Analysis of Ebola Glycoprotein Sequences from Strains of Varying Lethality

Biochem 218
Spring 2002
Tammy Doukas
tdoukas@stanford.edu

I. Background and Significance

Ebola hemorrhagic fever is a disease in humans, chimpanzees, and monkeys, caused by infection with Ebola virus, and associated with high mortality. This virus was first recognized in Zaire (now the Democratic Republic of Congo), Africa in1976. The exact origin and location of Ebola virus is still unknown.

Ebola virus is one of only two known members of a family of RNA viruses, the Filoviridae. (The other member is the Marburg virus.) The Filoviridae consists of enveloped viruses containing a non-segmented negative-stranded RNA genome. The Ebola genome shows a linear gene arrangement in the order 3' untranslated region -- nucleoprotein (NP) -- viral structural protein -- (VP) 35 -- VP40 -- glycoprotein (GP) -- VP30 -- VP24 -- polymerase (L) -- 5' untranslated region. Genes are flanked at their 3' and 5' ends by highly conserved transcriptional start and termination signal sequences, respectively, which all contain the pentamer 3'-UAAUU-5'. The genes show three overlaps that alternate with intergenetic sequences (Klenk et al, 2000).

To date, four species of Ebola virus have been identified: Ebola Zaire, Ebola Sudan, Ebola Ivory Coast, and Ebola Reston. Different strains have been identified among within the species. Ebola Zaire consists of four identified strains, Zaire Mayinga, Zaire-95, Eckron-76, and Gabon-94. Ebola Sudan consists of Sudan Boniface and Sudan Maleo-79. Ebola Reston consists of Reston and Reston Siena/Philippine-92. Ebola Ivory Coast consists of only one known strain, Ivory Coast-94.

All four known species of Ebola virus have infected humans, but with differing degrees of lethality between species and even among different strains of the same specie (Table 1, adapted from information from the CDC). Zaire Mayinga and Zaire-95 are the two most lethal forms of the Ebola virus, killing approximately 85% of all known infected humans. Zaire Gabon, Sudan Boniface, and Sudan Maleo-79 are less lethal, killing between 53 and 66% of its victims. Only two cases of Ivory Coast infection are known, and one person survived. Reston, which is lethal to monkeys, has infected several humans (as shown by the presence of antibodies), but no human deaths have occurred from this form of the virus.

Ebola Species and Strain	Number of Known Human Infections	Percentage of Deaths Among Cases
Zaire Mayinga	318	88%
Zaire-95	315	81%
Zaire Gabon-94	143	66%
Sudan Boniface	710	53%
Sudan Maleo-79	34	65%
Ivory Coast-94	1	0%
Reston	0 (however, 8 known infected due to presence of antibodies)	0%

Table 1. Ebola Strain Lethality

The difference in lethality of the different forms of Ebola virus must lie somewhere in the genetic code. But where it lies is still a mystery. To try to discern the differences between the different forms of the virus, a bioinformatics approach was used to analyze sequence differences, focusing on sequence variability and corresponding function.

II. Ebola Virus Species and Strain Identification

An NCBI search in the Taxonomy browser for Ebola sequences produces the species and strains listed below.

Under **Lineage** (abbreviated): root; Viruses; ssRNA negative-strand viruses; Mononegavirales; Filoviridae; Ebola-like viruses:

- **Ebola-like viruses**☜*Click on name to get more information.*

  - **Cote d'Ivoire Ebola virus**
    - **Ebola virus strain Ivory coast-94**
  - **Ebola virus**
  - **Reston Ebola virus**
    - **Ebola virus strain Reston**
    - **Ebola virus strain Reston Siena/Philippine-92**
  - **Sudan Ebola virus**
    - **Ebola virus strain Sudan Boniface**
    - **Ebola virus strain Sudan Maleo-79**
  - **Zaire Ebola virus**
    - **Ebola virus strain Eckron-76**

- **Ebola virus strain Gabon-94**
- **Ebola virus strain Zaire Mayinga**
- **Ebola virus strain Zaire-95**

The database provides eight different protein sequences of Zaire Mayinga, three protein sequences from Zaire-95, two from Eckron-76, four from Gabon-94, three from Sudan Boniface, three from Sudan Maleo-79, two from Ivory Coast-94, two from Reston, and two from Reston Siena/Philippine-92.

Protein Sequence Identification for Analysis

The two protein sequences included in the NCBI database from all known Ebola species are “small/secreted glycoprotein precursor” and “structural glycoprotein precursor (virion spike glycoprotein)”. The reason for this is likely due to the known role of glycoproteins in other viral infections. Ebola virus enters monocytes, macrophages, hepatocytes, and endothelial cells by employing their surface glycoproteins. Receptor binding and membrane fusion are initiated by a single glycoprotein. Glycoprotein is synthesized as a precursor that is cleaved by the trans-Golgi network by proprotein convertase furin into GP1 and GP2 (Klenk et al, 2000). This type of cleavage is a determinant of pathogenicity as has been observed with other viruses including influenza and paramyxoviruses (Klenk et al, 2000). A second type of cleavage, possibly due to metalloproteases, is reported to occur at the cell surface and cause the release of the entire ectodomain of glycoprotein in soluble form (Klenk et al, 2000). Also reported is the presence of two frame-shifted open reading frames (ORFs) in the glycoprotein gene. Transcription of only the first ORF results in expression of sGP, the small/secreted glycoprotein (Klenk et al, 2000).

III. Phylogenetic Comparisons

To determine whether a correlation exists between glycoprotein sequence variation and strain lethality, a phylogenetic analysis of the glycoprotein sequences was performed. Four different methods, Neighbor-Joining, UPGMA, ClustalW, and PileUp, were employed and compared in the construction of phylogenetic trees.

Secreted Glycoprotein Accession Number	**Structural Glycoprotein Accession Number**	**Species / Strain**
Q66819	Q05320	Zaire Mayinga
Q66819	P87666	Zaire-95
P87670	P87671	Zaire Eckron-76
O11458	O11457	Zaire Gabon-94
Q89455	Q66814	Sudan Boniface
Q89455	Q66798	Sudan Maleo-79
Q66811	Q66810	Ivory Coast
Q66800	Q66799	Reston
Q89569	Q89853	Reston Siena/Philippine-92

Table 2. Ebola Species/Strain Sequences and Corresponding Accession Numbers

Neighbor-Joining Tree

The first secreted and structural glycoprotein trees were constructed using the Neighbor-Joining method (Fig 1 and 2). Neighbor-joining is a method that does not require that all lineages have diverged by equal amounts (Saitou and Nei, 1987). The method is especially suited for datasets comprising lineages with largely varying rates of evolution. Neighbor-joining keeps track of nodes on a tree rather than taxa or clusters of taxa. The raw data are provided as a distance matrix and the initial tree is a star tree. Then a modified distance matrix is constructed in which the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes. The tree is constructed by linking the least-distant pair of nodes in this modified matrix. When two nodes are linked, their common ancestral node is added to the tree and the terminal nodes with their respective branches are removed from the tree. This process converts the newly added common ancestor into a terminal node on a tree of reduced size. At each stage in the process two terminal nodes are replaced by one new node. The process is complete when two nodes remain, separated by a single branch. Advantages of the Neighbor-Joining method are that it is fast and suited for large datasets, it permits lineages with largely different branch lengths, and it permits correction for multiple substitutions. Disadvantages include a reduction in sequence information, only one possible tree is produced, and the tree is strongly dependent on the model of evolution used.

Fig 1. **Neighbor-Joining Tree: Ebola Secreted Glycoprotein Precursor**

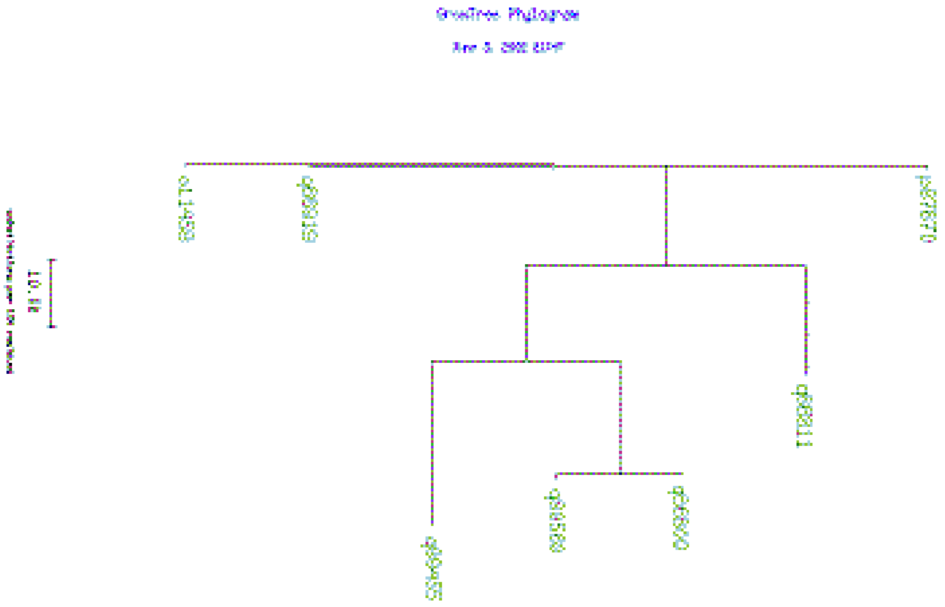
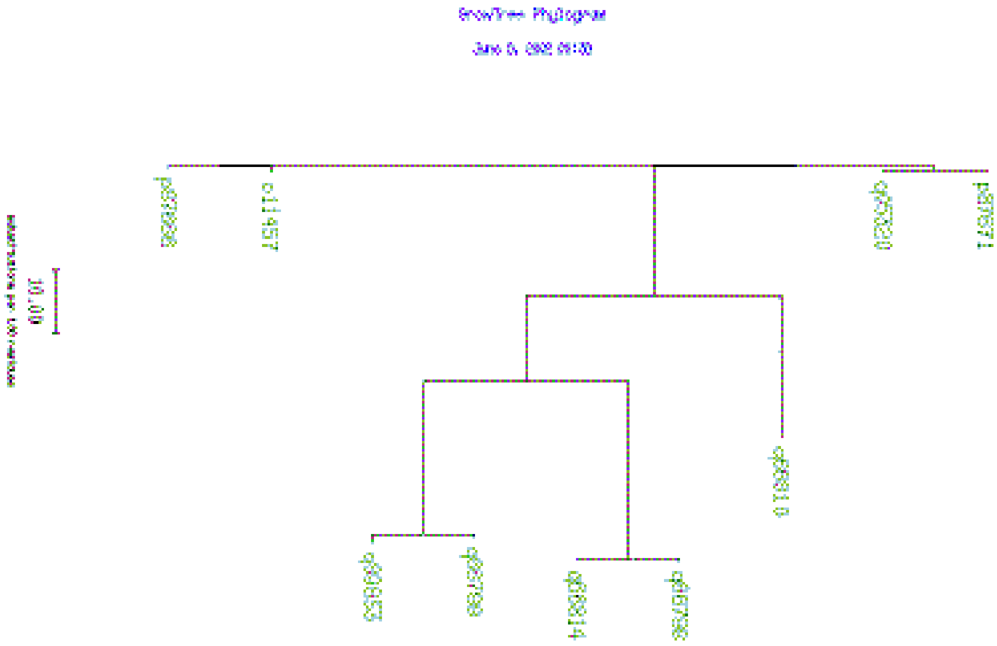


Fig 2. **Neighbor-Joining Tree: Ebola Structural Glycoprotein Precursor**



UPGMA Tree

The second tree was constructed using Unweighted Pair Group Method with Arithmetic Mean (UPGMA), (Fig 3 and 4). The UPGMA option constructs a tree by a sequential clustering algorithm, in which local topological relationships are identified in order of similarity, and the tree is build in a stepwise manner. It is a simple method, but it is very sensitive to unequal evolutionary rates.

Fig 3. **UPGMA Tree: Ebola Secreted Glycoprotein Precursor**

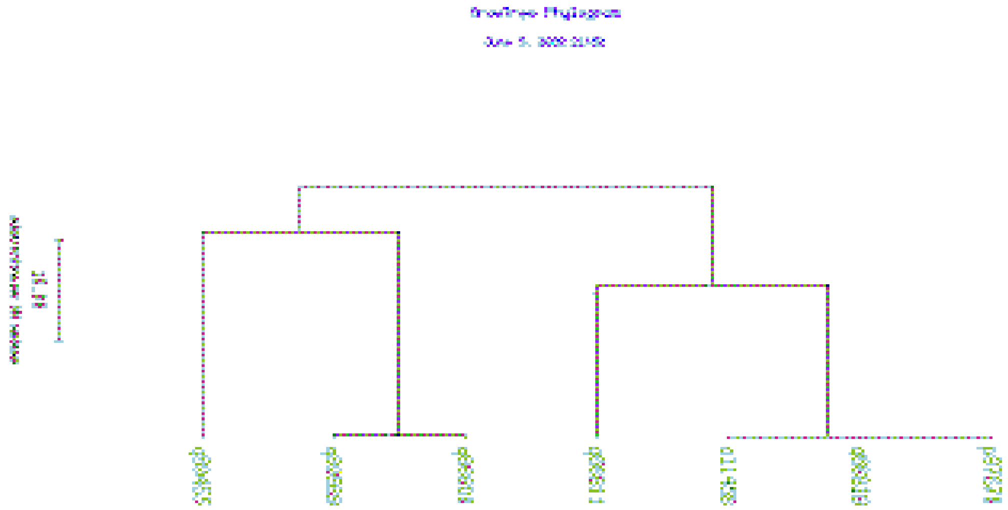
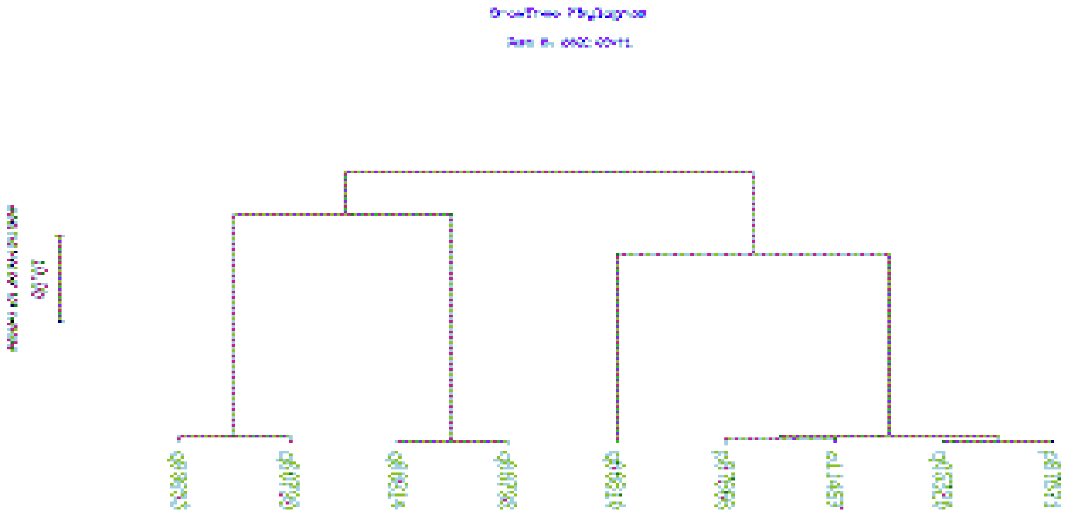


Fig 4. **UPGMA Tree: Ebola Structural Glycoprotein Precursor**



ClustalW Tree

The next tree was constructed using ClustalW Multiple Alignment (Fig 5 and 6). ClustalW is a heuristic algorithm, constructing trees using a progressive alignment method**,** built up in stages where a new sequence is added to an existing alignment (Thompson, Higgins, and Gibson, 1994). If the sequences are similar only in some smaller regions**,** while the larger parts are not recognizably similar, or if one sequence contains a large insertion compared to the rest, then ClustalW may have problems aligning all sequences properly. This is because ClustalW tries to find global alignments, not local. If one sequence contains a repetitive element, while another sequence only contains one copy of the element, then ClustalW may split the single domain into two half-domains to try to align the first half with the first the domain in the first sequence, and the other half to the second domain in the first sequence.

Fig 5. **ClustalW Tree: Ebola Secreted Glycoprotein Precursor**

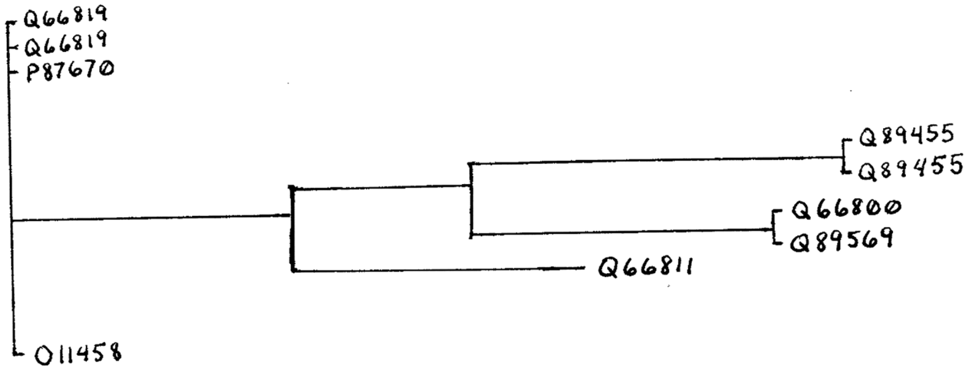
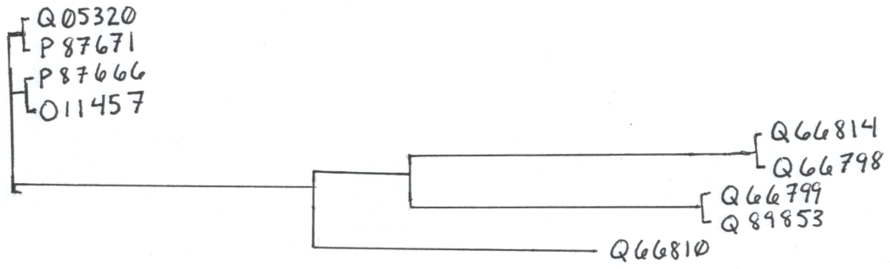


Fig 6. ClustalW Tree: Ebola Structural Glycoprotein Precursor



PileUp Multiple Sequence Alignment

A fourth tree was constructed using PileUp Multiple Sequence Alignment (Fig 7 and 8). PileUp creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments.

Fig 7. **Pileup Tree: Ebola Secreted Glycoprotein Precursor**

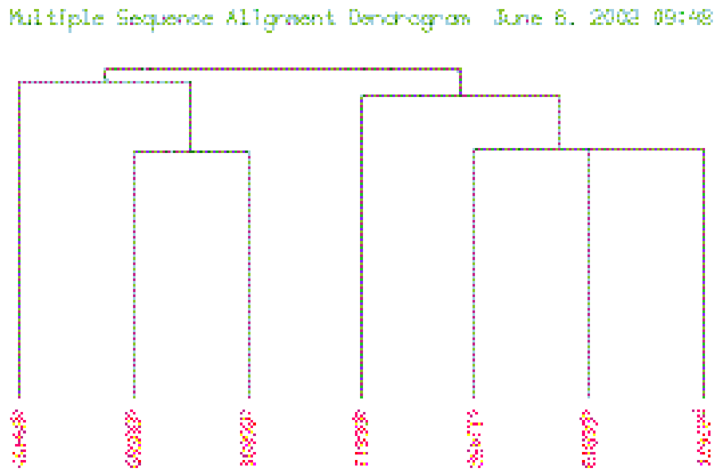
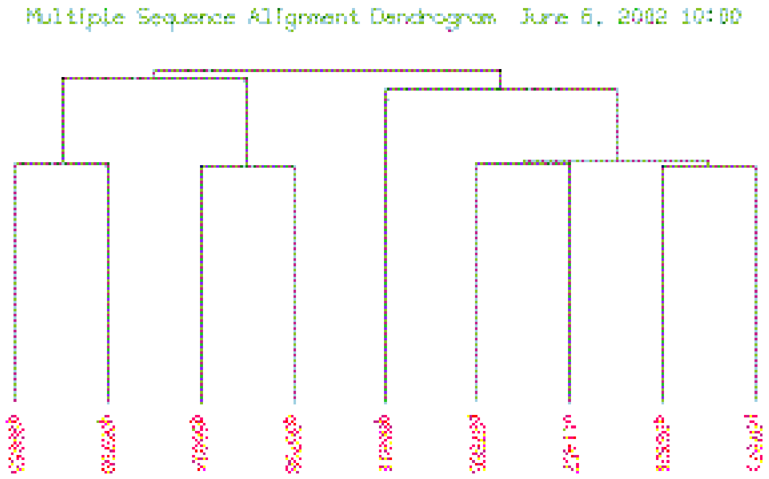


Fig 8. **Pileup Tree: Ebola Structural Glycoprotein Precursor**



Phylogenetic Comparison Results

Secreted Glycoprotein Results

All four tree-building methods produced similar results, clustering the different strains together within the corresponding species. However, the Neighbor-Joining method separated the Eckron-76 strain farther from the other Zaire species than did the other tree-building methods. Likewise, the ClustalW method separated the Gabon-94 strain farther from the other Zaire species.

Structural Glycoprotein Results

All four tree-building methods produced similar results, clustering the different strains together within the corresponding species. The Reston strains and Sudan strains clustered close together in all trees, while the Zaire strains were most distant from the other strains in all trees. In addition, all trees placed the Ivory Coast strain intermediate between the Zaire strains and the Reston / Sudan cluster. However, the Neighbor-Joining method put more distance between the Zaire Mayinga / Zaire Eckron-76 cluster and the Zaire-95 / Zaire Gabon-94 cluster than did the other tree-building methods.

Although all four tree-building methods produce results that are both biologically and taxonomically reasonable, there is one discrepancy in the Neighbor-Joining tree. Neighbor-Joining collapses the tree due to the average distance between each sequence and all others, resulting in a reduction in sequence information. This tree compares each sequence to the averaged remaining sequences, and so it may have identified a collapsed sequence as more distant to a remaining sequence such as the Zaire strain Eckron-76. The discrepancy could also have arisen as a result of the model of evolution used.

IV. Identification of Known Functional Regions of the Sequences

Identification of functional regions of the sequences is important in understanding the functional significance of sequence variability. First, sites and regions identified by the NCBI database as functional domains were noted. All Ebola sequences contained a glycosylation site around residue 40 as well as five glycosylation sites lying between residues 204 and 269 in the secreted glycoprotein. The main difference occurs toward the end of the sequence. Only the two most lethal forms of the virus, Zaire Mayinga and Zaire-95, contain what is described as the “Delta-Peptide”, spanning residues 325-364 of the secreted glycoprotein. None of the other less-lethal forms of the virus contain this Delta-peptide.

A search of the NCBI database for the Ebola structural glycoprotein precursor identified many glycosylation sites between residues 200 and 620, as well as a disulfide bond between cysteines 601 and 608. The database also identified the protein as extracellular, with a transmembrane region between 650 and 672 residues and a cytoplasmic region between residues 673 and 676. In addition, a GP1 sequence spanning residues 33 and 501 as well as a GP2 region spanning residues 502 and 676 were identified. All of the functional regions of the structural glycoproteins were identified in all of the Ebola strains.

Another program to look for the function of conserved regions is Accelrys Pattern Recognition “Motifs”. Motifs looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns. Motifs can display an abstract of the current literature on each of the motifs it finds. However, a Motifs search with both the secreted and structural glycoprotein sequences produced no hits. Likewise, a search of the PROSITE database with BLOCKS produced no known functional regions from the sequences. Several BLOCKS of high homology were produced, but all were of unknown function.

Delta-Peptide

The Delta-peptide was identified in only the two most lethal strains of the Ebola virus, Zaire Mayinga and Zaire-95. The Delta-peptide is the carboxy-terminal cleavage fragment of the secreted glycoprotein of Ebola (Volchkova et al, 1999). The secreted glycoproteins of all Ebola strains contain a site at position 295 or 296 that is cleaved by signalase. However, there is an additional cleavage site that has been found only in Zaire Mayinga and Zaire-95 that produces a smaller fragment between spanning residues 325 and 364 (Volchkova et al, 1999). Only the Delta-peptide cleavage fragment is highly O-glycosylated (Volchkova et al, 1999).

V. Sequence Variability

Methods used to identify regions of variability among the different strain sequences included Sequence Logos (Schneider et al, 1990), Wu-Kabat plot (Wu and Kabat, 1970), and PlotSimilarity. Sequence logos help identify the conserved residues, while the Wu-Kabat plot helps to identify regions of variability. PlotSimilarity plots the running average of the similarity among the sequences in a multiple sequence alignment. Optimally, plots should be designed from a large set of sequence data. However, a total of only nine secreted and nine structural Ebola glycoproteins are known.

Sequence Logos

Sequence Logos graphically plots the variability at each amino acid position (Fig 9 and 10). Strictly conserved regions are easy to identify, as only one letter corresponding to the appropriate residue will be present at that position. Highly variable positions will contain several compacted residues.

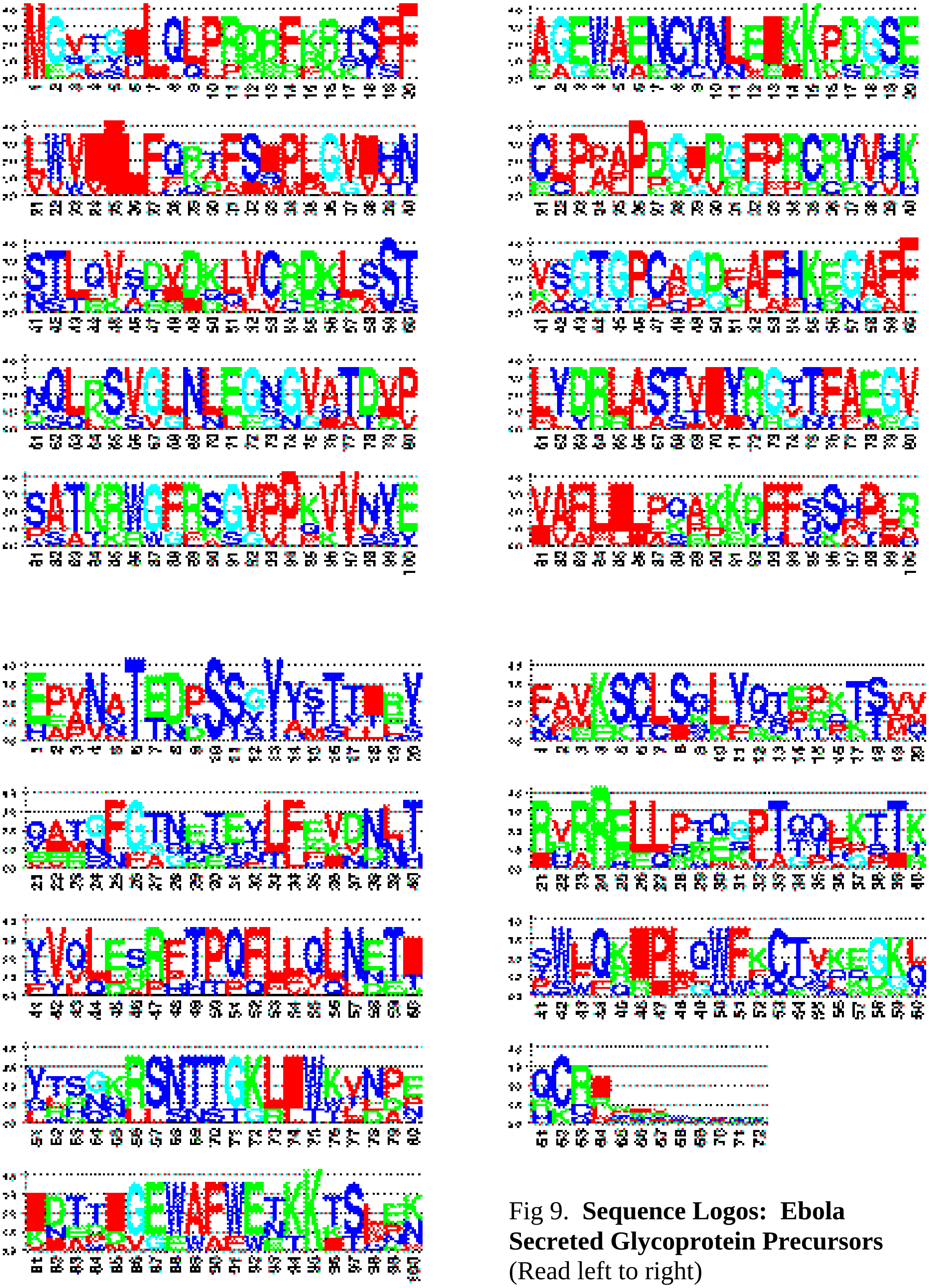
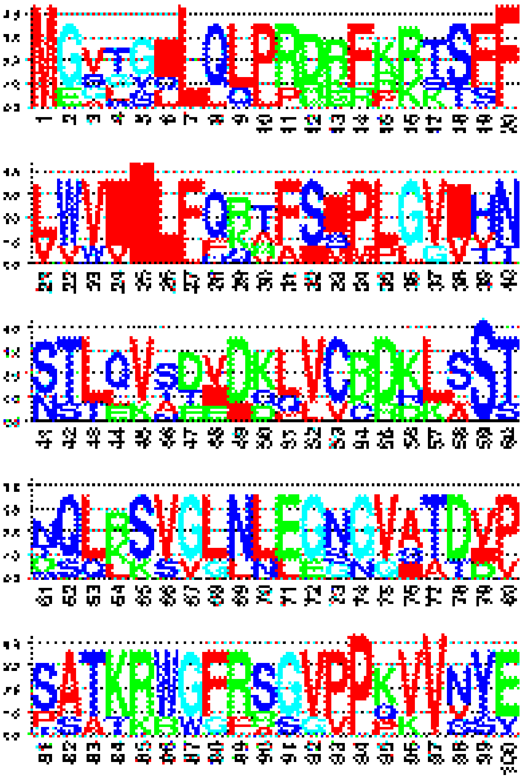
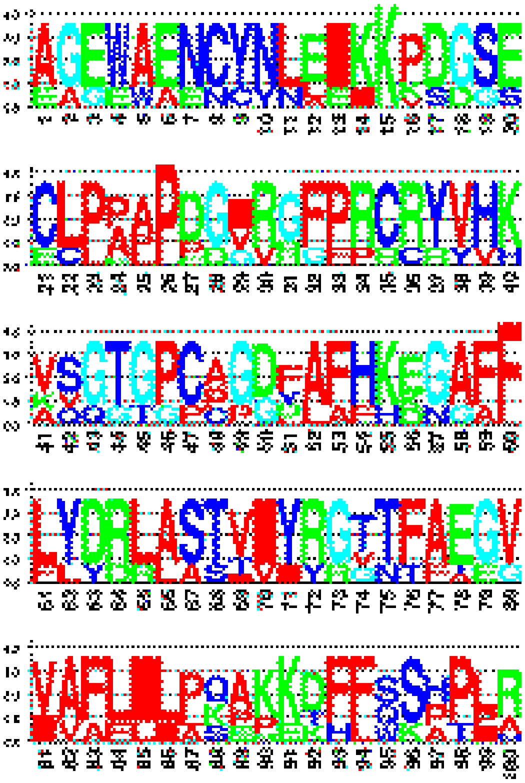
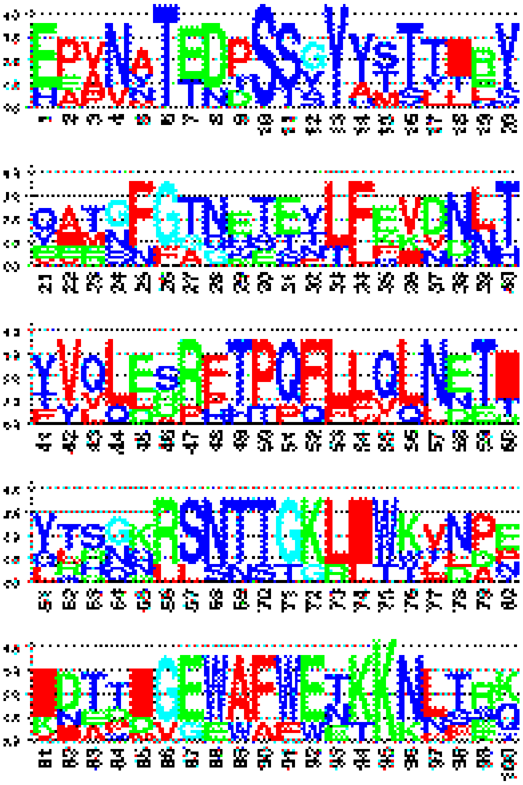
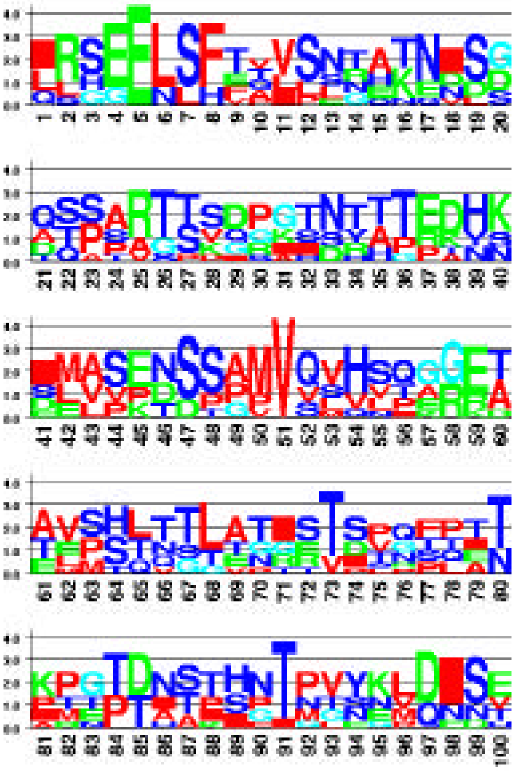


Fig 9. **Sequence Logos: Ebola Secreted Glycoprotein Precursors**
(Read left to right)









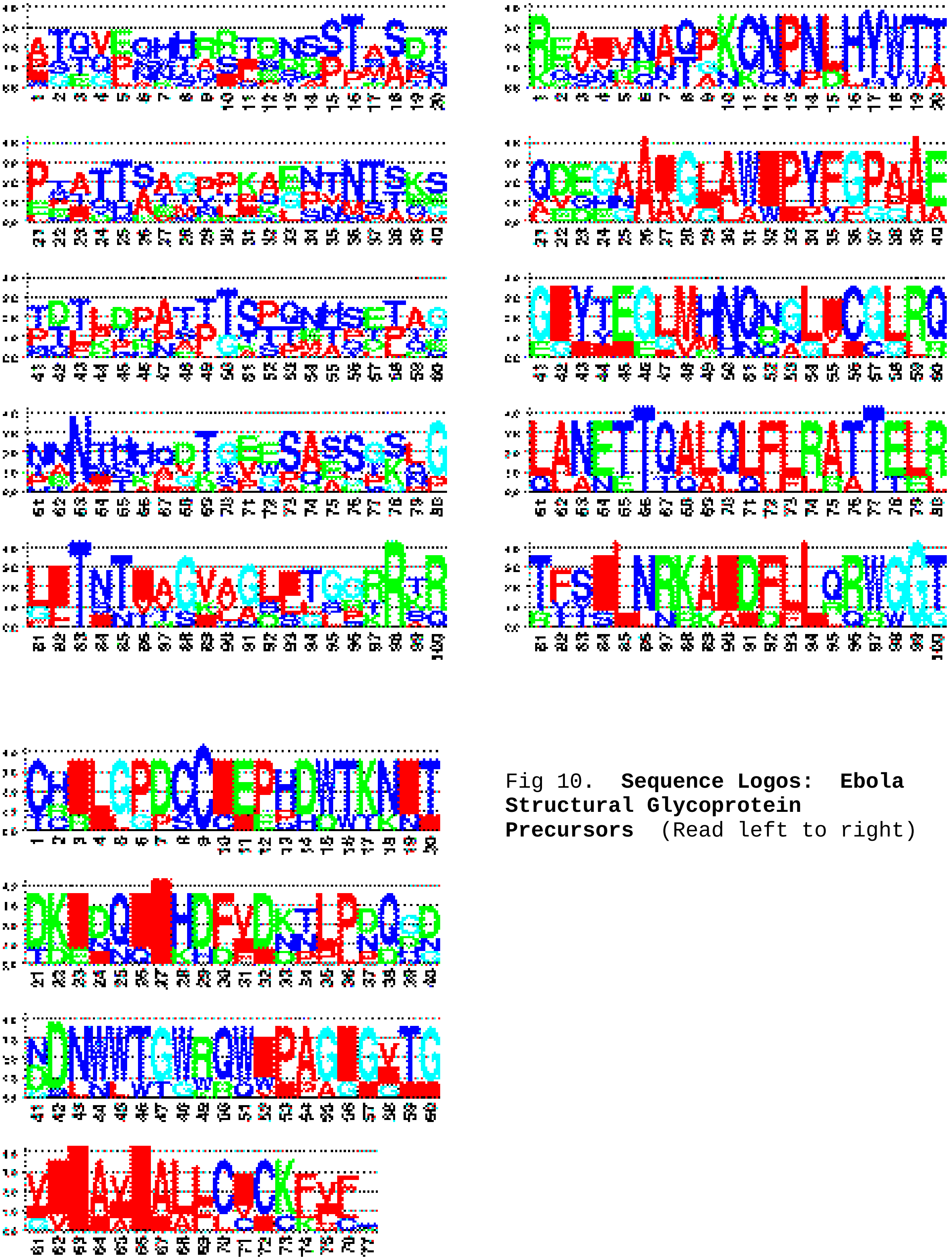
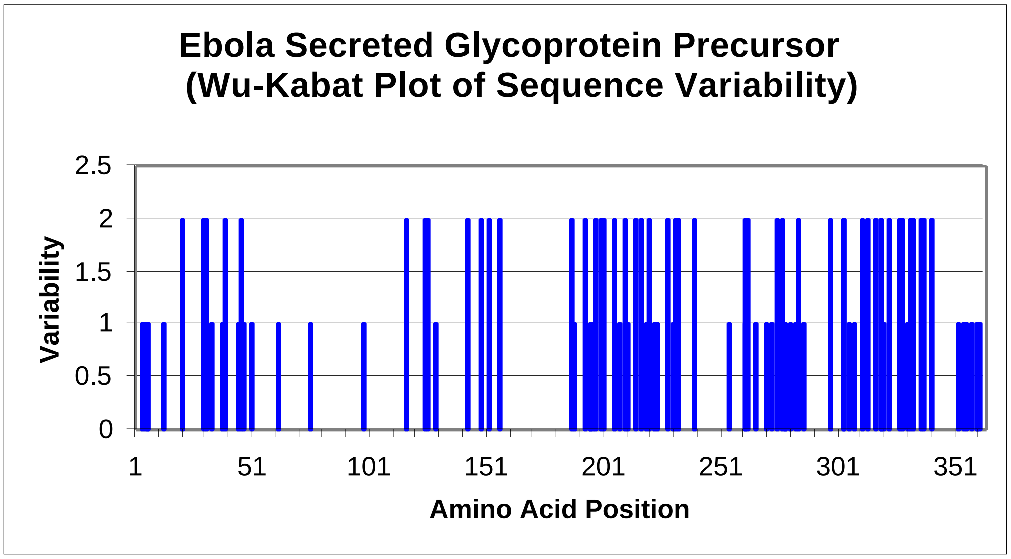
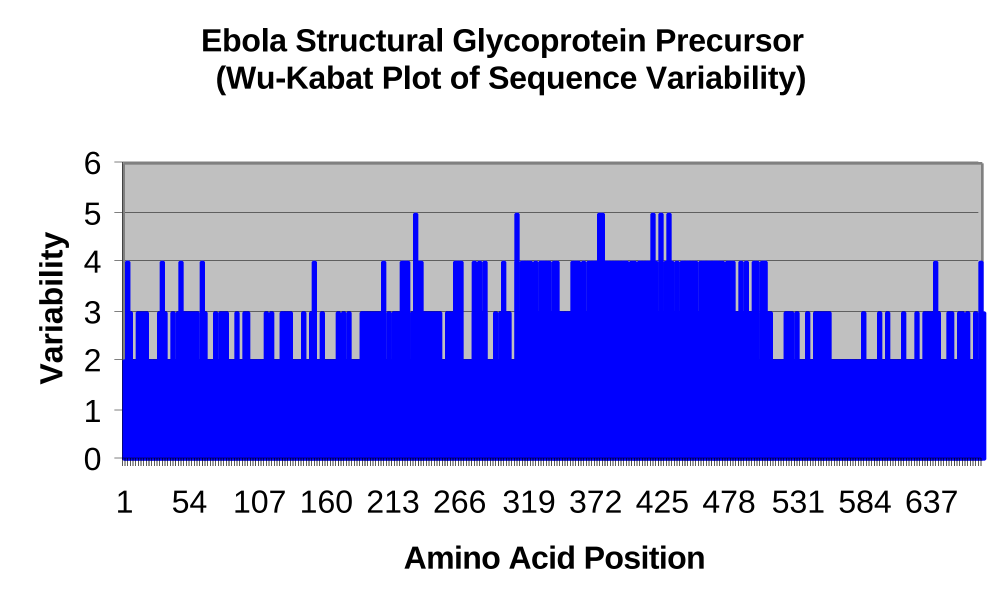


Fig 10. **Sequence Logos: Ebola Structural Glycoprotein Precursors** (Read left to right)

Wu-Kabat Plots

Wu-Kabat plots are useful in the identification of variable regions (Fig 11 and 12). The graph is created by plotting the ratio of the number of different residues found at a given position to the frequency of the most common residue at that position.





Accelrys PlotSimilarity

PlotSimilarity calculates the average similarity among all members of a group of aligned sequences at each position in the alignment, using a sliding window of comparison (Fig 13 and 14). The window is moved along all sequences one position at a time, and the average similarity over the entire window is plotted at the middle position of the window. The average similarity across the entire alignment is plotted as a dotted line.

Fig 13. **PlotSimilarity: Secreted Glycoprotein Precursors**

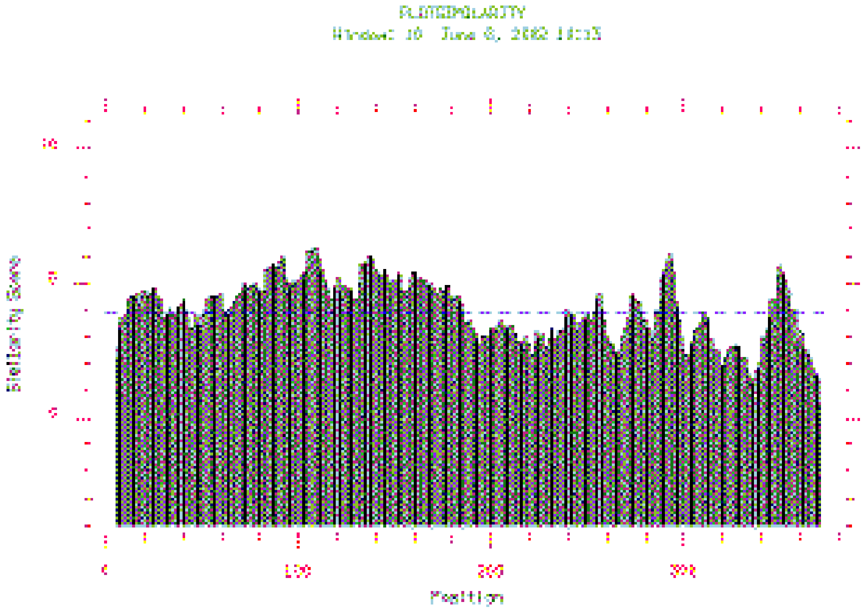
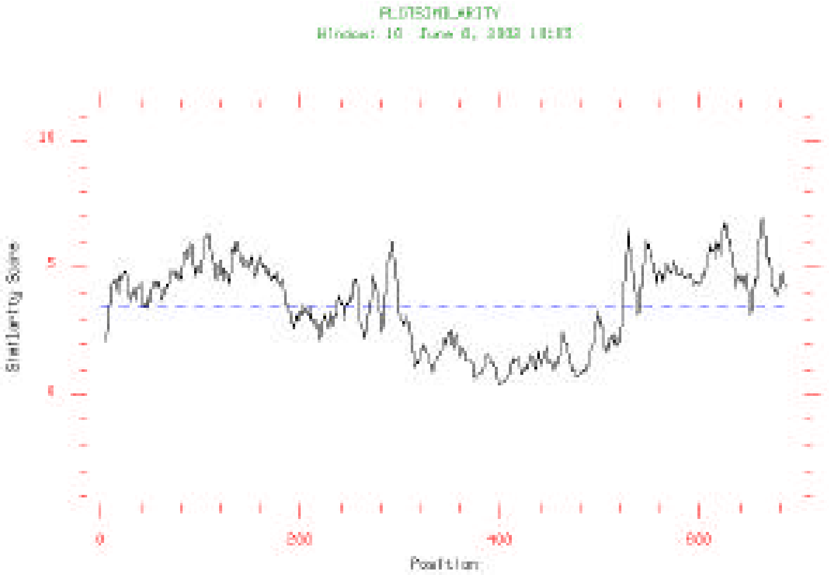


Fig 14. **PlotSimilarity: Structural Glycoprotein Precursors**



Sequence Variability Results

Even with the small subset of sequences used to create the variability plots, Sequence Logos, Wu-Kabat plot, and PlotSimilarity all recognize similar regions of variability (Table 3).

	Variable Regions of the Ebola Secreted Glycoprotein Precursor Sequences	**Variable Regions of the Ebola Structural Glycoprotein Precursor Sequences**
Wu-Kabat	185-230 260-285 300-340	300-505 (190-240 intermediately variable)
PlotSimilarity	180-230 250-270 300-345	300-520 (180-240 intermediately variable)
SequenceLogos	(similar to the regions listed above)	(similar to the regions listed above)

Table 3. Sequence Variability Results

VI. Functions of Variable Regions

A blast of the NCBI Database with the identified variable regions produced unknown functions for secreted glycoprotein precursor sequences 180-230 and 250-285, with the exception of several glycosylation sites. However, sequence 300-345 recognized a portion of the Delta-peptide discussed earlier.

A blast of the NCBI Database with structural glycoprotein precursor sequence 300-505 identified GP1. 300-520 also identified a portion of GP2. A blast with 180-240 also identified GP1. Several glycosylation sites were also recognized.

VII. Conclusions

The extremely high variability in lethality of the different Ebola virus species and strains has not yet been attributed to a specific cause. Identification and analysis of sequence variations of these strains may provide clues to the variability in viral function of such closely related and highly conserved strains. Secreted and structural glycoprotein precursors, the most well-characterized protein sequences of the Ebola virus family, and the only proteins of which all nine identified strains have been sequenced, were studied due to the known importance of glycoproteins in infections by other viruses.

Variability among the GP1 region of the structural glycoprotein precursor sequences may possibly be important for Ebola disease progression, as processing of GP1 in other viruses has been involved in viral pathogenicity. Likewise, the higher variability of the C terminal region of the secreted glycoprotein precursor sequences may provide a clue to the pathogenicity of the different strains of Ebola, where the two most lethal strains were found to contain a Delta-peptide that was not present in the less lethal or non-human lethal strains.

References

CDC
Special Pathogens Branch
Division of Viral and Rickettsial Diseases
National Center for Infectious Diseases
Centers for Disease Control and Prevention
U.S. Department of Health and Human Services

Klenk, H-D, W Slenczka, and Heinz Feldmann, 2000. Keystone Symposium “Cell Biology of Virus Entry, Replication, and Pathogenesis”, New Mexico.

Saitou, N and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul;4(4):406-25.

Schneider, TD and EM Stephens, 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18: 6097-6100.

Thompson, JD, DG Higgins, and TJ Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994 Nov 11;22(22):4673-80.

Valchkova, VA, H-D Klenk, and VE Volchkov, 1999. Delta-Peptide is the Carboxy-Teminal Cleavage Fragment of the Nonstructural Small Glycoprotein sGP of Ebola Virus. Virology 265 (1) 164-171.

Wu, TT and EA Kabat, 1970. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med (132) 211-250.