Information Entropy and Protein Secondary Structure in the ZEBOV-Makona Ebola Virus Glycoprotein

Joel K Weltman*
Alpert Medical School, Brown University, Providence, RI 02912 USA

Abstract

The current epidemic of Ebola virus disease (EVD) is caused by Zaire Ebola virus-Makona variant. Results are presented indicating that 88% of the information entropy (H) in the ZEBOV-Makona glycoprotein (GP1,2) was distributed to amino acids residing in random coil structures. In contrast, only 12% of the total H was due to mutations of amino acids in helical and extended sheet secondary structures. It is proposed that some of the H in random coils may represent mutational escape from host defense while the paucity of H in helical and extended sheet structures may reflect conformational constraints on mutation. By relating GP1,2 secondary structure and H in regions of GP1,2, this research helps to computationally identify potential targets for the design of anti-Ebola vaccines and drugs.

Keywords: Zaire Ebola virus; ZEBOV-Makona variant; Information entropy; Glycoprotein; GP; GP1,2; Protein secondary structure; EVD; Vaccines

Introduction

The high fatality of the 2014-2015 epidemic of Ebola virus disease (EVD) has been caused by the Zaire Ebola virus-Makona variant (ZEBOV-Makona) [1]. The research reported here focuses on distributions of information entropy (H) [2] and secondary structure [3] in the ZEBOV-Makona glycoprotein (GP1,2). The GP1,2 is the viral protein that mediates the binding and internalization of the virus by the target cell [4,5].

The high mortality, morbidity and public-health significance of EVD stresses the importance of developing effective vaccines, treatments and point-of-care diagnostics for this disease [6]. Results are presented that indicate most of the total H distribution in the GP1,2 protein is statistically accounted for by amino acids participating in random coil secondary structures. Knowledge of the H distribution in ZEBOV-Makona GP1,2, and knowledge of the structural features of that distribution, increase our insight into the functional biology of this virus. By helping to identify appropriate potential targets, this insight can facilitate the design of preventive and therapeutic anti-Ebola vaccines and drugs.

Materials and Methods

The complete set (N = 877) of Zaire Ebola virus ZEBOV-Makona variant GP1,2 full-length nucleotide sequences (length = 2031 nucleotides) was downloaded in FASTA format [7] on September 3, 2015 using the NCBI Ebola virus Resource (http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ebola/). The downloaded set of 877 GP1,2 nucleotide gene sequences was translated into amino acids with Biopython 1.65, using the IUPAC unambiguous DNA code. Eight hundred and fifteen (815) of these translated GP1,2 protein sequences were of full length 676 amino acids and without error characters. These 815 full-length, error-free GP1,2 protein sequences, comprising 92.93% of the initial download of nucleotide sequences, were utilized without further sequence re-alignment.

Information entropy (H) was calculated by the equation of Shannon [8]. Computation 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. Mann-Whitney U tests were computed with the Enthought Canopy 1.5.1, Python 2.7.6, Numpy 1.9.2-1, Scipy 0.15.1-2. Shannon [8]. Computation and graphing were performed with 64-bit Biopython 1.65, using the IUPAC unambiguous DNA code. Eight hundred and fifteen (815) of these translated GP1,2 protein sequences, comprising 92.93% of the initial download of nucleotide sequences, were utilized without further sequence re-alignment.

The consensus sequence of the GP1,2 dataset was determined with Jalview (2.8.2) [9]. Scores for random coil, helix and extended sheet predicted secondary structures were obtained for the consensus sequence with the PsiPred Protein Structure Prediction Server [10]. For representation of a variation of H distribution dependent upon protein secondary structure, each value of H was assigned to the mathematical array representing the secondary structure reported by PsiPred.

Results and Discussion

PsiPred scores summed along the length of the GP1,2 sequence are shown in Figure 1 for random coil, helical and extended sheet secondary structures. The summed random coil scores were significantly greater than the summed scores for helices (U=23297.0, p=3.232e-182) and the summed scores for extended sheets (U=49522.0, p=7.843e-138). The summed scores for helices were larger than those for extended sheets (127759.0, p=1.5488e-45). The total, summed PsiPred score was 380.959 for random coil, 95.239 for helix and 53.068 for extended sheets. Counting of PsiPred output showed that total number of amino acids in random coils (N=485) was greater than the number either helices (N=111) or in extended structures (N=80). The results in Figure 1 show that the predominant predicted secondary structure of GP1,2 is the random coil. The predicted distribution of each of three secondary structures along the length of the GP1,2 protein is shown in Figure 2. Mann-Whitney U tests on these unsummed distributions show that random coil scores were significantly greater than those for helices (U=111216.0, p=2.5389e-72) and for extended sheets (U=83790.0, p=4.2439e-112). The PsiPred scores for helices were slightly greater than those for extended sheets (U=215249.5, p=0.0023). Thus, the statistically ranked PsiPred scores for GP1,2 secondary structures are: random coils > helices > extended sheet. Plots of the PsiPred score

*Corresponding author: Joel K. Weltman, Clinical Professor Emeritus of Medicine, Alpert Medical School, Brown University, Providence, RI 02912 USA, Tel: 401-863-1000; E-mail: joel_weltman@brown.edu

Received October 08, 2015; Accepted November 09, 2015; Published November 16, 2015

Citation: Weltman JK (2015) Information Entropy and Protein Secondary Structure in the ZEBOV-Makona Ebola Virus Glycoprotein. J Health Med Informat 6: 207. doi:10.4172/2157-7420.1000207

Copyright: © 2015 Weltman JK. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
The predicted secondary structure of the ZEBOV-Makona glycoprotein is shown in Figure 1. The receptor binding domain (RBD) was observed in helices and sheets, with 7.27% in random coil structures, 88.00% in random coils, and 7.27% in extended sheets. The total summed information entropy (4.3742 bits) was greater than the total summed information entropy of the Makona variant reported here. These low entropy positions may reflect mutational constraints that are conformationally imposed. Extended sheet structures are concentrated in the ZEBOV-Makona glycoprotein (GP1,2) as random coil structures, with 7.27% in helices and 4.73% in extended sheets.

The GP1,2 amino acid H values were sorted into three sets, depending upon the secondary structure predicted for each of the 676 amino acid positions. The GP1,2 amino acid position, y=PsiPRED score for indicated protein secondary structure, was observed in helices and sheets, with high entropy values in the helical and random coil structures. These three amino acid positions recently have been shown to be associated with mutational escape. The presented results also suggest that the random coil glycoprotein regions with extended sheets are attractive and solvent accessible [18].

In contrast, the high entropy positions in the RBD, discussed above, were 32 low entropy, extended sheet structures within the ZEBOV-Makona RBD. These low entropy positions may reflect mutational constraints that are conformationally imposed. Extended sheets are concentrated in the RBD of the glycoprotein (Figure 2). The low values of H in the extended sheets consist with an increase in inter-amino acid side chain interactions that physically stabilize the sheet structure. Such physical stabilization would make emergence of successful mutations in the sheets less probable. Unlike the helix and extended sheet structures with intramolecular conformational constraints, random coil structures tend to be open and solvent accessible [18]. Thus, stabilized extended sheet amino acids, surrounded by penetrable coil structures, could be attractive targets for anti-EBOLA antibodies, especially if the sheets are located in the RBD which is essential for viral function.

Conclusions

The presented results show that information entropy in the ZEBOV-Makona glycoprotein is distributed mainly to amino acid positions in random coil structures. It is proposed that the random coil glycoprotein regions with H>0 represent sites of interaction of Ebola virus with its external environment, resulting in evolution towards mutational escape. The presented results also suggest that the observed low incidence of mutations in extended sheet positions may...
be caused by conformational constraints. Conformational epitopes displayed by the sheets, especially within the RBD, would be attractive targets for development of anti-Ebola vaccines and immunological therapies. Thus, this research helps provide insight for the development of vaccines against both linear and conformational Ebola epitopes.

References

1. Kuhn JH, Andersen KG, Baize S, Bao Y, Bavari S, et al (2014) 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 6: 4780-4799.

2. Cover TM and Thomas JA (2006) Entropy, Relative Entropy and Mutual Information. In Elements of Information Theory. (2ndedn) Wiley, USA.

3. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.

4. Lee JE and Saphire EO (2009) Ebola virus glycoprotein structure and mechanism of entry. Future Virol 6: 621-635.

5. Miller EH, Chandran K (2012) Filovirus entry into cells - new insights. Curr Opin Virol 2: 206-214.

6. Kaushik A, Tiwari S, Jayant RD, Marty A, Nair M (2016) Towards detection and diagnosis of Ebola virus disease at point-of-care. Biosens Bioelectron 75: 254-272.

7. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227: 1435–1441.

8. Shannon, Claude E (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27: 379–423.

9. 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.

10. Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 41:W340-W348.

11. Weltman JK (2015) Mutual Information-Based Cliques of Amino Acids in the Zaire Ebola Virus-Makona Glycoprotein; In Proceedings of the 2nd Int. Electron Conf. Entropy Appl. in the press.

12. Dube D, Brecher MB, Delos SE, Rose SC, Park EW, et al. (2009) The primed ebolavirus glycoprotein (19-kilodalton GP2): sequence and residues critical for host cell binding. J Virol 83: 2883-2891.

13. Lee JE, Fusco ML, Hessell AJ, Oswald WB, Burton DR, et al. (2008) Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. Nature 454: 177-182.

14. Vossen MT, Westerhout EM, Söderberg-Nauclér C and Wiertz EJ (2002)Viral immune evasion: a masterpiece of evolution. Immunogenetics. 54:527-542.

15. De Groot AS, Knopf PM, Rivera D, Martin W (2008) Immunoinformatics Applied to Modifying and Improving Biological Therapeutics; in Immunoinformatics; C Schönbach, S Ranganathan and V Brusic Springer : 109-131.

16. Weltman JK (2014) Identification of Invariant Peptide Domains within Ebola Virus Glycoprotein GP1,2. J Med Microb Diagn 4: 176.

17. Weltman JK (2014) Combined Use of Information Entropy and Bepipred Scores for Screening Ebola Virus Glycoprotein (GP) Sequences. In Proceedings of the 1st Int. Electron Conf. Entropy Appl. Sciforum Electronic Conference Series 1: d003.

18. Lins L, Thomas A and Brasseur R (2003) Analysis of accessible surface of residues in proteins. Protein Sci 12: 1406–1417.