFSEVEGRIQDLEK
Alpha helix (aa:459-470) len=12 VKNLYEKVRSQL EMNKLFEKTKKQ
Alpha helix (aa:541-553) len=13 LVLVVSLGAISFW FLLCVALLGFIMW
Extended strand (aa:192-196) len=5 VLWGI KLYIW
Extended strand (aa:246-251) len=6 YYWTLV ISIYWT
Random coil (aa:232-236) len=5 AIRPK NIGSR
Random coil (aa:305-309) len=5 TSLPF PNDKP
Random coil (aa:334-340) len=7 GLRNVPS TGMRNVP

Table 1: Influenza Hemagglutinin (HA) Consensus Amino Acid Sequences for Use as BLASTP Probes for Identical Subsequences in Human Proteins. H1N1 sequences are colored red and H3N2 sequences are colored blue. Amino acid positions refer to positions within the intact influenza HA protein molecule.

quences of length 20 and 23 amino acids to enable the BLASTP search of proteins of human origin. The longest subsequences each consist of 43 contiguous amino acids and reside in helical domains of the HA molecules. There are 4 pairs of subsequences with the minimum length of 5 contiguous amino acids. Two of these minimum length pairs have random coil structure, one of these pairs consist of extended strands and one pair consists of alpha helices. The subsequences of greatest length (11, 12, 13 and 43 amino acids) reside in alpha helical domains of the consensus sequences.

The H1N1 and H3N2 influenza HA subsequences shown above in Table 1 were next used as BLASTP probes for detection of matching subsequences in human proteins. Specific examples of the detected human: influenza virus matching subsequences are presented below. The observed length of matching contiguous amino acid components of influenza virus hemagglutinins and human proteins is presented as contiguous length. H1N1 sequences and associated information are coloured red and H3N2 sequences and associated information are coloured blue.

helix aa: 3-13 contiguous length=6, contiguous length=8

> XP_011542508.1 desumoylating isopeptidase 2 isoform X3 [Homo sapiens]

Influenza Virus ILVVLL-YT

Human ILVVLLSYT

> 6N4X_A Chain A, Metabotropic glutamate receptor 5 [Homo sapiens]

6N4X_B Chain B, Metabotropic glutamate receptor 5 [Homo sapiens]

Influenza Virus TIIALSYILCL

Human TIIALSYIFCL

helix aa:204-208 contiguous length = 5, contiguous length = 5

> XP_016874588.1 glutamate receptor-interacting protein 1 isoform X2 [Homo sapiens]

Influenza Virus DQQSL

Human DQQSL

> NP_872336.2 myelin regulatory factor-like protein [Homo sapiens]

Q96LU7.2 RecName: Full=Myelin regulatory factor-like protein

> XP_016874456.1 myelin regulatory factor-like protein isoform X1 [Homo sapiens]

> XP_016874457.1 myelin regulatory factor-like protein isoform X2 [Homo sapiens]

> XP_016874458.1 myelin regulatory factor-like protein isoform X3 [Homo sapiens]

> XP_016874459.1 myelin regulatory factor-like protein isoform X4 [Homo sapiens]

> XP_011537354.1 myelin regulatory factor-like protein isoform X5 [Homo sapiens]

Influemza Virus DKDQI

Human DKDQI

helix aa:385-427 contiguous length = 6, contiguous length = 6

> XP_011514027.1 alpha-aminoadipic semialdehyde synthase, mitochondrial isoform X1 [Homo sapiens]

Influenza Virus ITNKVNSV

Human ITNKVNMV

> 6BSZ_A Chain A, Human mGlu8 Receptor complexed with glutamate

6BSZ_B Chain B, Human mGlu8 Receptor complexed with glutamate

Influenza Virus AIDQIN

Human AIDQIN

helix aa:459-470 contiguous length = 5, contiguous length = 7

> XP_005255680.1 protein FAM234A isoform X3 [Homo sapiens]

XP_016879252.1 protein FAM234A isoform X3 [Homo sapiens]

XP_016879253.1 protein FAM234A isoform X3 [Homo sapiens]

Length = 543

Influenza Virus VKNLYEKV

Human VKGLYEKV

> NP_001018126.1 caveolae-associated protein 4 [Homo sapiens]

Q5BKX8.2 RecName: Full=Caveolae-associated protein 4; AltName: Full = Muscle-related coiled-coil protein; AltName: Full=Muscle-restricted

coiled-coil protein ACA62935.1 muscle-restricted coiled-coil protein [Homo sapiens]

Influenza Virus NKLFEKTKK

Human NKLFEKTRK

helix aa: 541-553 contiguous length = 5, contiguous length = 7

>NP_000584.2 antigen peptide transporter 1 isoform 1 [Homo sapiens]

Q03518.2 RecName: Full=Antigen peptide transporter 1; Short=APT1; AltName:

Full=ATP-binding cassette sub-family B member 2; AltName:

Full=Peptide supply factor 1; AltName: Full=Peptide transporter

PSF1; Short=PSF-1; AltName: Full=Peptide transporter

TAP1; AltName: Full=Peptide transporter involved in antigen

processing 1; AltName: Full=Really interesting new gene 4 protein

CAA40741.1 peptide transporter [Homo sapiens]

AAH14081.1 Transporter 1, ATP-binding cassette, sub-family B (MDR/TAP) [Homo sapiens]

EAX03647.1 transporter 1, ATP-binding cassette, sub-family B (MDR/TAP), isoform CRA_b [Homo sapiens]

ALQ33804.1 transporter 1 ATP-binding cassette sub-family B isoform 1, partial [Homo sapiens]

ARB08462.1 TAP1 [Homo sapiens]

ARB08463.1 TAP1 [Homo sapiens]

ARB08464.1 TAP1 [Homo sapiens]

ARB08465.1 TAP1 [Homo sapiens]

ARB08473.1 TAP1 [Homo sapiens]

ARB08474.1 TAP1 [Homo sapiens]

ARB08477.1 TAP1 [Homo sapiens]

ARB08481.1 TAP1 [Homo sapiens]

ARB08482.1 TAP1 [Homo sapiens]

ARB08484.1 TAP1 [Homo sapiens]

prf||1703418A RING4 gene

Influenza Virus LVLVV--SLG--AISF

Human LVLVVLSSLGEMAIPF

> AAH15195.1 Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide [Homo sapiens]

Influenza Virus LLCVALL

Human LLCVALL

extended aa:192-196 contiguous length = 5, contiguous length < 5

> AAI42999.1 Asparagine-linked glycosylation 11, alpha-1,2-mannosyltransferase homolog (yeast) [Homo sapiens]

Influenza Virus VLWGI

Human VLWGI

extended aa:246-251 contiguous length = 5, contiguous length = 5

> NP_036550.1 fascin-2 isoform 1 [Homo sapiens]

O14926.1 RecName: Full=Fascin-2; AltName: Full=Retinal fascin AAB86481.1 retinal fascin [Homo sapiens]

AAC18604.1 retinal fascin [Homo sapiens]

AAI26296.1 Fascin homolog 2, actin-bundling protein, retinal (Strongylocentrotus

purpuratus) [Homo sapiens]

AAI30331.1 Fascin homolog 2, actin-bundling protein, retinal (Strongylocentrotus purpuratus) [Homo sapiens]

Influenza Virus YWTLV

Human YWTLV

> CBX25812.1 cytochrome c oxidase subunit III, partial (mitochondrion) [Homo sapiens subsp. 'Denisova']

Influenza Virus ISIYW

Human ISIYW

coil aa:232-236 contiguous length = 5, contiguous length = 5

> NP_001034794.1 trophinin isoform 5 [Homo sapiens]

XP_006724663.1 trophinin isoform X1 [Homo sapiens]

XP_011529110.1 trophinin isoform X1 [Homo sapiens]

XP_011529111.1 trophinin isoform X1 [Homo sapiens]

XP_011529113.1 trophinin isoform X1 [Homo sapiens]

XP_011529114.1 trophinin isoform X1 [Homo sapiens]

XP_011529115.1 trophinin isoform X1 [Homo sapiens]

XP_016885256.1 trophinin isoform X1 [Homo sapiens]

Q12816.3 RecName: Full=Trophinin; AltName: Full=MAGE-D3 antigen

BAA83066.4 KIAA1114 protein [Homo sapiens]

Influenza Virus AIRPK

Human AIRPK

> XP_011525670.1 zinc finger protein 541 isoform X1 [Homo sapiens]

XP_011525671.1 zinc finger protein 541 isoform X1 [Homo sapiens]

XP_011525672.1 zinc finger protein 541 isoform X1 [Homo sapiens]

XP_011525673.1 zinc finger protein 541 isoform X1 [Homo sapiens]

Influenza Virus NIGSR

Human NIGSR

coil aa:305-309 contiguous length = 5, contiguous length = 5

> AAO15766.1 a disintegrin-like and metalloprotease with thrombospondin type

1 motifs 20 [Homo sapiens]

Influenza Virus TSLPF

Human TSLPF

> XP_016865452.1 G-protein coupled receptor 98 isoform X1 [Homo sapiens]

Influenza Virus PNDKP

Human PNDKP

coil aa:334-340 contiguous length = 6, contiguous length = 5

> EAW73283.1 BCL2-interacting killer (apoptosis-inducing), isoform CRA_b [Homo sapiens]

Influenza Virus LRNVPS

Human LRNVPS

>NP_078922.1 protein zyg-11 homolog B [Homo sapiens]

Q9C0D3.2 RecName: Full=Protein zyg-11 homolog B

BAB21821.2 KIAA1730 protein [Homo sapiens]

Influenza Virus TGMRN**V**P

Human TGMRN**H**P

Each of the 20 molecular probes of influenza HA protein origin detected matching identical subsequences of contiguous amino acids of human origin except **KLYIW**, a pentameric subsequence of contiguous amino acids found in H3N2 hemagglutinins (aa: 192-196) but which was not detected in human proteins. The largest detected human: influenza matching subsequence pair was **TIIALSYI** which is a contiguous octameric component of the amino acid 3-13 structural domain of the H3N2 hemagglutinin.

Discussion

This research is a continuation of the previous report of the occurrence of identical tetrapeptide subsequences in influenza hemagglutinins and in human proteins [4]. The present report extends the analysis to identical subsequences of at least pentapeptide length. The nineteen amino acid subsequences reported here as being identical in humans and influenza virus consist of amino acid polymers of contiguous amino acids between **5** and **8** in polymer length. The cut off for this study was increased to **5** amino acids because randomization of secondary structure has been reported to predominate in amino acid sequences below, but not above, pentameric length [9,10].

The methodology used is based upon the similar secondary structural features of H1N1 and H3N2 HA proteins (Figures 1 and 2). Structural features common to subtype H1N1 and H3N2 HA proteins were determined and were used to identify influenza virus HA subsequences for use as probes for subsequences in human proteins. Use of those influenza virus hemagglutinin subsequence probes enabled the detection of human subsequences identical those of the virus.

The occurrence of identical host-virus amino acid subsequences may reduce immunogenicity of the influenza virus in the infected host by favouring immunological tolerance to the organism [11]. Such tolerance would increase the susceptibility of the host to influenza infection and would reduce protective effectiveness of anti-influenza vaccines.

Conclusion

It is reported here that the influenza virus hemagglutinin and various proteins of human origin contain identical subsequences of amino acids. These observed subsequences are 5-8 amino acids in length and thus may have significant biological effects such as induction of immune tolerance to potentially protective epitopes of the influenza virus.

There is a need for an anti-influenza vaccine of greater efficacy. It is proposed that influenza hemagglutinin proteins and peptides lacking human subsequences should be prepared and tested for effective antiviral immunogenicity.

References

1. Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, et al. (2018) Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. Lancet 391: 1285-1300.

2. World Health Organization (2017) WHO Preferred Product Characteristics for Next-Generation Influenza Vaccines.

3. Doyle JD, Chung JR, Kim SS, Gaglani M, Raiyani C, et al. (2019) Interim estimates of 2018-19 seasonal influenza vaccine effectiveness - United States, February 2019. MMWR Morb Mortal Wkly Rep 68: 135-139.

4. Weltman JK (2018) Shannon entropy screening of influenza hemagglutinin for tetrapeptides with exact homology to human proteins. J Med Microb Diagn 7: 284-286.

5. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2- a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

6. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27: 379-423.

7. Källberg M, Wang H, Wang S, Peng J, Wang Z, et al. (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7: 1511-1522.

8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

9. Figureau A, Soto MA, Toha J (2003) A pentapeptide-based method for protein secondary structure prediction. Protein Eng 16: 103-107.

10. Turjanski P, Ferreiro DU (2018) On the natural structure of amino acid patterns in families of protein sequences.

11. Chirino AJ, Ary ML, Marshall SA (2004) Minimizing the immunogenicity of protein therapeutics. Drug Discov Today 9: 82-90.