Journal of Medical Microbiology & Diagnosis

Weltman, J Med Microb Diagn 2016, 5:4
DOI: 10.4172/2161-0703.1000247

Short Communication OMICS International

Sets and Subsets of Mutating Amino Acids in Zika Virus Polyprotein

Joel K Weltman*

Clinical Professor Emeritus of Medicine, Alpert Medical School, Brown University, USA

***Corresponding author:** Joel K Weltman, Clinical Professor Emeritus of Medicine, Alpert Medical School, Brown University Providence, RI 02912, USA, Tel: 401-2457588; E mail: joel_weltman@brown.edu

Rec Date: Dec 01, 2016; **Acc Date:** Dec 28, 2016; **Pub Date:** Dec 30, 2016

Short Communication

Zika virus disease (ZVD) is the cause of microcephaly in newborns (http://www.cdc.gov/zika/hc-providers/infants-children/zika-microcephaly.html) and other diseases in infected children [1-3]. Insight into the bioinformatics and molecular biology of the Zika virus (ZIKV) may facilitate development of anti-viral vaccines and drugs [4].

In this communication the results of a study of Shannon information entropy (H) of ZIKV polyprotein sequences downloaded from the NCBI Zika Virus Resource (http://www.ncbi.nlm.nih.gov/genome/viruses/variation/Zika/) on 21 Nov 2016. The dataset download consisted of the complete set of ZIKV full-length polyprotein sequences isolated either from humans (n=123) or from *Aedes aegypti* mosquitos (n=14). Sequence management was facilitated with Jalview 2.9.0b2 [5]. H was computed by the equation of Shannon [6], using Anaconda 2.4.0 (64-bit), Python 2.7.10, Numpy 1.10.1, Scipy 0.16.0 and Matplotlib 1.4.3. The Mann-Whitney non-parametric U test was performed with Scipy stats; a two-tail p-value is reported. Modeling of the data by sets and subsets was performed with Maple 18 (Maplesoft Group, Canada). Z-tests were performed using 1000 pseudo-random trials and are reported with two-tail probabilities.

Curves of H distribution are shown in Figure 1 for polyprotein H of ZIKV isolated either from humans (Figure 1, upper left) or from *Aedes aegypti* mosquitos (Figure 1, lower left). There were 284 amino acid positions with H>0 in the polyprotein of human origin, but there were only 52 positions with H>0 in the polyprotein of *Aedes aegypti* origin (Z=12.8695, 6.6809×10^{-38}). Despite this significant difference in number of mutating amino acids in the two datasets, the observed difference between the total, summed H of the polyproteins of human origin (39.5530 bits) and that of the polyproteins of *Aedes aegypti* mosquito origin (35.2936 bits) was not significant (Z=0.4875, p=0.6259). However, differences between the H distributions in polyproteins of human origin and those of Aedes origin were clearly detected by the Mann-Whitney nonparametric U test (U=5468496.5, $p=2.3924 \times 10^{-09}$).

As predicted from the differences in total number of amino acids at which H>0.0, discussed above, no significant correlation was observed between the overall distributions of H in the two polyproteins (Figure 1, upper right), whether measured parametrically (Pearson r=0.4743, 1.3699×10^{-191}) or non-parametrically (Spearman r=0.3470, $p=1.9517 \times 10^{-97}$). As shown next, these seemingly uncorrelated datasets were nevertheless useful for sorting the H distributions into meaningful subsets. It is seen in the figure that H values for polyproteins of human origin consist of vertically distributed H values greater than zero on the ordinate, but at zero on the abscissa (subset A).

There are analogous H value positions for the polyproteins of Aedes origin, but with a horizontal distribution on the abscissa and at zero on the ordinate (subset B). In addition to subsets A and B, there are H values that are distributed neither strictly vertically nor strictly horizontally; these H values are distributed on the surface between the ordinate and the abscissa (subset C). These subsets, all with positions at which H>0.0, are shown in Figure 1, lower right. There are 241 amino acid positions in the subset with exclusively human elements (subset A), 9 amino acid positions in the subset with exclusively *Aedes aegypti* elements (subset B) and 43 amino acid positions that are common to ZiKV polyproteins both of human origin and of *Aedes aegypti* origin (subset C). These amino acid position count subsets significantly differ from each other: subset A versus subset B, Z=14.6628, $p=1.1155 \times 10^{-48}$; subset A versus subset C, Z=12.0663, $p=1.5922 \times 10^{-33}$; subset B versus subset C, Z=4.7426, $p=2.1094 \times 10^{-06}$.

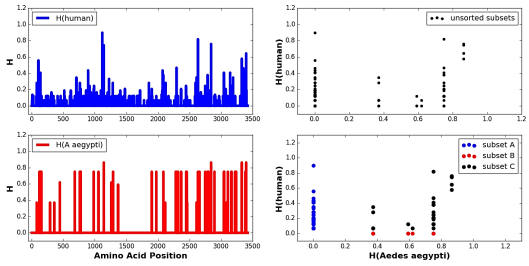

Figure 1: Information entropy (H) in ZIKV Polyproteins Isolated either from Humans or from *Aedes aegypti* Mosquitos. (upper left) H distribution in ZIKV polyproteins of human origin; (lower left) H distribution in ZIKV polyproteins of *Aedes aegypti* mosquito origin; (upper right); scatter plot of H of human-origin polyproteins as a function of H of polyproteins of *Aedes aegypti* origin; (lower right) subsets of amino acid positions where H>0.0 in human-origin and in Aedes-origin ZIKV polyproteins. The amino acid positions are sorted into subset A (exclusive human origin), subset B (exclusive *Aedes aegypti* origin) and subset C (common human and *Aedes aegypti* origin).

The distributions of information entropic ZIKV polyprotein amino acid positions described above can be expressed as sorting and partitioning of discrete sets and subsets of amino acid positions, i.e. indices. The set of indices of positions with H>0 in polyproteins isolated from humans is described by the union of subsets A and C:

$$human = A \cup C \quad [1]$$

The analogous set of indices of positions with H>0 in polyprotein isolated from *Aedes aegypti* mosquitos is the union of B and C:

$$mosquito = B \cup C \quad [2]$$

Thus, the C subset can be described as the intersection of indices of H values comprising the complete sets of ZIKV polyprotein of human and mosquito origin:

$$human \cap mosquito = ((A \cup C) \cap (B \cup C)) \text{ [3]}$$

For computational analysis, sets A, B and C are placed in a superset, k:

$$k = \{A, B, C\} \text{ [4]}$$

Superset k then provides p values for computation of H at the indexed polyprotein positions specific for ZIKV isolated from humans (subset A), specific for ZIKV isolated from *Aedes aegypti* mosquitos (subset B) and common to ZIKV isolated both from humans and from *Aedes aegypti* mosquitos (subset C):

$$H = k \rightarrow \sum_{i=1}^{20} p(k,i) * \log 2(p(k,i)) \text{ [5]}$$

where, i represents amino acids occupying a specific position in the ZIKV polyprotein chain. Equations 1-5 thus provide a concise logical framework for the mathematical analysis of H distributions in the polyproteins of ZIKV isolated either from a human source or from *Aedes aegypti* mosquitos.

It is proposed that the presented analysis shows that:

- The A subset of ZIKV amino acid positions represents biological forces and constraints expressed exclusively by infected human host cells but not by infected cells of the vector *Aedes aegypti* mosquitos;
- The B subset of ZIKV amino acid positions represents biological forces and constraints expressed exclusively by infected cells of the vector *Aedes aegypti* mosquitos but not by infected human host cells;
- The C subset of ZIKV amino acid positions represents biological forces and constraints expressed both by infected human host cells and by infected cells of the vector *Aedes aegypti* mosquitos.

It is recognized that these results, although statistically significant, may require modification as the ZIKV dataset is enlarged over time.

Acknowledgement

I thank the Brown University Center for Computation and Visualization (CCV) for providing access to Maple 18.

References

1. Rasmussen SA, Jamieson DJ, Honein MA, Petersen LR (2016) Zika Virus and birth defects-reviewing the evidence for causality. N Engl J Med 374: 1981-1987.
2. Dolma K (2016) Zika virus (ZIKV) Infection: A Review. J Res Development 4: 148.
3. Karwowski MP, Nelson JM, Staples JE, Fischer M, Fleming-Dutra KE, et al. (2016) Zika virus disease: A CDC update for pediatric health care providers. Pediatrics 137: e20160621.
4. Weltman JK (2016) Computer-assisted vaccine design by analysis of Zika virus e proteins obtained either from humans or from Aedes Mosquitos. J Med Microb Diagn 5: 235.
5. Waterhouse AM, Procter JB, Martin DM, 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.