Conference Proceedings Paper – Entropy

Combined Use of Information Entropy and Bepipred Scores for Screening Ebola Virus Glycoprotein (GP) Sequences

Joel K. Weltman [1],*

[1] Clinical Professor Emeritus of Medicine, Alpert Medical School, Brown University, Providence, RI 02912 USA

* E-Mail: joel_weltman@brown.edu

Abstract: Unprecedented epidemics caused by strains of Ebola virus (EBOV) are associated with extremely high mortality rates. Because EBOV glycoprotein (GP) is the precursor of the proteins mediating the binding and internalization of the virus, EBOV GP is an attractive target for vaccine development. It is reported here that by means of combined use of information entropy (**H**) and Bepipred predictive epitope screening of GP amino acid sequences, four invariant oligopeptide sequences were identified in GP of EBOVZ (EBOV Zaire) and EBOVS (EBOV Sudan), the two most epidemiologically prominent strains of EBOV. Three of these four oligopeptide sequences were subsequently identified in the GP protein of the Reston strain of EBOV (EBOVR), with a GP sequence length different from that of GP EBOVZ and GP EBOVS. It is suggested that the invariant oligopeptides identified by the combined information entropy and B-cell linear epitope screening procedure may be useful as components of oligopeptide anti-EBOV vaccines.

1. Introduction

A bioinformatic analysis of Ebola virus (EBOV) glycoprotein (GP) [1] is presented here based upon the combined distributions of information entropy [2] and predicted B-cell epitope score [3] in full length GP protein sequences. Most recognized cases of human infection with EBOV have been caused by EBOVZ (Zaire) and EBOVS (Sudan) strains of EBOV [4, 5]. Accordingly, this study focuses on GP amino acid sequences of those two EBOV strains. For comparison, an oligopeptide sequence search analysis was conducted on the GP protein of EBOVR (Reston); Ebola virus. EBOVR is a strain of EBOV that has not been reported to cause disease in humans [6].

2. Results and Discussion

2.1 Results

Figure 1. Immunobioinformatic Analysis of Ebola Virus (EBOV) Glycoprotein (GP).

(top) H distributions in EBOVS (black) and EBOVZ (red) GP sequence datasets; **(middle) H** distribution in the set of combined {EBOVS, EBOVZ} GP sequences; **(bottom)** β is the conditional Bepipred score of the combined {EBOVS, EBOVZ} GP sequences. The receptor binding domain (**RBD**) common to both sets of GP sequences extends between the vertical red lines at amino acids 54 and 185.
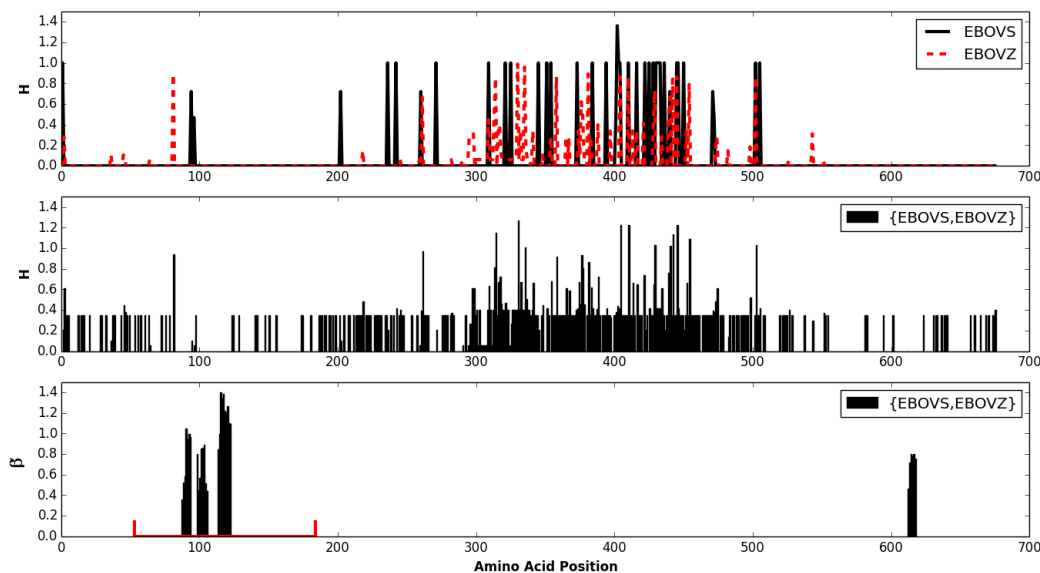
The **H** distributions in the GP(EBOVS) sequence set and in the GP(EBOVZ) sequence set are shown in Figure 1, top. The summed total entropy values (**ΣH**) for these GP sequences were: **ΣH**(EBOVS) = 36.3776) and **ΣH**(EBOVZ) =26.6990, respectively. These two **ΣH** values are indistinguishable (Z = 1.2300, p = 0.2187). However, although the patterns of distribution of H in the two GP datasets are similar, they clearly are not identical. The **H** distribution in the combined GP(EBOVS, EBOVZ) dataset is shown in Figure 1, middle. The summed **H** value obtained for the combined GP datasets was: **ΣH**(EBOVS, EBOVZ) =132.1436. This increased summed entropy value was significantly greater than that obtained for either the EBOVS GP dataset (Z = 7.6884, p= 1.4895e-14) or the EBOVZ GP dataset (Z=8.3938, p =4.7084e-17).

The distribution of screened Bepipred scores (β) in the EBOV GP sequences shown in Figure 1 (bottom) revealed four clusters of amino acids that fulfilled the β epitope screening criteria. Within those clusters, four oligopeptides, each consisting of at least six contiguous amino acids were identified in both EBOVZ and EBOVS GP proteins. (Table I). As shown in the table, three of those four GP oligopeptides were also identified in the EBOVR GP protein.

Table 1. Distribution of Predicted Epitopes in GP Protein of Strains of Ebola Virus (EBOV). Amino acid positions of epitopes for each EBOV strain are given by column. Absence of the predicted epitope is indicated by the dash (—).

Predicted Epitope	EBOVZ	EBOVS	EBOVR
FRSGVPP	88-94	88-94	89-95
YEAGEWAE	99-106	99-106	100-107
KKPDGSECLP	114-123	114-123	——
HDWTKN	613-618	613-618	614-619

2.2 Discussion

The glycosylated GP protein of the EBOV virus is cleaved into a GP1 protein and a GP2 protein [7, 8]. The GP1 protein is responsible for the binding of the EBOV virus to the membrane of the target cell. The GP2 protein is responsible for internalization of the virus into the target cell. Thus, GP protein is a reasonable subject for the design of a protective and ameliorative anti-EBOV vaccine.

By means of the combined use of a bioinformatic metric (**H**) and a screened Bepipred score (β), four GP peptides were identified as potential antigens in all of the complete EBOBZ and EBOVS GP sequences in the NCBI Genbank dataset. Three of these peptide sequences, a heptapeptide, an octapeptide and a decapeptide, reside within the receptor-binding domain (RBD) of the GP1 protein. The fourth identified peptide sequence, a hexapeptide. lies near the tail-region of the GP2 protein. The immunologic activity predicted in this current study differs from that of a set of antigens previously reported for EBOV GP [9]. Two of the predicted epitopic oligopeptides (FRSGVPP and HDWTKN) lie completely outside known antigenic regions in humans, except for the single N-terminal proline [10]. In contrast, the octapeptide (YEAGEWAE) and decapeptide (KKPDGSECLP) predicted epitopic oligopeptides lie within larger domains of GP proteins reported to be recognized by human antibodies in recovered EBOVZ patients and/or survivors [10]. Thus, these four predicted oligopeptide epitopes help detect and further define regions of the GP protein with immunologic potential.

Proline has been reported to play a significant role in Ebola antigen structure and function [11]. Proline seems to have played a significant role in the identified EBOV GP oligopeptide epitopes reported here. Two RBD oligopeptides with predicted epitopic potential, GP(88-94) and GP(114-123) each contains two proline residues. Proline has been reported to be a common component of epitope-containing segments of proteins, in association with specific effects of proline on protein secondary structure [12, 13].

Significant progress has been made towards a nucleotide-based anti-EBOV vaccine [14]. An oligopeptide based anti-EBOV vaccine would not require synthesis of the viral antigen in the vaccine recipient. This may shorten the response time to the vaccine, which would be a significant advantage given the brief incubation period and high mortality of EBOV in humans.

3. Experimental Section

The complete sets of EBOVZ (N=143) and EBOVS (N=10) GP protein sequences were downloaded from the NCBI Genbank [15] in FASTA format [16] on August 15, 2014. All of the GP sequences were full length, ie, 676 amino acids long. A set of EBOVR (Reston) sequences (N=9) was also downloaded, in which protein sequence full length was 677 amino acids.

Computation of information entropy (**H**) [2] and Z-tests on those computations were performed with the Enthought Canopy 1.4.1 distribution of 64-bit Python 2.7.6. Those computations were performed for both of the individual sets of GP amino acid sequences and for the combined set of EBOVS and EBOVZ GP sequences. **H** values were statistically evaluated by Z-test, using 1000 pseudo-random trials.

A consensus sequence for the combined GP sequence sets was determined with Jalview [17] and a B-cell predicted epitope score (**B**) was determined for each amino acid position of that consensus with Bepipred [3]. The **B** epitope score at each amino acid position was screened, conditional upon the following two criteria:

criterion #1 H = 0 at the selected amino acid position;
criterion #2 B ≥ 0.350, which is the Bepipred recommended cutoff,.

In this paper, the screened epitope **B** Bepipred score is denoted as β.

4. Conclusions

On the basis of the combined use of information entropy (**H**) and B-cell epitope score, four invariant oligopeptide epitopes, common to EBOVZ and EBOVS GP proteins were identified. Two of the identified oligopeptides each contain two proline residues. It is proposed that structural studies of these peptides can increase our understanding of EBOV structural biology and that vaccines based upon these oligopeptides may be useful for preventing and ameliorating disease caused by Ebola virus.

Acknowledgments

Author Contributions

Joel K Weltman is the only author of this research.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

1.   Beniac DR, Melito PL, deVarennes SL, Hiebert SL, Rabb MJ, et al. **2012** The Organisation of Ebola Virus Reveals a Capacity for Extensive, Modular Polyploidy. PLoS ONE 7(1): e29608. doi:10.1371/journal.pone.0029608.

2.   Shannon, Claude E. (July–October **1948**). "A Mathematical Theory of Communication". Bell System Technical Journal 27 (3): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

3.   Larsen JE, Lund O and Nielsen M **2006** Improved method for predicting linear B-cell epitopes. Immunome Res 2: 2.

4.   Dudas G and Rambaut A **2014** Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLOS Currents Outbreaks. doi: 10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.

5.   Sanchez A, Pierre E and Rollin PE **2005** Complete genome sequence of an Ebola virus (Sudan species) responsible for a 2000 outbreak of human disease in Uganda. Virus Res 113: 16–25.

6.   Marsh GA, Haining J, Robinson R, Foord A et al **2011** Ebola Reston Virus Infection of Pigs: Clinical Significance and Transmission Potential. J Inf Dis 204:S804–S809. doi: 10.1093/infdis/jir300

7.   Dube D, Brecher MB, Delos SE et al. **2009** The Primed Ebolavirus Glycoprotein (19-Kilodalton GP1,2): Sequence and Residues Critical for Host Cell Binding; J Virol 83(7):2883–2891; doi:10.1128/JVI.01956-08

8.   Lee JE, Fusco ML, Ann J. AJ et al. **2008** Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor; Nature 454:177-182;| 10 July 2008| doi:10.1038/nature07082

9.   Wilson JA, Hevey M, Bakken RB et al. **2000** Epitopes Involved in Antibody-Mediated Protection from Ebola Virus; Science 287:1664-1666.

10.   Becquart P, Mahlako T, Nkoghe D and Leroy EM **2014** Identification of Continuous Human B-Cell Epitopes in the VP35, VP40, Nucleoprotein and Glycoprotein of Ebola Virus. PLOS ONE June 2014 Volume 9 | Issue 6 |e96360.

11.   Rutledge RD, Huffman BJ, Cliffel DE and Wright DW **2008** Design and synthesis of an antigenic mimic of the Ebola glycoprotein. J Mater Res. 23(12): 3161–3168. doi:10.1557/JMR.2008.0384.

12.   Yu H ; Chen JK ; Feng S et al.**1994** Structural Basis for the Binding of Proline-Rich Peptides to SH3 Domains. Cell 76:933-945.

13.   VanHoof G, Goossens F, De Meester I et al. **1995** Proline motifs in peptides and their biological processing; FASEB J. 9:736-744.

14.   Stanley DA, Honko AN, Asiedu C et al. **2014** Chimpanzee adenovirus vaccine generates acute and durable protective immunity against ebolavirus challenge. Nature Medicine; advance online publication; 7 Sept 2014, doi:10.1038/nm.3702

15.   Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and David L. Wheeler **2006** GenBank; Nucleic Acids Res 34(Database issue): D16–D20. doi: 10.1093/nar/gkj15

16.   Lipman, DJ and Pearson, WR **1985** Rapid and sensitive protein similarity searches. Science 227, 1435–1441. doi:10.1126/science.2983426. PMID 2983426 (1985).

17.   Waterhouse, A.M., Procter, J.B., Martin, D.M.A, et al. **2009** Jalview Version 2 - a multiple sequence alignment editor and analysis workbench; Bioinformatics 25, 1189-1191 doi: 10.1093/bioinformatics/btp033 (2009).