Journal of
Medical Microbiology & Diagnosis

Weltman, J Med Microb Diagn 2018, 7:3
DOI: 10.4172/2161-0703.1000284

Short Communication Open Access

Shannon Entropy Screening of Influenza Hemagglutinin for Tetrapeptides with Exact Homology to Human Proteins

Joel Kenneth Weltman*

Warren Alpert Medical School, Brown University, USA

Abstract

Based upon a unique pair of non-mutating contiguous amino acids in the HA2 region of influenza H1N1 hemagglutinin, identical tetrapeptides were identified in the influenza hemagglutinin and in proteins of human origin. It is hypothesized that such peptide domains, present in both host and virus, increase the adaptability of the virus to the host.

Keywords: Influenza virus; Hemagglutinin; Shannon entropy; Human protein; Tetrapeptide; Homology

Introduction

Influenza virus remains a significant public health problem [1]. Understanding the biology of influenza virus may facilitate the design of new anti-viral therapeutic and preventive agents and strategies [2]. The present report is based upon an analysis of Shannon entropy and secondary protein structure in the hemagglutinin protein of H1N1 influenza virus. The work focuses on positions of zero Shannon entropy.

Materials and Methods

Sequences of Influenza virus H1N1 HA protein were downloaded from the Influenza Virus Resource (https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database) on 2 Aug 2018 [3]. Of a total 16678 HA protein sequences, 1915 sequences were of length between 560 and 565 amino acids, 30 sequences were of length between 567 and 575 amino acids and 14733 HA protein sequences were of length 566 amino acids. The largest subset, consisting of the HA sequences of length 566 amino acids, was used for this study.

Computations were performed with Anaconda Python 2.7.14. Information entropy (H) was computed by the method of Shannon and is reported in bits [4]. Protein secondary structure was computed on the RaptorX server [5]. Sequence management and calculation of consensus sequence were performed with the Jalview application [6]. The domains of the HA protein were assigned as signal sequence (Positions 1-17), HA1 (Positions 18-344) and HA2 (positions 345-566) according to reference sequence Influenza A virus (A/Puerto Rico/8/1934(H1N1)) segment 4, complete sequence NP_040980.1. The Mann-Whitney U test and the Z-test with 1000 pseudorandom trials were performed with Scipy [7].

Protein-protein searches for human protein sequences (Homo sapiens, taxid 9606) were performed on the National Library of Medicine-National Center for BioInformatics website (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) using BLASTP, with the Blosum 62 substitution matrix and the NLM-NCBI Reference Proteins (refseq_protein) database; organism was set to Homo sapiens (taxid: 9606). Searches of the refseq_protein database of human-origin used two tetrapeptides as query peptide sequences: GLY TRP TYR GLY and GLY TRP PHE GLY. These two tetrapeptides were detected in H=0.0 distributions in the HA2 domain of H1N1 influenza hemagglutinin.

Results and Discussion

The distribution of H in the H1N1 HA protein with intact signal sequence is shown in the top graph of Figure 1. Maximum H is at position 391, with a value of 1.1462 bits. The total H of the dataset was 37.5704 bits. For the complete dataset, the mean H equaled 0.0664 bits, median H equaled 0.0099 bits and a standard deviation of the mean equaled 0.1508 bits. There were 532 positions at which H>0.0 and 34 positions at which H=0.0. Sorting the complete set of 566 positions into subsets, based on either H>0.0 or H=0.0, yielded two subsets with non-overlapping H values that significantly differed from each other statistically (Mann-Whitney U=0.0, $p=6.5946 \times 10^{-23}$).

The distributions of the subset of positions at which H=0.0 at each position is shown in Figure 1 (bottom three graphs) as functions of helix (h), extended strand (e) and random coil (c) secondary structures of the consensus HA protein. As shown in Figure 1, seven of the positions where H=0.0 occurred in helices, 12 occurred in extended strands and 15 occurred in random coils. The 34 positions at which H=0.0 and their secondary structures are listed in Table 1.

As shown in Table 1, non-mutating amino acids, i.e. with H=0.0, occurred in all three domains of the HA protein. Almost all the non-mutating amino acids were noncontiguous. There were two pairs of almost contiguous non-mutating amino acids in which the amino acids were separated by a single amino acid position at which H was greater than 0.0: (LEU67, LEU69) and (CYS107, PRO109). Each of the other amino acid positions at which H=0.0 was separated from other amino acid positions at which H=0.0 by more than a single position, with the following unique exception: positions GLY364 and TRP365. H equaled 0.0 at each of these two contiguous positions. Moreover, H also equaled 0.0 at neighboring, albeit non-contiguous GLY367. At the single intervening position 366 (H=0.0044 bits), two amino acids occurred: TYR366 (n=14728) and PHE366 (n=5). Thus, H equaled 0.0 at 3 of these 4 positions 364, 365, and 367 and H was greater than 0.0 bits at position 366. HA positions 364-367 were thus the site of the following two tetrapeptides:

***Corresponding author:** Joel Kenneth Weltman, Clinical Professor Emeritus of Medicine, Warren Alpert Medical School, Brown University, USA, Tel: 1-401-245-7588; E-mail: joel_weltman@brown.edu

Received August 31, 2018; **Accepted** September 14, 2018; **Published** September 19, 2018

Citation: Weltman JK (2018) Shannon Entropy Screening of Influenza Hemagglutinin for Tetrapeptides with Exact Homology to Human Proteins. J Med Microb Diagn 7: 284. doi:10.4172/2161-0703.1000284

Position Count	Amino Acid Position	Hemagglutinin Domain	Amino Acid	Secondary Structure
1	1	signal peptide	MET	Random coil
2	23	HA1	GLY	Extended strand
3	35	HA1	THR	Random coil
4	59	HA1	CYS	Extended strand
5	67	HA1	LEU	Extended strand
6	69	HA1	LEU	Random coil
7	75	HA1	ALA	Helix
8	93	HA1	TRP	Extended strand
9	107	HA1	CYS	Random coil
10	109	HA1	PRO	Random coil
11	140	HA1	TRP	Random coil
12	153	HA1	CYS	Random coil
13	191	HA1	LEU	Extended strand
14	243	HA1	ARG	Extended strand
15	322	HA1	LYS	Random coil
16	349	HA2	ALA	Helix
17	353	HA2	PHE	Helix
18	364	HA2	GLY	Random coil
19	365	HA2	TRP	Extended strand
20	367	HA2	GLY	Extended strand
21	375	HA2	GLY	Random coil
22	380	HA2	ALA	Random coil
23	386	HA2	GLN	Helix
24	424	HA2	LEU	Helix
25	427	HA2	LYS	Helix
26	434	HA2	ASP	Random coil
27	448	HA2	ASN	Random coil
28	476	HA2	GLU	Extended strand
29	481	HA2	CYS	Random coil
30	501	HA2	TYR	Random coil
31	512	HA2	LEU	Helix
32	531	HA2	LEU	Extended strand
33	534	HA2	TYR	Extended strand
34	560	HA2	LEU	Extended strand
The “Position Count” column shows the running total count of amino acid positions at which H=0.0, beginning with the N-terminal MET.				

Table 1: H1N1 Hemagglutinin (HA) amino acid positions at which H=0.0.

Protein of Human Origin (>NP_001258683.1 integrin beta-like protein 1 isoform 4 [Homo sapiens])	MCKNSQDIICSNAGTCHCGRCKCDNSDGSGLVYGKFCECDDRE CIDDETEEICGGHGKCYCGNCYCKAGWHGDKCEFQCDITPWES KRRCTSPDGKICSNRGTCVCGECTCHDVDPTGDWGDIHGDTCE CDERDCRAVYDRYSDDFCSGHGQCNCGRCDCKAGWYGKKCE HPQSCTLSAEESIRKCQGSSDLPCSGRGKCECGKCTCYPPGDRR VYGKTCECDDRRCEDLDGVVCGGHGTCSCGRCVCERGWFGK LCQHPRKCNMTEEQSKNLCESADGILCSGKGSCHCGKCICSAEE WYISGEFCDCDDRDCDKHDGLICTGNGICSCGNCECWDGWNG NACEIWLGSEYP

Table 2: A protein of human origin containing a GLY TRP TYR GLY Tetrapeptide (GWYG) and a GLY TRP PHE GLY Tetrapeptide (GWFG) selected by observed clustered amino acid positions where H=0.0 in influenza H1N1 HA2 protein.

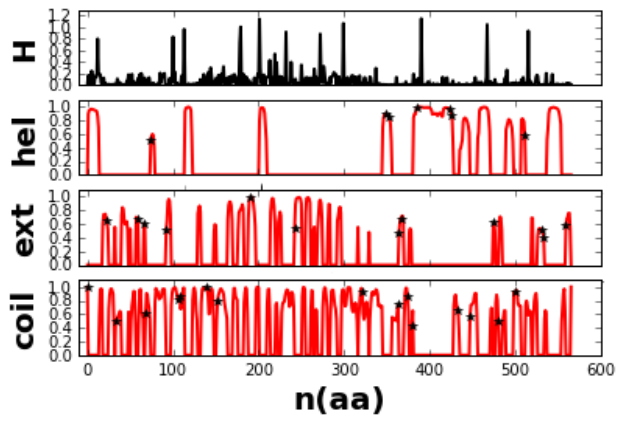

Figure 1: Zero information entropy (H=0.0) and secondary structure distributions in full-size (n=566) H1N1 hemagglutinin protein. Secondary structure is represented as helix (hel), extended strand (ext) and random coil (coil). Each position at which H=0 is represented by a star-shaped symbol (*) on the maximum secondary structure at that position.

(1) GLY TRP TYR GLY

(2) GLY TRP PHE GLY

The occurrence of these two influenza tetrapeptides as possible components of human proteins was next addressed. Screening human protein sequences for the presence of the two tetrapeptides detected in H1N1 influenza HA2 proteins yielded a total of 68 hits ($Z=8.0183$, $p=1.0722 \times 10^{-15}$). This total consisted of 57 occurrences of tetrapeptide_1 in the absence of tetrapeptide_2 ($Z=7.6666$, $p=1.7666 \times 10^{-14}$) and 11 hits for tetrapeptide 1 in the presence of tetrapeptide 2 ($Z=3.3351$, $p=0.0009$). These Z-test results indicate that each of the tetrapeptide counts was statistically greater than zero. Tetrapeptide_2 was not observed in human proteins in the absence of tetrapeptide_1.

An example (Integrin beta-like protein 1 isoform 4) of one of the human protein sequences detected by the presence of a tetrapeptide detected in influenza HA protein (tetrapeptide 2) is shown in Table 2. A complete list of reference human proteins detected because of the presence of either a tetrapeptide 1 or a tetrapeptide 2 component is given in Supplementary Information.

Conclusion

It is proposed that the tetrapeptides expressed both in human proteins and in influenza H1N1 HA are structural features that are associated with immunological or other disguising features of the virus in the human host, thereby permitting viral replication and function [8]. The effects of such peptides on HA-based vaccines and treatments should be determined.

Acknowledgment

The author thanks the Brown University Center for Computation and Visualization (CCV) for providing resources and services that facilitated this research.

References

1. Grohskopf LA, Sokolow LZ, Broder KR, Walter EB, Bresee JS (2017) Prevention and control of seasonal influenza with vaccines: Recommendations of the advisory committee on immunization practices-United States, 2017-18 influenza season. MMWR Recomm Rep 66: 1-20.
2. Bouvier NM, Palese P (2008) The biology of influenza viruses. Vaccine 4: D49-D53.
3. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. J Virol 82: 596-601.
4. Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379-423.
5. Källberg M, Wang H, Wang S, Peng J, Wang Z, et al. (2012) Template-based protein structure modeling using the Raptor-X web server. Nat Protoc 7: 1511-1522.
6. Waterhouse AM, Procter JB, Martin DMA (2009) Jalview version 2: A multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.
7. Mann HB, Whitney DR (1947) On a test of whether one of two random variables are stochastically larger than the other. Ann. Math. Stat 18: 50-60.
8. Pérez-Cañamás M, Hernández C (2015) Key importance of small RNA binding for the activity of a glycine-tryptophan (GW) motif-containing viral suppressor of RNA silencing. J Biol Chem 290: 3106-3120.