OPEN ACCESS

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

www.sciforum.net/conference/ecea-2

*Conference Proceedings Paper – Entropy*

# Mutual Information-Based Cliques of Amino Acids in the Zaire Ebola Virus-Makona Glycoprotein

**Joel K. Weltman**

Clinical Professor Emeritus of Medicine, Alpert Medical School, Brown University, Providence, RI 02912, USA; E-Mail: joel_weltman@brown.edu

*Published: 13 November 2015*

---

**Abstract:** Zaire Ebola virus Makona variant (**ZEBOV-Makona**) is the cause of the 2014-2015 high-mortality epidemic of Ebola virus disease (**EVD**). The viral glycoprotein (**GP**) is the component that mediates binding and internalization of the virus. By means of the coordinated determination of information entropy (**H**), mutual information (**MI**) and protein secondary structure (**PsiPred Score**), three non-interconnected **MI** cliques of amino acids were detected within the **ZEBOV-Makona GP** isolated from humans. Three amino positions (82, 230 and 371) were identified with **H** values significantly greater than those of all the other 673 **GP** amino acids in the **GP** molecule. These three amino acid positions formed **MI** network cliques with additional sets of four, nine and five amino acid positions, respectively. Each **MI** clique was complete but no inter-clique **MI** connections were detected. Each wild type amino acid member of a **MI** clique was of one of the essential amino acids **VAL**, **THR**, **ILE**. Since essential amino acids are not synthesized by humans, use of these essential amino acids by the **MI** cliques may be indicative of human host factors, e.g., diet and nutritional state, that influence the occurrence and survival of **ZEBOV-Makona** mutations.

**Keywords:** Ebola virus; ZEBOV-Makona variant; information entropy; mutual information; glycoprotein; GP; protein secondary structure; networks; cliques; EVD

**PACS Codes:** 87.19.xd; 89.70.Cf; 87.19.lo; 89.75.Hc; 87.14.E-

---

## 1. Introduction

The 2114-2115 epidemic of Ebola virus disease (**EVD**) is the largest and most severe that has been recorded. As of this writing, there have been a greater number of **EVD** cases and deaths in this

epidemic than all of the cases and all of the deaths summed since 1976, the year of the discovery of Ebola virus [1]. The 2014-2015 epidemic of **EVD** is caused by the Zaire Ebola virus Makona variant (**ZEBOV-Makona**) [2]. The glycoprotein (**GP1,2**) of Ebola virus mediates the binding of the virus to the target cell membrane and the subsequent internalization of the virus into the cell [3, 4].

Reported here are distributions of Shannon information entropy (**H**) and mutual information (**MI**) in **ZEBOV-Makona GP1,2** [5, 6]. The **H** and **MI** distributions are reported for **GP1,2** amino acid subsets selected by additional, concomitant application of secondary structure parameters [7] which facilitated the sorting process. Finally, the detected subsets were treated as network cliques [8] in order to gain insight into the functional organization of **ZEBOV-Makona GP1,2** on a bioinformatic level.

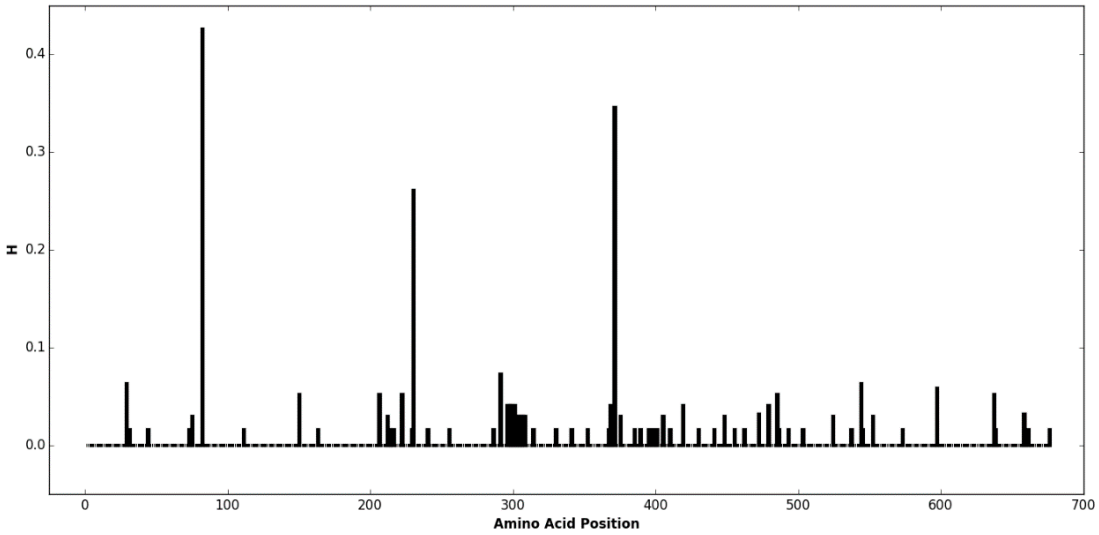
## 2. Results and Discussion

The distribution of **H** and the relations of protein secondary structure to that distribution were first determined for **ZEBOV-Makona GP1,2**. The combinations of these methods of analysis, which were found to be useful for studying those relations of **H,** were then subsequently applied to **MI**.

### *2.1. **H** Distribution in the Ebola Glycoprotein*

**H** values computed for **ZEBOV-Makona GP1,2** are shown in Figure 1; the **H** values are presented according to amino acid position. Three amino acid positions (positions 82, 230 and 371) had **H** values (0.4260, 0.2612 and 0.3462 bits) which were greater than those of all of the other 72 non-zero **H** values. For those 72 amino acid positions with **H**>0, median=0.0160, mean=0.0271, std=0.0145, min=0.0160 and max=0.0730. The amino acid positions 82, 230 and 371 thus formed a subset consisting of three amino acid positions with **H** values significantly greater than those of all 72 of the other non-zero positions (t=27.1845, p(t)=6.3278e-40). Inclusion of the 601 amino acid positions at which **H**=0.0 further increased the significance to t=55.6453, p(t)=3.5149e-254.

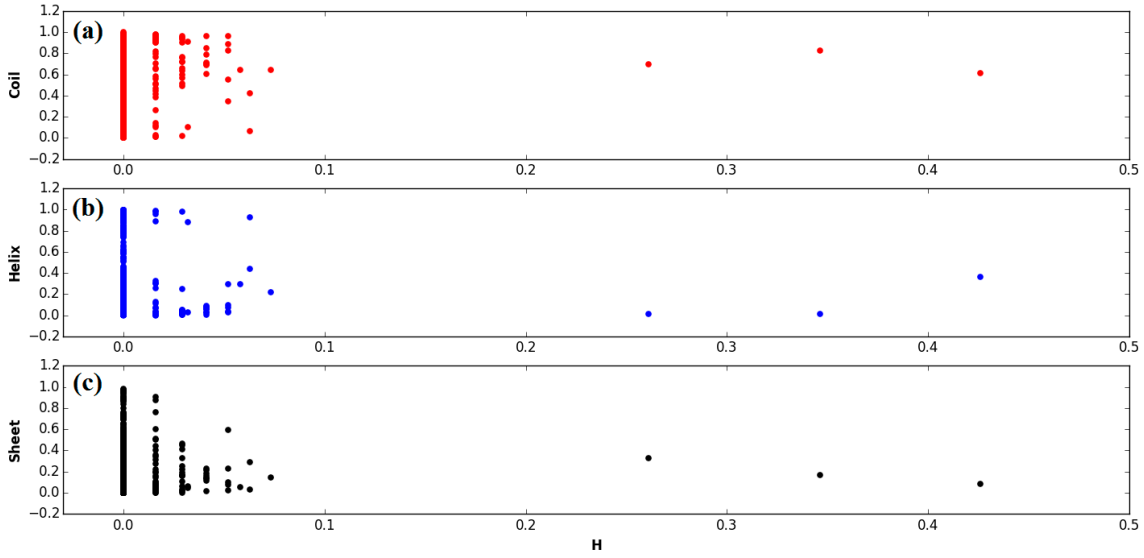
**Figure 1. Information Entropy (H) Distribution in ZEBOV Makona Variant Virus Glycoprotein.** x=amino acid position, y=information entropy (**H**), in bits.



*2.2. **H** and the Secondary Structure of the Ebola Glycoprotein*

The clustering of H values into subsets was made more apparent by introduction of structural parameters into the analysis. Clustering of the three greatest **H** values positions was evident in plots for random coil (Figure 2a), helix (Figure 2b) and extended sheet (Figure 2c) secondary structures as functions of **H**. The clustering of these three **H** values is evident in all three plots in Figure 2. However, the **PsiPred scores** for random coil were greater than the corresponding scores for helix (t=4.4105, p(t)=0.0116), for extended sheet (t=5.4759, p(t)= 0.0054) and for the combined (helix, extended sheet) datasets((t=5.5026, p(t)=0.0009). Thus, considering **PsiPred score** as a function of **H** achieved two goals: **(1)** the set of data for the 676 GP1,2 amino acid positions was conveniently sorted into three statistically distinguishable subsets (**H=0.0**, **H>0.0** and **H>>0.0**); **(2)** the predicted secondary structure of all members of a particular subset could readily be determined. The **PsiPred score** was next similarly applied to analysis of **MI** in the **GP1,2** protein molecule.

**Figure 2. Structurally-Assisted Sorting of ZEBOV-Makona Glycoprotein Information Entropy (H). (a)** y = random coil PsiPred score (red**). (b)** y= helix Psipred score (blue). **(c)** y= extended sheet PsiPred score (black). x=information entropy (**H**), in bits.

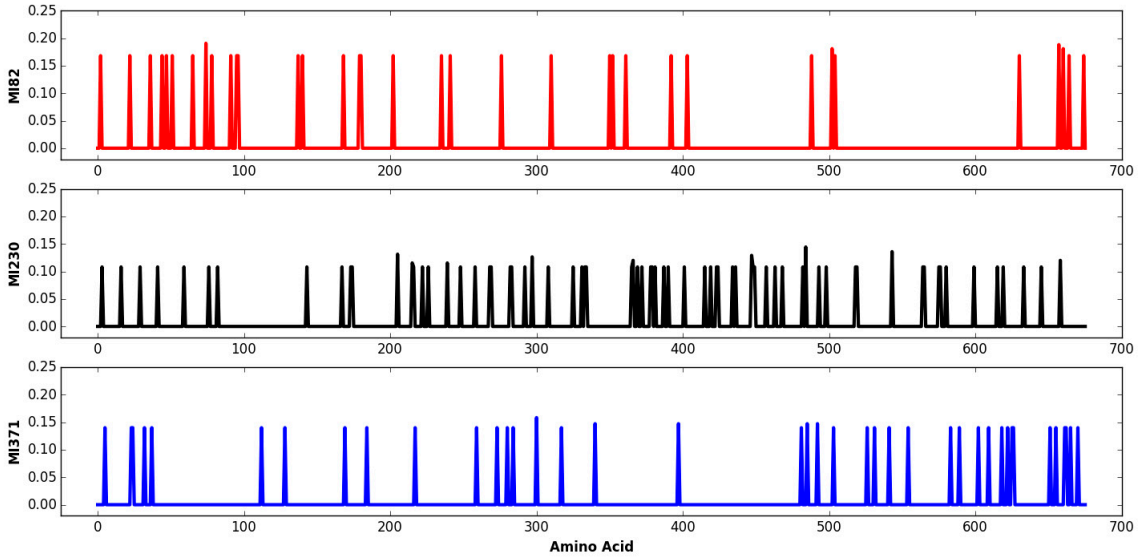



*2.3. Three **MI** Subsets of Amino Acids in the Ebola Glycoprotein*

The **MI** of all 676 amino acid positions of **GP1,2** was determined with each of the three high **H** positions discussed in the previous paragraph as reference position. The **MI** distribution with position 82 as reference position (**MI82**) is shown in Figure 3a, with position 230 as reference position (**MI230**) in Figure 3b and with position 371 as reference position (**MI371**) in Figure 3c. The **MI82** dataset contained 35 positions with **MI>0.0**, the **MI230** dataset contained 70 positions with **MI>0.0** and the **MI371** dataset contained 40 positions with **MI>0.0**. However, it was found that **MI=0.0** between each of the three possible pairs of observed maximum **H** reference positions, ie, (82, 230), (82, 371) and (230, 371). Furthermore, there was no significant correlation between the distributions of **MI** in any of the pairs of **MI** subsets, with near-zero values of the Pearson parametric (**r**) and Spearman non-parametric (**ρ**) correlation coefficients, ranging from only -0.0852 to -0.0585 with

associated probabilities ranging from 0.0268 to 0.1281. Structurally-assisted sorting was next applied to each of the three sets of **MI** data shown in Figure 3 as had been done above for **H** (Figure 2).

**Figure 3. Mutual Information (MI) Distributions in ZEBOV-Makona Glycoprotein Data Sets. (top) MI** distribution with reference amino acid 82 (**MI82**, red); **(middle) MI** distribution with reference amino acid 230 (**MI230**, black) **(bottom) MI** distribution with reference amino acid 371 (**MI371**, blue). **x**=amino acid position; **y** = **MI** (bits).

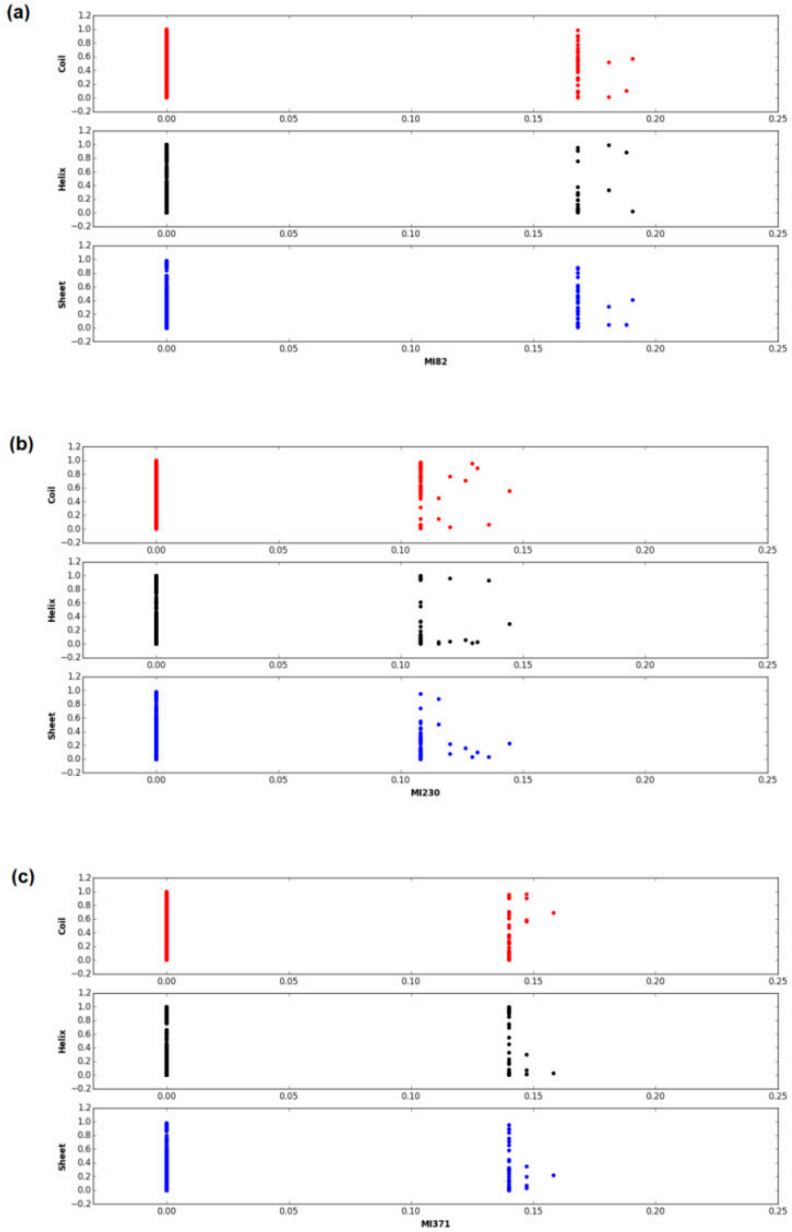



*2.4. Structurally-Assisted Sorting **MI** Subsets of Amino Acids in the Ebola Glycoprotein*

Plots of **PsiPred** secondary structure scores of **GP1,2** are shown as functions of **MI** for the **MI82** dataset (Figure 4a), the **MI230** dataset (Figure 4b) and for the **MI371** dataset (Figure 4c). Using **PsiPred** secondary structure scores as independent variables, the **MI82** set was sorted into a relatively high **MI** subset (n=4) and a low **MI** subset (n=31), where the difference between the higher and lower **MI** values was statistically significant (**t**= 21.0609, **p(t)** =1.0485e-20). Similarly, the **MI230** dataset was sorted into a relatively high **MI** subset (n=9) and a low **MI** subset (n=61) where the difference between **MI** low and **MI** high was again statistically significant (**t**=15.2853, **p(t)** =1.0978e-23); finally, the **MI371** set was sorted into a relatively high **MI** subset (n=5) and a low **MI** subset (n=35) where the difference between high and low **MI** values was again statistically significant (**t**= 12.3762, **p(t)** =6.6613e-15).

The secondary structures of the amino acids in the high **MI** subsets were obtained directly from the **PsiPRED** output. In the **MI82** subset, there were two amino acid positions with random coil structure and two with helical structure. In the **MI230** subset, there were five amino acid positions with random coil structure, two with helical structure and two positions with extended sheet secondary structure. All of the five amino acid positions with relatively high **MI** in the **MI371** subset had random coil structure.

**Figure 4. Structurally-Assisted Sorting of ZEBOV-Makona Glycoprotein Mutual Information (MI) Subsets.** (**a**) amino acid 82 in reference position (**MI82**); **(b)** amino acid 230 in reference position (**MI230**). **(c)** amino acid 371 in reference position (**MI371**). In **a**, **b** and **c**: x= **MI** (bits); y=**PsiPRED** random coil score (top, red), helix score (middle, black) and y=extended sheet score (bottom, blue).

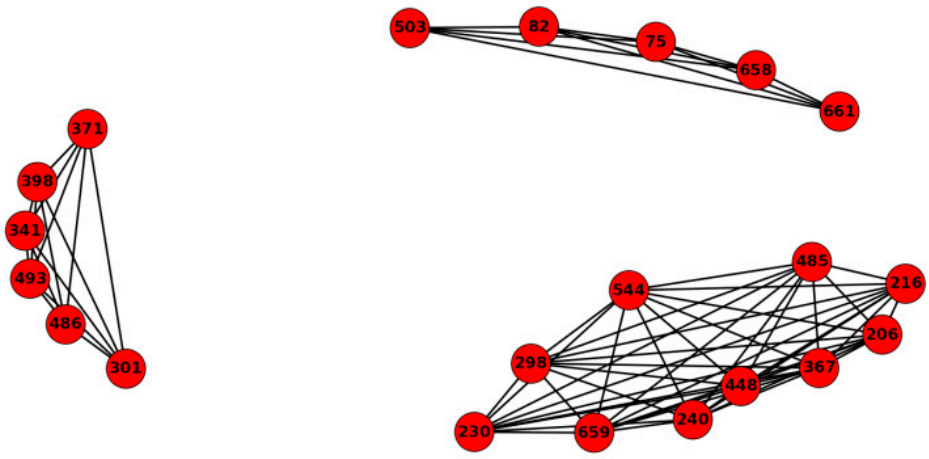



*2.5. Network Analysis of the Ebola Virus Glycoprotein **MI** Subsets of Amino Acids*

Graph results of network analysis of the combined data for the 21 high **MI** positions in the **MI82**, **MI230** and **MI371** subsets are shown below in Figure 5. The network is presented as an undirected, weighted graph, where weight was set equal to 0.176/exp($\mathbf{MI_{i,j}}$ ) for visualization purposes, and where $\mathbf{MI_{i,j}}$ is the observed mutual information connection between the i-th and the j-th amino acids. The three **MI** subsets formed three disjoint cliques, each of which was complete, ie, completely connected

within the clique. It should be noted that the nodes of each network clique are the nodes of the corresponding **MI** subset plus the single, corresponding reference node (82, 230 and 371). Degree centrality was 0.30 for each of the five members of clique **MI82**, 0.55 for each of the 10 members of clique **MI230** and 0.35 for each of the six members of clique **MI371**. These degree centrality values support the interpretation of the cliques as disjoint and complete. For the 21 amino acid positions in the high **MI** subsets, there were 21x21=441 **MI** determinations; **MI>0.0** for 161 of these 441 determinations. For all 161 of the determinations where **MI>0.0, clique(i)$_i$ = clique(j)**. For the remaining 280 determinations, where **MI = 0.0, clique(i) ≠ clique(j)**. There were five members of the **MI82** clique, 10 members of the **MI230** clique and six members of the **MI371** clique; $5^2 + 10^2 + 6^2 =$ 161 positions, so that all observed, positive **MI** values are completely accounted for by intra-clique interactions.

**Figure 5.** Disjoint Subsets of Amino Acids Forming Complete Cliques within ZEBOV-Makona Glycoprotein. Clique MI82 (top), clique MI230 (bottom right) and clique MI371 (bottom left).




*2.6. Metadata for the Ebola Glycoprotein Cliques*

Metadata that characterize the amino acids in each of the three **MI** cliques are given in Table 1. The secondary structure of each of the 21 wild type amino acids is given. There were 15 random coil positions, four helices and two extended sheets. The three wild type amino acids were encoded by nine different codons. Two of the amino acids reside in the receptor binding domain [9, 10], six in the glycan cap [11, 12], eight in the mucin-like domain [12], two in the fusion loop [10, 13] and three in the hydrophobic tail. There was no codon usage, protein structure or functional domain that was exclusively associated with one of the three amino acid cliques. However, an amino acid, either **VAL**, **THR** or **ILE**, characterized the wild type member of each clique. Each of these three amino acids is an essential amino acid, ie, is not synthesized by the human host (https://en.wikipedia.org/wiki/Essential_amino_acid). Essential amino acids must be ingested in order for normal protein metabolism to be maintained. As shown in Table 1, not only the three central,

reference positions were essential amino acids, but all 21 of the wild type amino acids in the **MI** network cliques were essential amino acids. It should be noted, however, that as shown in Table 1, amino acid essentiality was not uniformly maintained by amino acids mutating away from the wild type.

**Table 1.** Metadata for Cliques of Amino Acid Mutual Information (MI) Observed in ZEBOV-Makona Glycoprotein

| Network Clique | Amino Acids | Codons | Protein Secondary Structure | GP1,2 Domain |
|---|---|---|---|---|
| MI82 | V75A | GTG→GCG | Coil | RBD |
| MI82 | V82A | GTG→GCG | Coil | RBD |
| MI82 | V503A | GTA→GCA | Coil | FL |
| MI82 | V658A<br>V658I | GTT→GCC<br>GTT→ATT | Helix | GP2 Tail |
| MI82 | V661A | GTT→GCT | Helix | GP2 Tail |
| MI230 | T206M | ACG→ATG | Coil | GC |
| MI230 | T216P | ACC→CCC | Extended | GC |
| MI230 | T230A | ACA→GCA | Coil | GC |
| MI230 | T240N | ACC→AAC | Extended | GC |
| MI230 | T298L | ACT→CTA | Coil | GC |
| MI230 | T367A | ACC→GCC | Coil | MLD |
| MI230 | T448A | ACC→GCC | Coil | MLD |
| MI230 | T485A | ACT→GCT | Coil | MLD |
| MI230 | T544I | ACA→ATA | Helix | FL |
| MI230 | T659A | ACA→GCA | Helix | GP2 Tail |
| MI371 | I301F | ATT→TTC | Coil | GC |
| MI371 | I341N | ATC→AAC | Coil | MLD |
| MI371 | I371V | ATC→GTC | Coil | MLD |
| MI371 | I398T | ATC→ACC | Coil | MLD |
| MI371 | I486T | ATT→ACT | Coil | MLD |
| MI371 | I493T | ATC→ACC | Coil | MLD |

**RBD**=receptor binding domain; **FL**=fusion loop; **GC**=glycan cap; **MLD**=mucin-like domain. Helices are indicated by the color red, extended sheets by the color yellow.

## 3. Experimental Section

A complete set of full-length Zaire Ebola virus **ZEBOV-Makona** variant **GP1,2** nucleotide sequences (N = 729) was downloaded in FASTA format on June 18, 2015 using the NCBI Ebolavirus Resource (http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ebola/). The downloaded **GP1,2** nucleotide gene sequences were translated into 679 amino acid sequences with Biopython 1.65, using the IUPAC unambiguous DNA code. Each **GP1,2** sequence was of length 676 amino acids and without error characters. Information entropy (**H**) was calculated with the equation of Shannon [5]. Mutual information (**MI**) was computed as $\mathbf{MI=H_i + H_j - H_{i,j}}$.[6] Computations and graphing were performed with 64-bit **Enthought Canopy** 1.5.1, **Python** 2.7.6, **Numpy** 1.9.2-1, **Scipy** 0.15.1-2 and

**matplotlib** 1.4.2-2 Network analysis was performed with **Networkx** 1.9.1-3 [8]. The consensus sequence of the **GP1,2** dataset was determined with **Jalview** (2.8.2) [14]. **PsiPred scores** for the **GP1,2** consensus sequence random coil, helix and extended sheet protein secondary structures were obtained with the **PSIPRED** Protein Structure Prediction Server [7, 15].

## 4. Conclusions

Information entropy (**H**) and mutual information (**MI**) distributions in the **ZEBOV-Makona** glycoprotein (**GP**), isolated from humans, were sorted by protein secondary structure (random coil, helix and extended sheet). Inclusion of protein secondary structure in the analytical process significantly simplified detection and identification of groups of **H** and **MI** statistical outliers, without the necessity of prior statistical assumptions. These statistical outliers formed three complete but disjoint **MI** cliques of **GP** amino acids. Each clique was characterized by wild type use of an essential amino acid. Since essential amino acids are not synthesized by the body but must be ingested, the observed **ZEBOV-Makona** glycoprotein **H** and **MI** distributions may reflect human host nutritional and dietary factors.

## Author Contributions

Joel K Weltman is the sole author of this publication.

## Conflicts of Interest

The author declares no conflicts of interest.

## References and Notes


1. WHO Ebola Virus Disease Fact Sheet, N°103, http://www.who.int/mediacentre/factsheets/fs103/en/
2. Kuhn JH, Andersen KG, Baize S et al, Nomenclature and Database-Compatible Names for the Two Ebola Virus Variants that Emerged in Guinea and the Democratic Republic of the Congo in 2014. *Viruses* **2014**, 6, 4760-4799; doi:10.3390/v6114760
3. Lee JE and Saphire EO, Ebolavirus glycoprotein structure and mechanism of entry. *Future Virol*. **2009**,(6):621-635.
4. Miller EH, Chandran K. Filovirus entry into cells - new insights. *Curr Opin Virol*. **2012**, 2:206-214. doi: 10.1016/j.coviro.2012.02.015
5. Shannon, Claude E. A Mathematical Theory of Communication.. *Bell System Technical Journal* 1948, 27: 379–423.
6. Cover TM and Thomas JA. Entropy, Relative Entropy and Mutual Information. In Elements of Information Theory, 2nd ed; Wiley, USA, 2006**,** 19-25**.**
7. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **1999**, 292: 195-202.

8. Hagberg AA, Schult DA and PJ Swart (**2008**) Exploring Network Structure, Dynamics, and Function using NetworkX; Proceedings of the 7th Python in Science Conference (SciPy 2008) pp 11-16. G Varoquaux, T Vaught, and Jd Millman (Eds), (Pasadena, CA USA)
9. Dube D, Brecher MB, Delos SE, Rose SC, Park EW, Schornberg KL, Kuhn JH, White JM () The Primed Ebolavirus Glycoprotein (19-Kilodalton GP1,2): Sequence and Residues Critical for Host Cell Binding; J Virol **2009**, 83:2883–2891; doi:10.1128/JVI.01956-08
10. Lee JE, Fusco ML, Hessell AJ, Oswald WB, Burton DR & Ollmann Saphire E. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. Nature **2008**, 454:177-182; doi:10.1038/nature07082
11. Lee JE and Ollmann Saphire E. Ebolavirus glycoprotein structure and mechanism of entry; Future Virol. **2009**; 4(6): 621–635. doi: 10.2217/fvl.09.56
12. Lennemann NJ, Rhein BA, Ndungo E, Chandran K, Qiu X, Maury W. 2009 Comprehensive functional analysis of N-linked glycans on Ebola virus GP1. mBio **2014**, 5(1):e00862-13. doi:10.1128/mBio.00862-13.
13. Gregory SM, Harada E, Liang B, Delos SE, White JM, Tamm LK. Structure and function of the complete internal fusion loop from Ebolavirus glycoprotein 2. Proc Natl Acad Sci U S A. **2011**, 108(27):11211-11216. doi: 10.1073/pnas.1104760108
14. Waterhouse AM, Procter JB, Martin DM, et al, Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **2009,** 25:1189-1191. doi: 10.1093/bioinformatics/ btp033.
15. Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT. **2013**, Scalable web services for the PSIPRED Protein Analysis Workbench, *Nucleic Acids Res* 41 (W1): W340-W348