Haley Hauptfeld, Jess Kelly, Amanda Lowther, Andrew Ressler, Meadow Wicke
BCB4001
Professor Korkin
14 December 2019

Comparative Sequence-Based Analysis of EEEV and its Relatives

Abstract

Eastern Equine Encephalitis (EEE) is a virus with severe and sometimes fatal consequences for the humans it infects. There has been a local outbreak of EEE in Massachusetts in the summer of 2019. Currently, there is very little research on comparative sequence analysis related to EEE. Advancements in this research could contribute greatly to developing treatments. We performed comparative sequence analysis between EEE sequences from different years, different strains of the same classification, and found in different host species. We also performed phylogenetic analysis to show evolutionary behaviors among different strains. These analyses identified several distinctions in important amino acid residue chains. The goal of this research is to understand genetic differences and potential weaknesses in EEEV. This could play important roles in developing a vaccine and saving human lives.

1: Introduction

Comparative sequence analysis is a useful method for analyzing and deciphering biological sequences. It can be used between any sequences of interest, whether they be from the same virus at two different points in evolutionary time (such as several years apart), comparable sequences across different viruses, or between different strains of a virus. One of the most important applications of comparative sequence analysis is to understand genetic differences and potential weaknesses in the genes of pathogenic bacteria and viruses. These weaknesses can be exploited to create or improve vaccines and other treatments, and differences found between strains can shed insight on what precisely will be an effective treatment as strains evolve and diverge.

Eastern Equine Encephalitis (EEE) presents a promising area for research: it can have devastating consequences (including death) for humans infected with it, yet the relatively small number of cases present means that there has not been much research into it due to the lower demand for treatments when compared to other, more prevalent diseases. These factors are what allow original research to be done here, and the severity of the infections caused makes this of fundamental interest to the scientific community in order to improve human life.

2: Background

Eastern Equine Encephalitis is a single-stranded RNA virus transmitted via the bite of an infected mosquito. The local outbreak of EEE in the summer of 2019 in Massachusetts resulted in 3 deaths, and has left most of the survivors with traumatic brain deficits. Symptoms include high fever, headaches, and vomiting, and can progress into seizures, inflammation of the brain, and comas (Eastern Equine Encephalitis, 2019). Although there are treatments to address the symptoms, there is currently no vaccine for humans to target the actual virus. Massachusetts has an incredibly dense concentration of red maple and white cedar trees that propel EEE. Every year, birds infected with EEE migrate from Florida to New England, and live in these trees. From there, mosquitoes feed on the blood of these birds, and carry on EEE to mammals such as horses and humans (Saplakoglu, 2019).

In relation to comparative sequence-based analysis, other diseases have been tackled. For example, after the human genome project, biologists are now testing the drug, Herceptin, to see if it can be a treatment for breast cancer. Scientists were also able to leverage the tactic to compare three strains of Lumpy skin disease virus (LSDV). They compared the South African vaccine strain (LW), a virulent strain from an outbreak in South Africa (LD), and a virulent Kenyan 2490 strain (LK). They found that the LW strain had 438 amino acid residue substitutions in the virulent area of its genome as compared to the virulent LD strain. This led to deletions, insertions, and an altered open reading frame in regions that coded for gene expression, immune responses of the host, DNA repair, and more. Researchers suggested that the gene products in question be subjected to further study to improve the LSDV vaccine (PD, 2003).

Currently, not much research has been dedicated to EEEV in the field of bioinformatics, specifically comparative sequence based analyses. However, there have been studies related to the phylodynamic analysis of EEEV in the United States. Phylodynamic analysis finds the correlations between the epidemiological and evolutionary behavior of viral pathogens when they are within the immune system of a host. One study used phylodynamic analysis to find that EEEV evolves slowly, and is spread zoonotically in Florida. An advantage of phylodynamic analysis is that researchers can view the relations between the evolution of a virus and its ability to spread itself to other hosts. A disadvantage of this method is that you can only analyze data from a large scope. You cannot necessarily find a detailed answer as to why a specific outbreak occurs. Comparative sequence based analysis combats this disadvantage because it allows researchers to compare outbreaks of similar strains to EEEV that are occurring at the same time (Chang, 1987).

In order to provide a sequence-based analysis on EEEV, an online web tool called T-Coffee was used. T-Coffee is a popular multi-sequence alignment tool used to compare and analyze similar sequences. It has the ability to both combine multiple alignments as well as generate a library of

pairwise alignments which guides the multi-sequence alignment. It returns color-coded sequences that can be analyzed to determine the regions of high conservation between the sequences. Expresso, a special alignment feature offered by T-Coffee, uses protein data bank 3D structures as templates for the alignment. Expresso searches for closely related structures between the provided sequences and uses pairwise sequence alignment for the structures that do not match (Di Tommaso, 2011). Using this program to compare the evolutionary and regional strains of Equine Encephalitis will provide information to better understand the virus.

3: Methods

An analysis of sequence alignments were completed to compare protein sequences of various EEEV strains. The conservation of domains was determined within strains of the same year from different locations, across years, and between different host species. The differences in residues from the reference were observed across these alignments and were summarized in the form of potential effect of the change in residue. Brief analyses of secondary structure effects and phylogenetic relationships of the related EEEV strains were also completed to complement the multiple sequence alignment results.

3.1: Standardizing Frames of Analysis

The complete reference genome of the EEE virus was obtained by searching "Eastern equine encephalitis" in viruSITE (Stano et. al., 2016). This genome comes from the Volchkov et.al paper. Sequences for specific strains were obtained by searching the reference genome sequence using a BLASTN search against the unclustered U-RVDBv17.0 database in the Reference Viral Database (RVDB) (Goodacre et. al., 2018), and clicking the links to GenBank for the desired strains. From the GenBank page for each strain, the FASTA file was downloaded for the translated structural polyprotein (Benson, 2004). The full sequences were aligned using Expresso through the T-Coffee web service for multiple sequence analysis. All default settings were accepted except for the expect threshold; this value was changed from 10 to 0.001. From these results, domain A and domain B sequences of the E1 glycoprotein were isolated and used for consistency across analysis of all biological questions. These sequences were determined using NCBI's BLASTp tool by entering the structural polyprotein FASTA sequence of the reference EEEV genome and editing the database parameter to the PDB database (National Center for Biotechnology Information).

The T-Coffee results for the sequences were then exported to Core/TCS, and the resulting text was used to add the sequence alignment to Jalview, a software for multiple sequence alignment editing, for each analysis. Within Jalview, a secondary structure prediction was completed using the JPred web service to determine the predicted secondary structure and conservation of local structures within the alignment. Jalview was also used to generate a phylogenetic analysis of the

strains to show their evolutionary relationships. The calculation parameters “Neighbor Joining” using BLOSUM62 were used in creating the tree. Neighbor joining is a method for constructing phylogenetic trees that was first proposed by Saitou and Nein in 1987. The method attempts to find relationships that minimize the total branch length at each stage of the star-like tree (Saitou and Nei, 1987).

3.2: Comparison Between Strains of EEEV by Year

A comparative analysis of the domains was performed on all EEEV sequences from the year 2014 and the year 2016, because there are several strains available for these years, allowing for the largest possible sample size. The sequences were obtained from RVDB per the protocol described in section 3.1, filtering the results to choose strains from these specific years for analysis.

After navigating to the GanBank results for each selected strain, the FASTA sequences were obtained for the structural polyprotein and compiled into two FASTA files: one for each year containing all of the sequences for that year. See Table 1 below for a summary of the sequences analyzed using the GenBank protein identifiers. The first entry in each comprehensive FASTA file was the sequence for the structural polyprotein of the reference genome. The resulting FASTA files were uploaded to the T-Coffee Expresso service for structural alignment. Mutations between sequences were noted in a separate table and analyzed according to the properties of each residue (Di Tommaso, 2011).

Table 1: Structural polyprotein sequences under study for species comparison.
Protein identifier number and year of structural polyprotein sequences used to conduct sequence-based analysis across different years (Benson, 2004).

Protein Identifiers		
2014		**2016**
AHL83687.1	AHL83769.1	AMT80038.1*
AHL83721.1	AHL83707.1	AMT80100.1*
AHL83711.1	AHL83789.1	AMT80088.1
AHL83653.1	AHL83695.1*	AMT80058.1*
AHL83727.1	AHL83735.1	AMT80016.1
AHL83739.1	AHL83799.1	AMT79966.1
AHL83719.1	AHL83649.1	AMT79954.1

AHL83635.1	AHL83753.1	AMT79998.1
AHL83791.1*	AHL83731.1	AMT80296.1
AHL83793.1	AHL83779.1	AMT79990.1
AHL83755.1*	AHL83787.1	AMT79992.1
AHL83743.1		AMN91567.1
AHL83655.1		AMN91599.1
AHL83781.1		AMN91497.1
AHL83679.1		AMN91617.1
AHL83667.1		AMN91521.1

* Sequences chosen for analysis between years (see explanation below)

Next, a comparison was done using 3 representative sequences for each of the two years. See Table 1 above for the chosen sequences, noted with an asterix. The representative sequences were selected for having the most residues that were different than the reference and each other to make sure the diversity of the sequences was best captured in the analysis (Di Tommaso, 2011). The 6 sequences chosen plus the reference were added to a new FASTA file, which was uploaded to the T-Coffee Expresso service for structural alignment. The resulting alignment was visually analyzed for patterns in mutations between the two years, and was uploaded to Jalview for secondary structure and phylogenetic analysis.

3.3: Comparison Between Strains of EEV by Host Species

A comparative analysis was also performed on EEEV sequences that were collected from different host species. These sequences were obtained by running a BLASTN search against the unclustered U-RVDBv17.0 database on the Reference Viral Database (RVDB) (Goodacre et. al., 2018). The query sequence was the EEEV complete reference genome that was derived from the Volchkov et. al. paper on viruSITE. All default settings were accepted except for the expect threshold; this value was changed from 10 to 0.001.

Search results were sorted by percent sequence identity (from highest to lowest). Sequences from birds, horses, humans, and mosquitoes were included for analysis. The three sequences from each species with the highest percentage of sequence identity to the query sequence were selected. The only exception to this rule was for humans; only one sequence was included for the human species.

RVDB provided links to the GenBank page for each sequence (Benson, 2004). This link was used to navigate to the GenBank page for the sequence of interest. From here, the translated structural polyprotein sequences were saved as individual files in FASTA format. See Table 2 below for a summary of the sequences selected for analysis.

Table 2: Structural polyprotein sequences under study for species comparison.
Protein identifier number, class, and species of structural polyprotein sequences used to conduct sequence-based analysis across different host species.

Protein Identifier	Class	Species
AHL83719.1	Bird	Vireo olivaceus
AHL83791.1		Phasianus
AMT80016.1		Phasianus colchicus
AHL83687.1	Horse	Equus ferrus caballus
AHL83711.1		Equus ferrus caballus
AHL83721.1		Equus ferrus caballus
AHL83727.1	Human	Homo sapiens
AHL83755.1	Mosquito	Culex restuans
AMT79966.1		Culiseta
AMT80088.1		Culiseta

Next, each sequence in FASTA format was compiled into one text file. The text file was uploaded to the T-Coffee Expresso service for structural alignment. Mutations between sequences were noted in a separate table and analyzed according to the properties of each residue, and the alignment was uploaded to Jalview for secondary structure and phylogenetic analysis.

3.4: Comparison Between Alphaviruses

A comparative sequence-based analysis was also performed across the Alphavirus classification. Sequences that corresponded to Eastern equine encephalitis, Western equine encephalitis, Venezuelan equine encephalitis, Sindbis virus, Semliki forest virus, Chikungunya virus, and Ross River virus were included in the comparison. For each virus, the full reference genome was obtained by searching for the name of the virus in the viruSITE Keyword Search (Stano et. al., 2016). The full nucleotide sequence of the virus was saved as an individual file in FASTA format. See Table 3 below for information on each full genome sequence selected for analysis.

Table 3: Complete reference genomes used for Alphavirus comparison.

Summary of complete reference genomes used for comparison across Alphavirus classification. The name of each virus/the search term used on viruSITE is listed in the first column. The most recent papers that cited each genome and the year that those papers were published are listed in the second column.

Virus/Search Term on Virusite	Authors of Most Recent Paper and Year
Eastern equine encephalitis (EEV)	Volchkov V.E., Volchoka V.A., Netesov S.V. (1991)
Western equine encephalitis (WEEV)	Netolitzky *et. al.* (2000)
Venezuelan equine encephalitis (VEEV)	Li *et. al.* (2016)
Sindbis virus	Ramsey J., Renzi E.C., Arnold R.J., Trinidad J.C., Mukhopadhyay S. (2017)
Semliki forest virus	Schulte T. *et. al.* (2017)
Chikungunya virus	Tossavainen H., Aitio O., Hellman M., Saksela K., Permi P. (2017)
Ross River virus	Faragher S.G., Meek A.D., Rice C.M., Dalgarno L. (1988)

The protein translation of each genome sequence was obtained by running BLASTN searches against the unclustered U-RVDBv17.0 database on the Reference Viral Database (RVDB) (Goodacre et. al., 2018). One search was run for each virus; the complete reference genome downloaded from viruSITE for each genome served as the query sequence (Stano et. al., 2016). All default settings except for the expect threshold were accepted; this value was changed from 10 to 0.001. The first result for each BLASTN search corresponded to the complete genome of the virus. When the link was followed to the GenBank profile for the genome, the translated structural polyprotein sequences were saved as individual files in FASTA format (Benson, 2004).

Each of the individual files was consolidated into one text file in FASTA format. The text file was uploaded to the T-Coffee Expresso service for multiple sequence alignment (Di Tommaso, 2011). The resulting alignment was uploaded to Jalview for phylogenetic analysis to determine evolutionary relationships between the different viruses.

4: Results

The T-Coffee results were analyzed by observation of the residues that differed from the reference. T-Coffee indicates the potential impact of polymorphisms by assigning one of four symbols below the alignment. An asterisk (*) indicates that the position corresponds to a residue that is fully conserved across all sequences (Di Tommaso, 2011). A colon (:) indicates that while residues may not match in that position, the properties of the residues are similar, scoring over 0.5 in the Gonnet PAM 250 matrix. A period (.) indicates that residues do not match in the position. Furthermore, the residues of each sequence do not have similar properties, scoring under 0.5 in the Gonnet PAM 250 matrix. Finally, a blank space ( ) indicates that residues may be different and have dramatically different properties; it also may indicate that a gap exists (Di

Tommaso, et.al 2011). These symbols denoted at the bottom of each residue of the aligned sequences were used to help in the determination of the potential effect of the residue change on the EEEV structural protein. The specific residue changes were also noted and summarized.

Then, Jalview was used to run a secondary structure-based analysis of the alignment using the JPred 4 web tool. The conservation scores assigned to residues with low conservation were used to determine which local structures are most likely to differ amongst the different aligned strains. Jalview was also used to run used to construct a phylogenetic tree as an alternative way to analyze the conservatory evolution of the strains between years, alphaviruses, and species.

4.1: Comparison Between Strains of EEEV by Year

The multiple sequence analysis that was performed within the years 2014 and 2016 both closely resembled the alignments of the models generated in PDB. T-Coffee reported high levels of consistency between the final alignments and the libraries derived from PDB 3D structures; each sequence had a total consistency value of 99. All residues were highlighted in red, indicating that the alignment produced by T-Coffee was strongly correlated with the alignment produced from the templates in the PDB 3D structure library (Di Tommaso, 2011).

Although the majority of the sequences were highly conserved, there were several local differences in residues in some of the strains that could affect the produced protein. The residues observed to be different than those of the reference template mainly affect the charge, rigidity, and polarity of the protein (Voet et.al 2016).

Within the sequences aligned from 2014, 1,259 loci were analyzed when generating the multiple sequence alignment. There were 39 single nucleotide polymorphisms (SNPs) observed across loci. Of the SNPs observed, 9 were observed on Chain A and 26 were observed on Chain B of the E1 glycoprotein, and the remaining were observed within regions of the structural polyprotein outside of Chains A and B.

The discrepancies in the residues from the reference template in E1 Glycoprotein Chain A of the strains from 2014 are partially summarized in Table 4 below (See Appendix A for the full summary). Of the 9 SNP's observed, 3 of the T-Coffee evaluation symbols were blank, indicating the potential to have great effect on the protein, 1 was a period, indicating the residues do not match, and 5 were colons, indicating there is not likely to be a great effect on the protein. Of the blanks, two of them were changes that affect polarity, while the third is a change from Proline to Leucine that affects rigidity of the protein. The period symbol was assigned to protein ID AHL83653.1, which had a change from Lysine to Threonine that could affect both charge and shape of the protein (Voet et. al 2016).

Table 4: Discrepancies between residues in Chain A of E1 glycoprotein for 2014: Summary of the discrepancies in residues between sequences in the alignment across strains from 2014. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et. al 2016). These results are partial to get a general idea of what the results look like. The full table of analysis can be found in Appendix A.

Protein ID	TCoffee Evaluation	Residue Change	Potential Effect (General)	Potential Effect (Specific)
AHL83735.1	BLANK	Leucine instead of Proline	Rigidity	Leucine is much less rigid than the cyclic proline, which could affect the flexibility of the protein.
AHL83791.1	:	Glutamine instead of Histidine	Charge, rigidity	Histidine is positively charged, while glutamine is uncharged. Additionally, histidine is cyclical and more rigid.
AHL83719.1	:	Isoleucine instead of Valine	None	Isoleucine has one more C, making it a little bulkier, but overall, there is no major difference.
AHL83653.1	.	Threonine instead of Lysine	Charge, shape	Threonine is uncharged and bulky, while lysine is long and positively charged. Interactions between neighboring residues will likely change, affecting the shape of the protein.fi
AHL83791.1	BLANK	Threonine instead of Isoleucine	Polarity	Isoleucine is nonpolar, while threonine is polar.

The discrepancies in the residues from the reference template in E1 Glycoprotein Chain B of the strains from 2014 are partially summarized in Table 5 below (See Appendix B for the full summary). Of the 21 unique SNPs observed, 8 of the T-Coffee evaluation symbols were blank, indicating the potential to have great effect on the protein, 4 were a period, indicating the residues do not match, and 9 were colons, indicating there is not likely to be a great effect on the protein. Most of the observed differences from the template for this year are in residues of this chain for this, and most of the changes affect charge. Many of the blanks are also associated with a change in polarity, which could potentially have a great effect on the protein (Voet et. al 2016).

Table 5: Discrepancies between residues in Chain B of E1 glycoprotein for 2014: Summary of the discrepancies in residues between sequences in the alignment across strains from 2014. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et. al 2016). These results are partial to get a general idea of what the results look like. The full table of analysis can be found in Appendix B.

Protein ID	TCoffee Evaluation	Residue Change	Potential Effect (General)	Potential Effect (Specific)
AHL83695.1	.	Threonine instead of Lysine	Charged, size	Threonine is uncharged and bulky, while lysine is long and positively charged.
AHL83755.1 AHL83743.1	.	Serine instead of Serine	Protein Interactions	Serine does not have the amine group or carboxyl groups that glutamine has, but serine has a hydroxyl group.
AHL83649.1	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged. Histidine is also more rigid in its cyclical nature.
AHL83695.1	:	Lysine instead of Glutamic Acid	Charge	Glutamic is negatively charged, while lysine is negatively charged.
AHL83727.1	BLANK	Arginine instead of Histidine	Shape	Histidine is cyclic and arginine is long and branched, but they are otherwise very similar.

Within the sequences aligned from 2016, 1,259 loci were analyzed when generating the multiple sequence alignment. There were 27 single nucleotide polymorphisms (SNPs) observed across loci. Of the SNPs observed, 8 were observed on Chain A and 17 were observed on Chain B of the E1 glycoprotein, and the remaining were observed within regions of the structural polyprotein outside of Chains A and B.

The discrepancies in the residues from the reference template in E1 Glycoprotein Chain A of the strains from 2016 are fully summarized in Table 6 below. Of the 7 unique SNPs observed, 2 of the T-Coffee evaluation symbols were blank, indicating the potential to have great effect on the protein, and 5 were colons, indicating there is not likely to be a great effect on the protein. Most of the observed residue differences from the template are associated with a change in charge or polarity, as are both changes between Threonine and Isoleucine in which T-Coffee produced a blank symbol (Voet et. al 2016).

Table 6: Discrepancies between residues in Chain A of E1 glycoprotein for 2016: Summary of the discrepancies in residues between sequences in the alignment across strains from 2016. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID	TCoffee Evaluation	Residue Change	Possible Effect (General)	Possible Effect (Specific)
AMN91521.1	:	Valine instead of Isoleucine	Slightly size	Isoleucine has one more C, making it a little bulkier, but overall, there is no major difference.
AMT80058.1 AMT80296.1	:	Glutamine instead of Histidine	Charge, rigidity	Histidine is positively charged, while glutamine is uncharged. Additionally, histidine is cyclical and more rigid
AMT79992.1	:	Asparagine instead of Lysine	Charge	Lysine is positively charged, while Asparagine is uncharged.
All	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged. Histidine is also more rigid due to its cyclical nature.
AMT80058.1	BLANK	Threonine instead of Isoleucine	Polarity	Isoleucine is nonpolar, while Threonine is polar.
AMT80038.1	:	Threonine instead of Alanine	Polarity	Alanine is nonpolar, while Threonine is polar.
AMT80058.1	BLANK	Isoleucine instead of Threonine	Polarity	Isoleucine is nonpolar, while Threonine is polar.

The discrepancies in the residues from the reference template in E1 Glycoprotein Chain B of the strains from 2016 are partially summarized in Table 7 below. Of the unique SNPs observed, of the T-Coffee evaluation symbols were blank, indicating the potential to have great effect on the protein, and 5 were colons, indicating there is not likely to be a great effect on the protein. As was true for Chain A for 2016 strains, most of the observed residue differences from the template are associated with a change in charge. However, the blanks are associated with different changes. All of the strains were different from the reference in a change from Isoleucine to Threonine, which has a great potential to cause a change in polarity (Voet et. al 2016). The other blank is associated with a gap that occurred instead of Phenylalanine in proteins of ID

AMT80100.1. AMT80088.1, andAMT80016.1. This gap is likely to cause significant changes in the protein's final folded shape and potentially its interactions with other molecules, as was confirmed later in the JPred analysis.

Table 7: Discrepancies between residues in Chain B of E1 glycoprotein for 2016: Summary of the discrepancies in residues between sequences in the alignment across strains from 2016. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et. al 2016). These results are partial to get a general idea of what the results look like. The full table of analysis can be found in Appendix C.

Protein ID	TCoffee Evaluation	Residue Change	Potential Effect (General)	Potential Effect (Specific)
AMT80100.1	:	Aspartic Acid instead of Asparagine	Charge	Asparagine is not charged, while aspartic acid is negatively charged. This could affect interactions with other residues and protein function.
AMT79998.1	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged.
All	BLANK	Isoleucine instead of Threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.
AMT80100.1 AMT80088.1 AMT80016.1	BLANK	Gap instead of Phenylalanine	Folding and protein interactions	Could cause change in folding and protein interactions due to missing residue.
AMT80296.1	:	Isoleucine instead of Valine	Slightly size	Isoleucine has one more C, making it a little bulkier, but overall, there is no major difference.

The multiple sequence analysis that was performed to compare 3 representative sequences from 2014 and 2016 showed that the structural polyproteins are very highly conserved between the two years. Figure 1 below shows the total consistency values generated by T-Coffee, where all sequences aligned produced a value of 99.

Figure 1: Consistency scores for each sequence of the alignment and a 32 residue example of the alignment results. Each sequence had a consistency score of 99, indicating high levels of similarity between the sequence and the template in the PDB library. There were few observed residue changes, and the only pattern observed was a similarity between AMT80058.1 and AHL83791.1 (Di Tommaso, 2011).

It was found that most of the residue differences from the reference that were the same between other strains were in proteins with IDs AMT80058.1 and AHL83791.1. Both of these proteins come from the pheasant host, which is likely the reason for the similarities between the two (Goodacre et. al., 2018). Otherwise, there was no pattern between the observed differences.

An analysis of the secondary structure conservation completed using the JPred web service through Jalview showed that the structure of the proteins were also highly conserved between the strains compared between years. Figure 2 below shows a small portion of the secondary structure analysis.

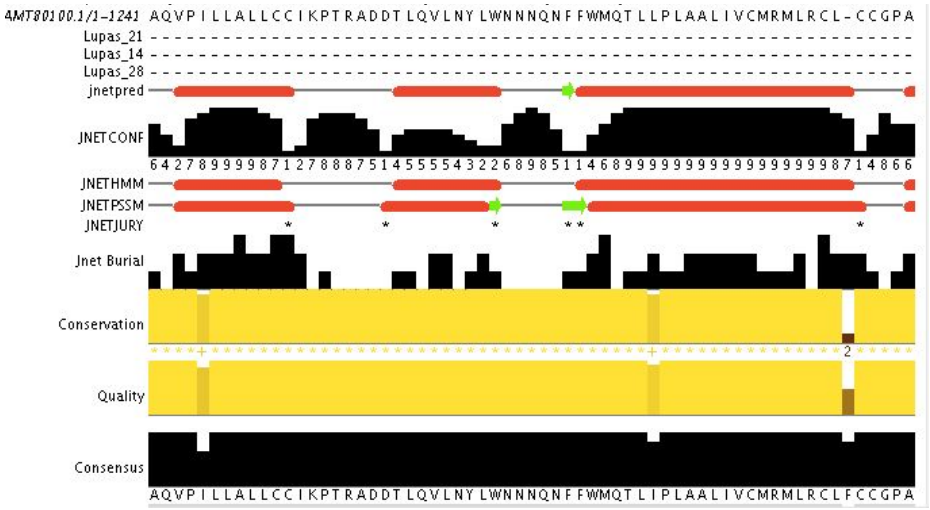

Figure 2: JPred secondary structure analysis of compared EEEV strains from the years 2014 and 2016. Alpha helices are represented by green arrows and beta sheets are represented by red lines. The conservation of the structure within the aligned sequences is also represented by a conservation score noted in yellow **(**Drozdetskiy, 2015). Most of the residues that were different from the reference showed conservation denoted with a "+", similar to the first two shown above, or a score of 8-9. However, the gap observed in some of the 2016 Chain B sequences produced a very low conservation score of 2, as shown in the far right above.

Based on the secondary structure analysis results, most of the residue differences from the reference will have little to no effect on the structure, as they received high conservation scores. However, the introduction of a gap in three of the 2016 Chain B sequences is likely to have a great effect on the secondary structure of the protein for these strains, as given by the low conservation score of 2. The local structure likely to be affected in this protein is where the end of a beta sheet meets the beginning of a loop structure. Otherwise, the secondary structure of all structural polyproteins compared between the years 2014 and 2016 are highly conserved.

A phylogenetic analysis of all sequences from both years was completed in Jalview to also show the relationship between the strains. Due to their similarity, nearly all of the proteins were placed on the same branch as each other, as can be seen in Figure 3 below.

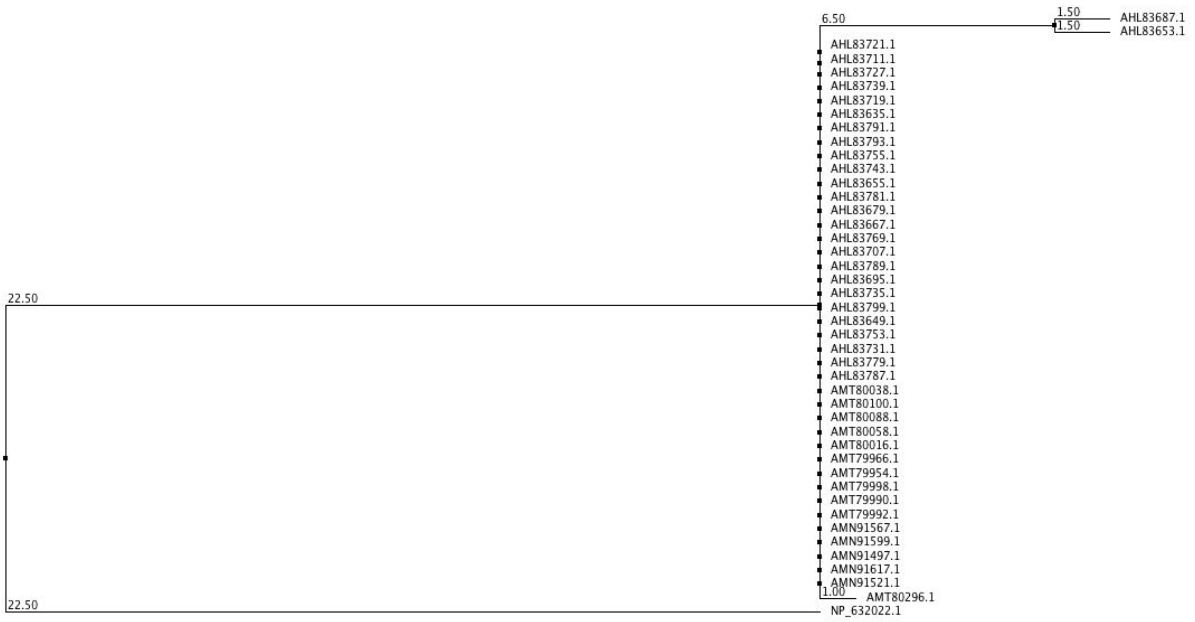

Figure 3: Neighbor-joining phylogenetic tree of all sequences from 2014 and 2016 based on calculated scores using the BLOSUM62 matrix. Based on the resulting branches, all of the proteins are very closely related to each other, with AHL83687 and AHL83653 having diverged from the common ancestor in evolution more than the rest. The reference sequence is equal in distance from the common ancestor of the other proteins, but has its own branch.

Proteins AHL83687 and AHL83653 were branching off of the branch containing the rest of the sequences, suggesting they have accumulated more changes from the common ancestor of the strains. Both of these proteins came from strains of the virus isolated in Eastern New Jersey in 2014 from equine, explaining their similarities due to relatedness (Goodacre et. al., 2018)..

4.2: Comparison Between Strains of EEEV by Host Species

The multiple sequence analysis that was performed between different host species closely resembled the alignment of the models generated in PDB. T-Coffee reported high levels of consistency between the final alignment and the library derived from PDB 3D structures; each sequence had a total consistency value of 99. Higher consistency values correspond to higher levels of similarity between the sequence that was entered and the corresponding sequence in the PDB library. All residues were highlighted in red, indicating that the alignment produced by T-Coffee was strongly supported by the alignment produced with the templates in the PDB 3D structure library. Figure 4 below displays the consistency scores assigned to each sequence and the first row of the sequence alignment.

```
T-COFFEE, Version_11.00.d625267 (2016-01-11 15:25:41 - Revision d625267 - Build 507)
Cedric Notredame
SCORE=99
*
 BAD  AVG  GOOD
*
AHL83719.1       :  99
AHL83791.1       :  99
AMT80016.1       :  99
AHL83687.1       :  99
AHL83711.1       :  99
AHL83721.1       :  99
AHL83727.1       :  99
AHL83755.1       :  99
AMT79966.1       :  99
AMT80088.1       :  99
AHL83727.1_1     :  99
cons             :   9

AHL83719.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AHL83791.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AMT80016.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AHL83687.1       MFPYPTLNYPPMASINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AHL83711.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AHL83721.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AHL83727.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AHL83755.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AMT79966.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AMT80088.1       MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA
AHL83727.1_1     MFPYPTLNYPPMAPINPMAYRDPNPPRRRWRPFRPPLAAQIEDLRRSIANLTLKQRAPNPPAGPPA

cons             *************.**************************************************
```

Figure 4: Consistency scores for each sequence and the first row of the alignment. Each sequence had a consistency score of 99, indicating high levels of similarity between the sequence and the template in the PDB library. All sequences were identical in the first row with the exception of one residue; AHL83687.1 had serine where the other sequences had proline (Di Tommaso, 2011).

While most of the EEEV sequences derived from different host species were highly conserved, some local variations were observed. There were 1,259 loci analyzed when generating the multiple sequence alignment. There were 17 single nucleotide polymorphisms (SNPs) observed on 14 loci. Of the 17 SNPs, 4 were observed on Chain A and 7 were observed on Chain B of the E1 glycoprotein. The remaining 6 SNPs were associated with other structural polyproteins.

Of the four residue discrepancies observed within Chain A, two instances were assigned a blank space in the T-Coffee analysis. The remaining two instances were assigned colons. See Table 8 below for a summary of the discrepancies observed within Chain A. The sequence with the Protein ID of AHL83791.1 was the only one with discrepancies in Chain A; the host species was a type of bird.

Table 8: Discrepancies between residues in Chain A of E1 glycoprotein: Summary of the discrepancies in residues between sequences in the alignment. Sequence AHL83791.1 hosted all four residue discrepancies; the rest of the sequences were identical on all loci. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change, the specific residue change, and the potential general and specific effects of the change (Di Tommaso, 2011).

Protein ID/Species	T-Coffee Evaluation	Residue Change	Potential Effect: General	Potential Effect: Specific
AHL83791.1 (Bird)	:	Glutamine instead of histidine	Charge, rigidity	Histidine is positively charged, while glutamine is uncharged. Additionally, histidine is cyclical and more rigid.
AHL83719.1 (Bird)	:	Isoleucine instead of valine	Size	Isoleucine has one more carbon atom, making it a little bulkier, but overall, there is no major difference.
AHL83791.1 (Bird)	BLANK	Threonine instead of isoleucine	Polarity	Threonine is polar, while isoleucine is nonpolar.
AHL83791.1 (Bird)	BLANK	Isoleucine instead of threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.

Jalview was used to annotate the multiple species alignment according to the conservation of secondary structure across each sequence. Although the change from glutamine to histidine observed in the first row of Table 8 occurred within a loop structure, the discrepancies observed on the remaining three loci of Chain A were correlated with changes in both alpha helices and beta sheets. The change from isoleucine to valine in AHL83719.1 occurred within an alpha helix structure. However, the software rated the conservation of the sequence with a score of 9 out of 10 possible points. This indicates that the effect of the discrepancy on the overall function of the protein is likely small. This evaluation supports the potential effect predicted in Table 8; although the size of the protein may be slightly affected by the substitution, the overall change in function is likely small.

The first discrepancy observed within AHL83791.1 on Chain A occurred within a beta sheet structure. The change from threonine to isoleucine was assigned a gap on the T-Coffee alignment, suggesting that the properties of the residues were dramatically different. In this case, the discrepancy between residues may lead to a change in polarity: threonine is nonpolar, and isoleucine is polar. This is reflected in the conservation score assigned to the matrix; the locus scored 7 out of 10 possible points. The second discrepancy observed within AHL83791.1 on Chain A occurred within an alpha helix. Since the discrepancy was also observed between threonine and isoleucine, the locus was assigned a score of 7 out of 10 possible points.

See Figure 5 below for a portion of the secondary structure analysis of the multiple sequence alignment of Chain A. The protein sequences are aligned at the top of the feature. The residues that match within each locus are highlighted in blue; residues that do not agree with the rest of the sequences are not highlighted. The green arrows at the bottom of the figure denote which portions of the sequence are associated with beta sheets; the red tubes denote which portions are associated with alpha helices. The bars that are located in the row labeled “Conservation” evaluate the conservation between sequences.

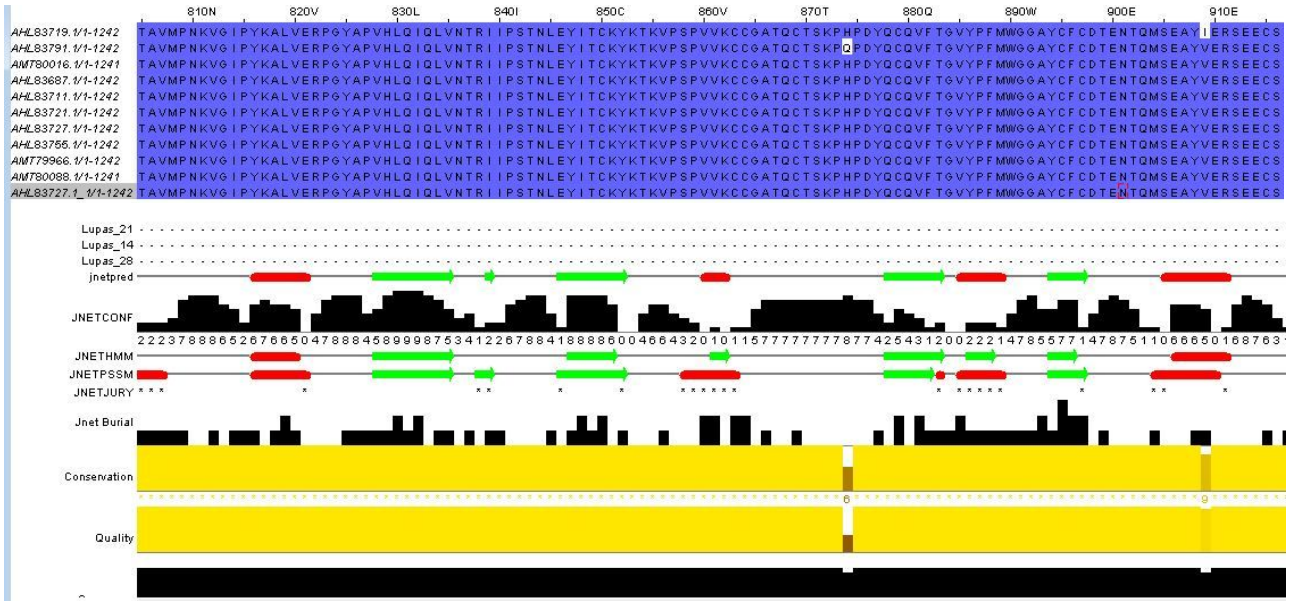

Figure 5: Secondary structure prediction for Chain A of E1 glycoprotein: JPred secondary structure analysis of strains of EEEV derived from a variety of host species. Alpha helices are represented by green arrows and beta sheets are represented by red lines. The conservation of the structure within the aligned sequences is also represented by a conservation score noted in yellow (Drozdetskiy, 2015). Although the discrepancy pictured on the left was observed in a loop structure rather than an alpha helix or beta sheet, it was given 6 out of 10 possible points for conservation. The discrepancy pictured on the right was observed within an alpha helix structure, but it was given 9 out of 10 possible points for conservation.

Of the ten residue discrepancies observed within Chain B, seven instances were assigned a colon. This indicated that the properties of the residue on each sequence were similar. The next discrepancy was assigned a period to indicate the the residue properties were not conserved. Two instances were assigned gaps; this indicated that there was a significant difference between the residues observed on each sequence. These blank spaces corresponded to the only gaps that were inserted into the alignment. See Table 9 below for a summary of the residue discrepancies observed in Chain B.

Table 9: Discrepancies between residues in Chain B of E1 glycoprotein: Summary of the discrepancies in residues between sequences in the alignment. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID/Species	T-Coffee Evaluation	Residue Change	Potential Effect: General	Potential Effect: Specific
AMT80016.1 (Bird) AMT80088.1 (Mosquito)	BLANK	Gap instead of phenylalanine	Size	Phenylalanine has a nonpolar, cyclical side chain that is fairly bulky.
AHL83755.1 (Mosquito)	.	Serine instead of histidine	Charge, rigidity	Serine has an uncharged polar side chain; histidine has a charged polar side chain. Histidine has a cyclical side chain, so it may be more rigid than serine as a result.
AHL83727.1 (Human)	:	Arginine instead of histidine	Rigidity, size	Both residues are nitrogen-containing amino acids with charged polar side chains. Histidine is cyclical and more rigid than the linear arginine molecule. Arginine is also heavier than histidine.
AHL83791.1 (Bird)	:	Methionine instead of leucine	None	Both residues are very similar in properties, so the effect should be small
AHL83755.1 (Mosquito)	:	Phenyalanine instead of leucine	Rigidity	Phenylalanine is cyclical and more rigid.
AHL83791.1 (Bird)	:	Leucine instead of isoleucine	None	Both residues are very similar in properties, so the effect should be small
AMT80016.1 (Bird) AHL83755.1 (Mosquito) AMT80088.1 (Mosquito)	:	Leucine instead of isoleucine	None	Both residues are very similar in properties, so the effect should be small

Six of the ten residue discrepancies observed within Chain B were associated with alpha helix structures. Two residue discrepancies occurred within a beta sheet, and the remaining two were observed within loop structures. The change from histidine to arginine on AHL83727.1 occurred just one residue away from a predicted beta sheet structure. The JPred software assigned this prediction a confidence score of 6; it is possible that this residue could also be associated with a beta sheet structure.

Most of the residue discrepancies were labeled with high conservation scores of 8, 9, or a "+." However, the gaps that were inserted into the sequence rather than phenylalanine were assigned a conservation score of two, indicating that the properties of the original chain were not conserved. Although JPred predicted that this residue would be located on a loop structure, it is located just one locus away from a region that was predicted to be an alpha helix. JPred assigned just 1 out of 10 points for confidence that its evaluation was correct at this point; this residue could be a part of the alpha helix. See Figure 6 below for a screenshot of the secondary structure analysis for the locus that contained the gaps.

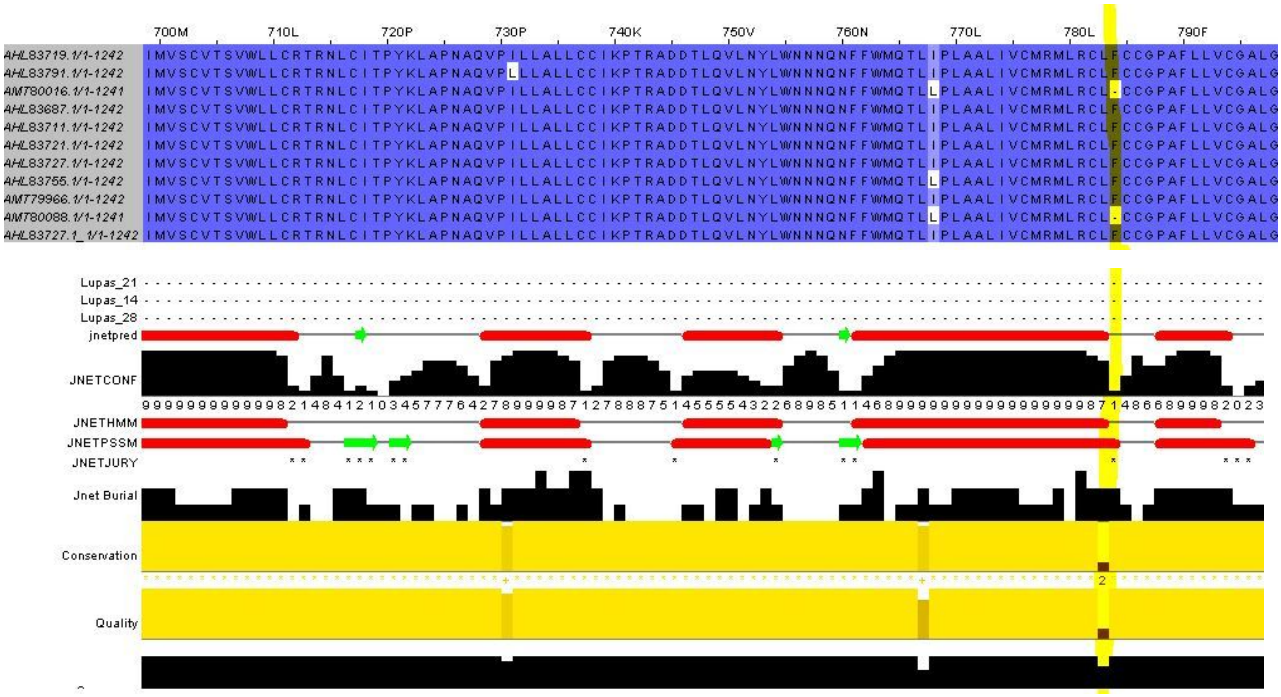

Figure 6: Secondary structure prediction for Chain B of E1 glycoprotein: JPred secondary structure analysis of strains of EEEV derived from a variety of host species. Alpha helices are represented by green arrows and beta sheets are represented by red lines. The conservation of the structure within the aligned sequences is also represented by a conservation score noted in yellow (Drozdetskiy, 2015). The locus highlighted in yellow contained two instances in which a gap was inserted rather than phenylalanine; this locus was assigned a conservation score of 2 out of 10 possible points.

Of the remaining four discrepancies observed in the structural polyprotein region, one instance was assigned a period that indicated different properties between residues on one loci; three instances were assigned colons that indicated similar properties. See Table 10 below for a summary of the discrepancies observed that did not correspond to Chain A or Chain B of the E1 glycoprotein.

Table 10: Discrepancies between residues in structural polyprotein region outside Chains A and B of the E1 glycoprotein: Summary of the discrepancies in residues between sequences in the alignment. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID/Species	T-Coffee Evaluation	Residue Change	Potential Effect: General	Potential Effect: Specific
AHL83687.1 (Horse)	.	Serine instead of proline	Polarity	Serine has an uncharged polar side chain while proline has just an uncharged side chain. This could affect interactions with other residues and protein function. Serine is more flexible than the cyclic proline residue, which could affect function.
AHL83687.1 (Horse)	:	Isoleucine instead of methionine	Size	Isoleucine is a little bit bulkier, but overall they are very similar in properties.
AMT79966.1 (Mosquito)	:	Isoleucine instead of valine	Size	Isoleucine has one more carbon atom, making it a little bulkier, but overall, there is no major difference.
AHL83755.1 (Mosquito)	:	Asparagine instead of aspartatic acid	Charge	Asparagine is not charged, while aspartic acid is negatively charged. This could affect interactions with other residues and protein function.

Of the four residue discrepancies observed outside Chains A and B of the E1 glycoprotein, one locus was located in a region that JPred predicted would be a loop. Two loci were in regions that were predicted to be beta sheets, and the remaining locus was predicted to be an alpha helix. JPred assigned high conservation scores to all four loci; the locus in the loop region was assigned a 7, both loci in the beta sheet regions were assigned a 9, and the locus in the alpha helix region was assigned an 8.

The sequences associated with the bird host species had seven residue discrepancies across the alignment. One of those discrepancies would likely lead to a difference in charge. Two of the discrepancies would lead to a difference in rigidity, two would lead to a difference in size, two would lead to a difference in polarity, and two would likely have no effect on the properties of the chain.

The sequences associated with the horse host species had two residue discrepancies across the alignment. One discrepancy would lead to a change in polarity of the structural polyprotein; another would lead to a change in size.

The sequences associated with the human host species had one residue discrepancy across the alignment. This discrepancy would likely lead to a difference in rigidity and size of the chain as compared to the other sequences.

The sequences associated with the mosquito host species had six residue discrepancies across the alignment. Two of the discrepancies were associated with charge, two were associated with size,

one was associated with rigidity, and two would likely have no effect on the properties of the polyprotein.

A phylogenetic analysis of the strains between species, run by Jalview, confirms the secondary structure analysis. As shown below in figure 7 below, the host AHL83791.1 is completely branched off from the other host sequences, indicating low conservation between itself and the other host species, which is consistent with discrepancies as discussed above. Unique gap discrepancies shared between AMT80016.1 and AMT80088.1 were also noted above and are present in the tree as those two strains are within their own sub-branch with a greater distance between their closer relative sequences.

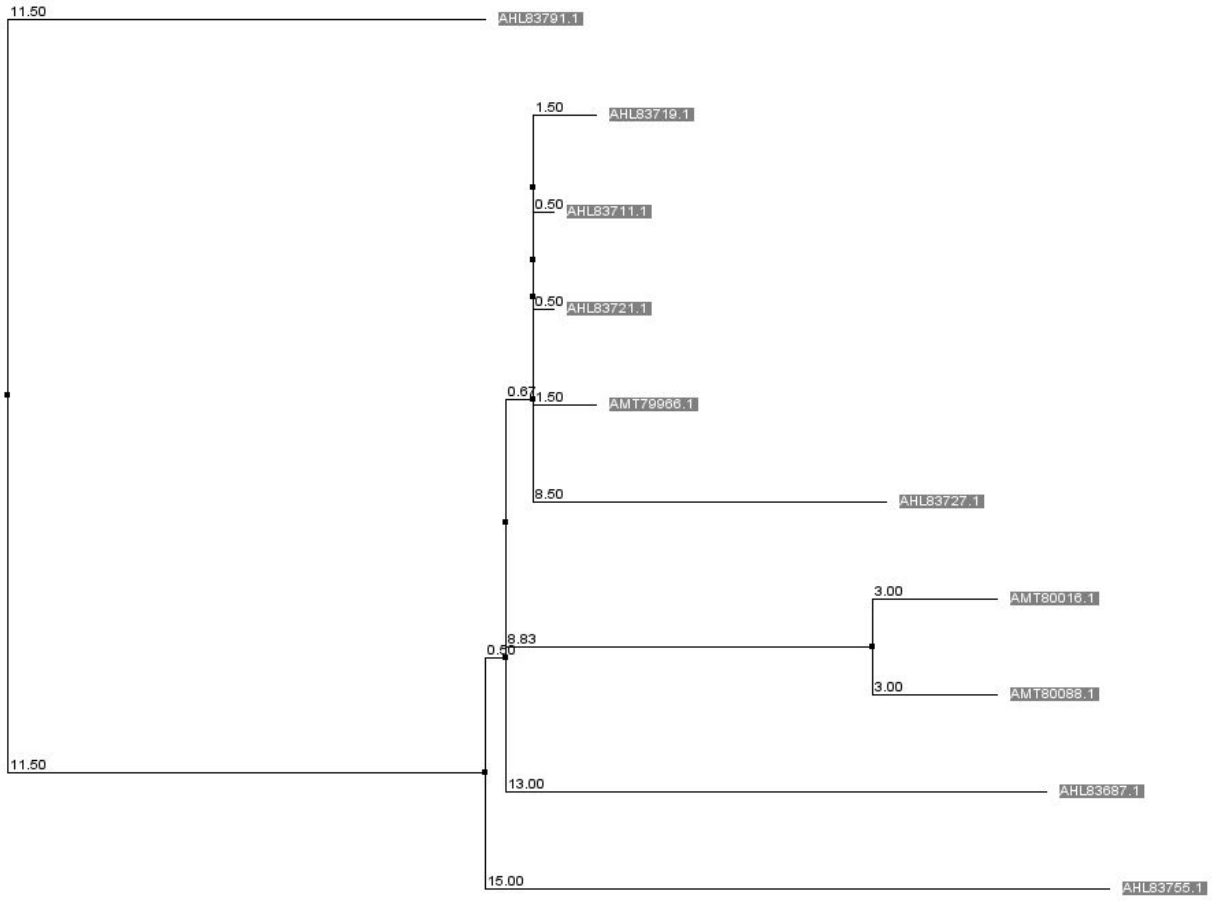

Figure 7: Neighbor-joining phylogenetic tree of host species based on calculated scores using the BLOSUM62 matrix. The resulting branches indicate that the pheasant host, AHL83791.1, diverged substantially from the alternate species, consistent with the discrepancy observed by the T-Coffee alignment, while the other sequences share a more closely related ancestor.

4.3: Comparison Between Alphaviruses

The multiple sequence analysis that was performed across the Alphavirus classification did not resemble the alignment of the models generated in PDB as closely as the year-to-year and host

species analyses. T-Coffee reported lower levels of consistency between the final alignment and the library derived from PDB 3D structures; the consistency values of the sequences ranged from 89 to 93. The Ross River virus, Semliki forest virus, Sindbis virus, and Western equine encephalitis virus (WEEV) had the highest consistency scores. Eastern equine encephalitis virus (EEEV) had the lowest consistency score. The residues at the beginning of the structural polyprotein region were primarily highlighted in yellow and green, indicating that the alignment produced by T-Coffee was not strongly supported by the alignment produced with the templates in the PDB 3D structure library. See Figure 8 below for a screenshot of the alignment produced by T-Coffee for the Alphavirus comparison. The consistency scores are listed above the alignment; the first two rows of the alignment are visible below the scores.

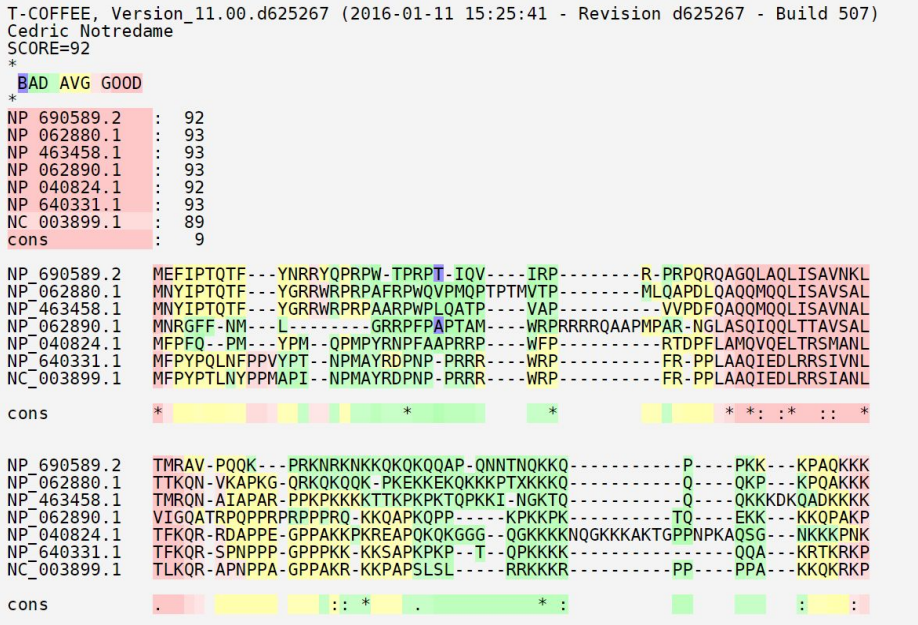

Figure 8: Consistency scores and first two rows of the T-Coffee alignment produced across the Alphavirus classification. The first two rows of the alignment are primarily labeled with blank spaces, indicating that the residues have significantly different properties (Di Tommaso, 2011).

The majority of the loci in the first two rows were labeled with blank spaces by T-Coffee, indicating that the residues in each position were significantly different. The loci in the next three rows were labeled with asterisks, indicating that the section was conserved across the Alphavirus classification. This section See Figure 9 below for a screenshot of the third, fourth, and fifth rows in the T-Coffee alignment.

Figure 9: Third, fourth, and fifth rows of the T-Coffee alignment produced across the Alphavirus classification. The rows are primarily labeled with asterisks, indicating that several loci have residues that are conserved across all sequences in the analysis (Di Tommaso, 2011).

While the next seven rows did have some residues that were conserved across all sequences, the majority of loci were labeled with blank spaces, colons, or periods. The residues were highlighted in red, indicating that the alignment was highly supported by the reference library generated using the PDB 3D database. See Figure 10 below for a sample screenshot of the next seven rows in the alignment.

Figure 10: Representative screenshot of the next seven rows in the alignment. While the asterisks indicate that some portions of the chain are conserved across all sequences in the alignment, the majority of the loci are labeled with blank spaces by T-Coffee (Di Tommaso, 2011).

The remaining rows of the multiple sequence alignment did not indicate any degree of similarity between the aligned polyprotein structures. T-Coffee labeled each locus with a blank space, suggesting that the residues were significantly different in properties. While the residues were highlighted in red for most of the alignment, the highlight color switched to yellow for the end of the alignment. This suggests that the alignment did not conform as readily to that created within the template library and may not be reliable. See Figures 11 and 12 below for screenshots of the end of the alignment on T-Coffee.

Figure 11: End of the multiple sequence alignment produced by T-Coffee for the Alphavirus classification. Each locus towards the end of the alignment was labeled with a blank space (Di Tomaso, 2011).

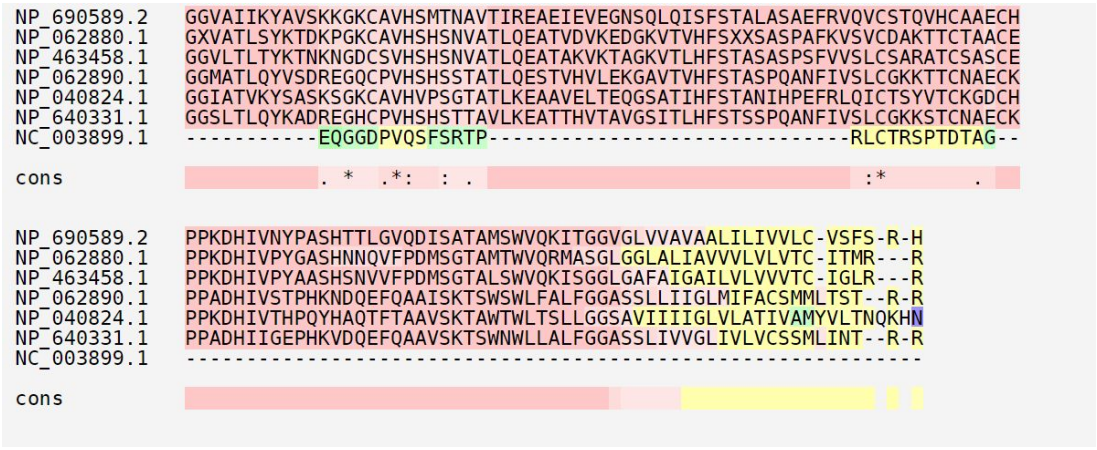

Figure 12: Last two rows of the multiple sequence alignment produced by T-Coffee for the Alphavirus classification. The residues are highlighted in yellow for the last residues, suggesting that the alignment does not conform to the one produced from the PDB 3D library (Di Tommaso, 2011).

The multiple sequence alignment annotations produced by Jalview support this analysis. See Figure 13 below for a screenshot of the annotations. The bars in various shades of yellow and brown represent the conservation and quality observed at each locus. The brown bars represent low levels of conservation across Alphaviruses; the yellow bars represent high levels of conservation. As discussed when analyzing the T-Coffee multiple sequence alignment, the residues at the beginning of the sequence demonstrate low levels of conservation. Most of the bars are dark brown and score less than 5 of 10 possible points for conservation and quality. In the middle of the sequence, there are extended portions that are coded entirely in yellow. The conservation scores within these regions are either 9 or 10. The yellow sequences are bordered by sequences coded in light brown that score between 6 and 8 out of 10 possible points for conservation (Drozdetskiy, 2015).

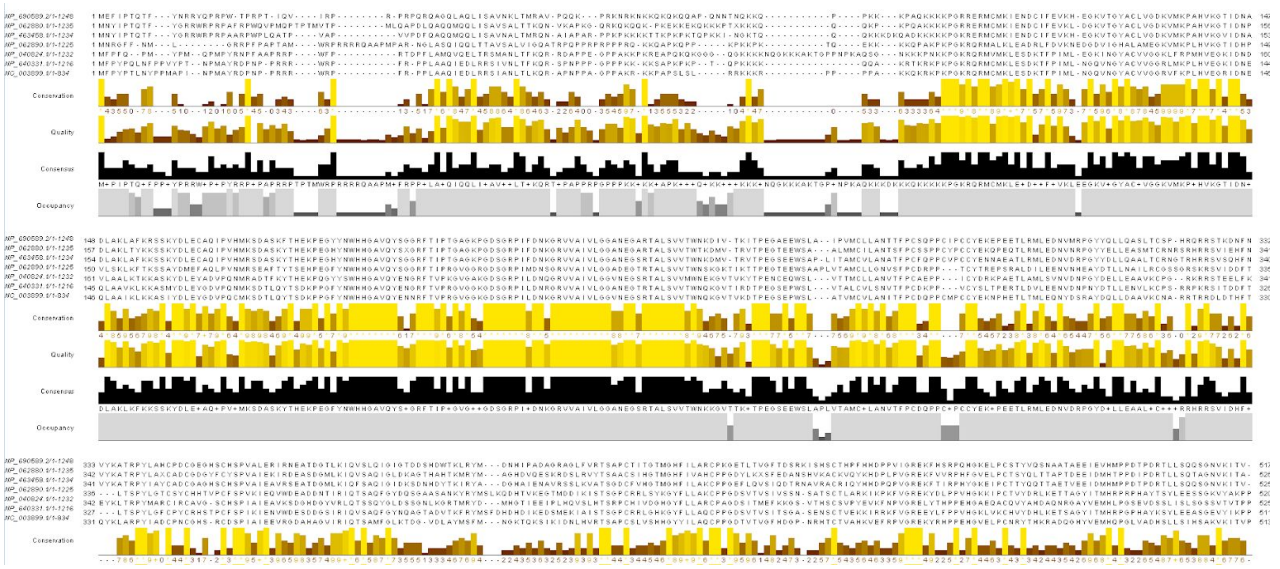

Figure 13: Conservation and quality of loci in multiple sequence alignment between Alphaviruses. The sequences demonstrate low levels of conservation at the beginning of the alignment, high levels of conservation in the middle of the sequence, and low levels of conservation at the end of the sequence.

A phylogenetic tree analysis between the alphaviruses, shown below in figure 14, agrees with the previous findings of low levels of conservation among the alphaviruses. The longer length, or further distance, of the branches indicate that the sequences are not closely related and are further evolved from their relative common ancestor.

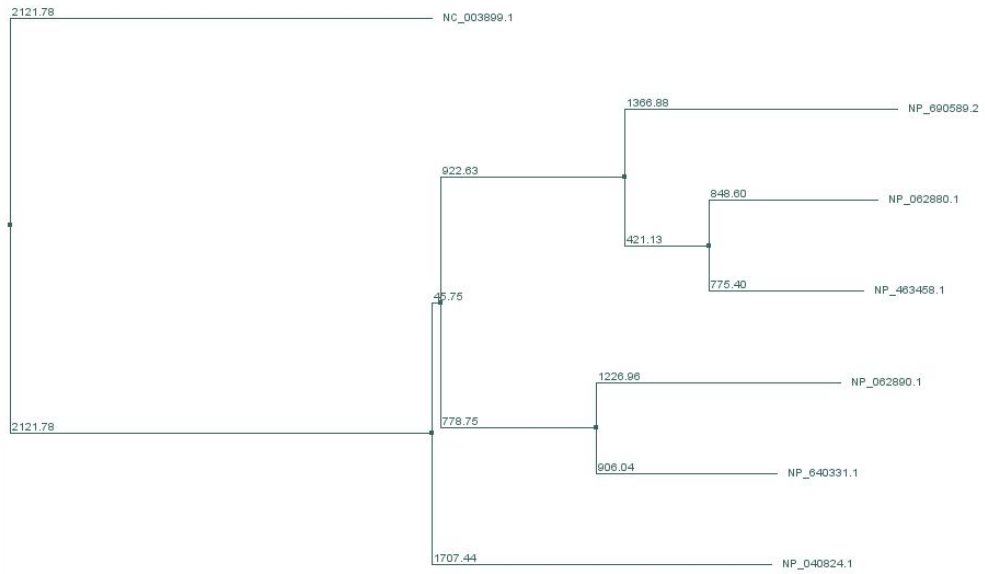

Figure 14: Neighbor-joining phylogenetic tree of alphaviruses based on calculated scores using the BLOSUM62 matrix. The greater lengths in the presented branches indicate lower conservation between the alphaviruses.

5: Discussion & Conclusion

In relation to the different strains of EEEV from 2014 and 2016, a comparative analysis of the domains was performed on all EEEV sequences from these years because there are many strains available, which allows for large populations for sample sizes. In relation to the strains of EEEV within different host species, these strains had a consistency value of 99 when running a multiple sequence alignment. Therefore, these sequences are extremely similar. In relation to the comparison of alphaviruses, the highest consistency scores were found in the Ross River virus, Semliki forest virus, Sindbis virus, and Western equine encephalitis virus (WEEV). Meanwhile, Eastern equine encephalitis virus (EEEV) had the lowest consistency score. Currently, there is no human vaccine for the Ross River virus or WEEV. However, there are vaccines for Semliki forest virus, as well as Sindbis virus.

The goal of this research is to use comparative sequence analysis is to understand genetic differences and potential weaknesses in EEEV. This is completed by comparing strains of EEEV at different points in evolutionary time, across different host species, and similar sequences across different species of viruses. Any weaknesses that are found in these strains can be exploited to create or improve vaccines and other treatments. Differences found between strains can reveal potential effective treatments for strains as they evolve. An advantage from this research is the large amount of sequencing data available. This allows for several ways to

interpret this data. A disadvantage is that when comparing strains of EEEV across different years, the evolutionary history can be incongruent with the genealogy of a single gene.

In terms of future research, comparative sequence based analyses have great potential for combatting EEEV. More research should be focused on the relations of strains across different host species, as well as comparisons between different species that are similar to EEEV.

6: References

BLAST: Basic Local Alignment Search Tool. Bioinformatics.

Benson,D.A. (2004) GenBank, Nucleic Acids Research. Bioinformatics, 33.

CDC (2019) Eastern Equine Encephalitis. Bioinformatics.

Chang,G.-J.J., and Trent,D.W. (1987) Nucleotide Sequence of the Genome Region Encoding the 26S mRNA of Eastern Equine Encephalomyelitis Virus and the Deduced Amino Acid Sequence of the Viral Structural Proteins. Bioinformatics, Journal of General Virology, 68(8), 2129–2142.

Di Tommaso,P. and Moretti,P. et al. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Bioinformatics.

Drozdetskiy,A. and Cole,C. et al. (2015) JPred 4, Nucl. Acids Res. Bioinformatics.

Goodacre,N. and Aljanahi,A. et al. (2018) A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. Bioinformatics, 3, 2.

PD,K. and CL,A. et al. (2003) Comparative sequence analysis of the South African vaccine strain and two virulent field isolates of Lumpy skin disease virus.Bioinformatics, 148, 7.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Bioinformatics, 4, 4.

Saplakoglu,Y. (2019) Why a Rare But Deadly Mosquito-Borne Virus Is Hitting Massachusetts So Hard. Bioinformatics.

Stano, M. and Beke,G et al. (2016) viruSITE—integrated database for viral genomics. Bioinformatics, 2016.

Waterhouse,A.M. and Procter,J.B., et al. (2009) Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench. Bioinformatics, 25, 1189-1191.

Supplemental Materials

Appendix A: Discrepancies between residues in Chain A of E1 glycoprotein for 2014.
Summary of the discrepancies in residues between sequences in the alignment across strains from 2014. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID	TCoffee Evaluation	Residue Change	Potential Effect (General)	Potential Effect (Specific)
AHL83735.1	BLANK	Leucine instead of Proline	Rigidity	Leucine is much less rigid than the cyclic proline, which could affect the flexibility of the protein.
AHL83791.1	:	Glutamine instead of Histidine	Charge	Histidine is positively charged, while glutamine is uncharged.
AHL83719.1	:	Isoleucine instead of Valine	None	Isoleucine has one more C, making it a little bulkier, but overall, there is no major difference.
All	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged.
AHL83791.1	BLANK	Threonine instead of Isoleucine	Polarity	Isoleucine is nonpolar, while threonine is polar.
AHL83653.1	.	Threonine instead of Lysine	Charge, shape	Threonine is uncharged and bulky, while lysine is long and positively charged. Interactions between neighboring residues will likely change, affecting the shape of the protein.
AHL83793.1	:	Serine instead of Threonine	None	There is not much difference in these residues, so there should be little effect on the protein
AHL83793.1	:	Arginine instead of Lysine	None	The only difference is arginine is slightly bulkier.
AHL83791.1	BLANK	Isoleucine instead of Threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.

Appendix B: Discrepancies between residues in Chain B of E1 glycoprotein for 2014.
Summary of the discrepancies in residues between sequences in the alignment across strains from 2014. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID	TCoffee Evaluation	Residue Change	Potential Effect (General)	Potential Effect (Specific)
AHL83695.1	.	Threonine instead of Lysine	Charged, size	Threonine is uncharged and bulky, while lysine is long and positively charged.
AHL83755.1 AHL83743.1	.	Serine instead of Serine	Protein Interactions	Serine does not have the amine group or carboxyl groups that glutamine has, but serine has a hydroxyl group.
AHL83649.1	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged. Histidine is also more rigid in its cyclical nature.

AHL83695.1	:	Lysine instead of Glutamic Acid	Charge	Glutamic is negatively charged, while lysine is negatively charged.
AHL83727.1	BLANK	Arginine instead of Histidine	Shape	Histidine is cyclic and arginine is long and branched, but they are otherwise very similar.
AHL83655.1	BLANK	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged.
All	BLANK	Isoleucine instead of Threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.
All	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged. Histidine is also more rigid in its cyclical nature.
AHL83679.1	:	Arginine instead of Lysine	None	The only difference is arginine is slightly bulkier.
All	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged.
AHL83667.1	BLANK	Tyrosine instead of Aspartic Acid	Charge, size, rigidity	Tyrosine is uncharged and bulky aromatic, while Aspartic acid is charged and a chain.
AHL83769.1	BLANK	Isoleucine instead of Threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.
AHL83755.1 AHL83743.1	:	Phenylalanine instead of Leucine	Rigidity	Phenylalanine is aromatic and very bulky, which could affect the flexibility of the protein.
AHL83791.1	:	Methionine instead of Leucine	None	There are no great differences between these two residues that should affect the structure.
AHL83667.1 AHL83789.1 AHL83695.1	BLANK	Isoleucine instead of Threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.
AHL83753.1	BLANK	Leucine instead of Proline	Rigidity, size	Proline is cyclical and very rigid, while Leucine is branched and much less bulky
AHL83649.1	.	Alanine instead of Valine	None	Both residues are very similar in properties, so the effect should be small.
AHL83791.1	:	Leucine instead of Isoleucine	None	Both residues are very similar in properties, so the effect should be small.
AHL83755.1 AHL83743.1	:	Leucine instead of Isoleucine	None	Both residues are very similar in properties, so the effect should be small.
AHL83707.1	BLANK	Threonine instead of Methionine	Polarity	Methionine is nonpolar, while threonine is polar.
AHL83781.1	.	Valine instead of Alanine	None	Both residues are very similar in properties, so the effect should be small.

Appendix C: Discrepancies between residues in Chain B of E1 glycoprotein for 2016.

Summary of the discrepancies in residues between sequences in the alignment across strains from 2016. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID	TCoffee	Residue Change	Potential Effect	Potential Effect (Specific)

	Evaluation		(General)	
AMT80100.1	:	Aspartic Acid instead of Asparagine	Charge	Asparagine is not charged, while aspartic acid is negatively charged. This could affect interactions with other residues and protein function.
AMT79998.1	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged.
All	BLANK	Isoleucine instead of Threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.
All	:	Histidine instead of Tyrosine	Charge	Tyrosine is uncharged, while histidine is positively charged.
AMT80296.1	:	Isoleucine instead of Valine	Slightly size	Isoleucine has one more C, making it a little bulkier, but overall, there is no major difference
AMT80038.1	:	Tyrosine instead of Histidine	Charge	Tyrosine is uncharged, while histidine is positively charged..
All	:	Histidine instead of Tyrosine	Charge	Tyrosine is uncharged, while histidine is positively charged..
AMN91617.1	:	Alanine instead of Threonine	Polarity	Alanine is nonpolar, while threonine is polar.
AMT80058.1	:	Methionine instead of Leucine	None	There are no great differences between these two residues that should affect the structure.
AMT80058.1 AMT80296.1	:	Leucine instead of Isoleucine	None	Both residues are very similar in properties, so the effect should be small
AMT80100.1 AMT80088.1 AMT80016.1	:	Leucine instead of Isoleucine	None	Both residues are very similar in properties, so the effect should be small
AMT80100.1 AMT80088.1 AMT80016.1	BLANK	Gap instead of Phenylalanine	Folding and protein interactions	Could cause change in folding and protein interactions due to missing residue.

Appendix D: Discrepancies between residues in Chain A of the E1 glycoprotein between host species. Summary of the discrepancies in residues between sequences in the alignment across host species. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID/Species	T-Coffee Evaluation	Residue Change	Potential Effect: General	Potential Effect: Specific
AHL83791.1 (Bird)	:	Glutamine instead of histidine	Charge, rigidity	Histidine is positively charged, while glutamine is uncharged. Additionally, histidine is cyclical and more rigid.
AHL83719.1 (Bird)	:	Isoleucine instead of valine	Size	Isoleucine has one more carbon atom, making it a little bulkier, but overall, there is no major difference.
AHL83791.1 (Bird)	BLANK	Threonine instead of isoleucine	Polarity	Threonine is polar, while isoleucine is nonpolar.
AHL83791.1 (Bird)	BLANK	Isoleucine instead of threonine	Polarity	Isoleucine is nonpolar, while threonine is polar.

Appendix E: Discrepancies between residues in Chain B of the E1 glycoprotein between host species. Summary of the discrepancies in residues between sequences in the alignment across host species. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID/Species	T-Coffee Evaluation	Residue Change	Potential Effect: General	Potential Effect: Specific
AMT80016.1 (Bird) AMT80088.1 (Mosquito)	BLANK	Gap instead of phenylalanine	Size	Phenylalanine has a nonpolar, cyclical side chain that is fairly bulky.
AHL83755.1 (Mosquito)	.	Serine instead of histidine	Charge, rigidity	Serine has an uncharged polar side chain; histidine has a charged polar side chain. Histidine has a cyclical side chain, so it may be more rigid than serine as a result.
AHL83727.1 (Human)	:	Arginine instead of histidine	Rigidity, size	Both residues are nitrogen-containing amino acids with charged polar side chains. Histidine is cyclical and more rigid than the linear arginine molecule. Arginine is also heavier than histidine.
AHL83791.1 (Bird)	:	Methionine instead of leucine	None	Both residues are very similar in properties, so the effect should be small
AHL83755.1 (Mosquito)	:	Phenyalanine instead of leucine	Rigidity	Phenylalanine is cyclical and more rigid.
AHL83791.1 (Bird)	:	Leucine instead of isoleucine	None	Both residues are very similar in properties, so the effect should be small
AMT80016.1 (Bird) AHL83755.1 (Mosquito) AMT80088.1 (Mosquito)	:	Leucine instead of isoleucine	None	Both residues are very similar in properties, so the effect should be small

Appendix F: Discrepancies between residues outside Chains A and B of the E1 glycoprotein between host species. Summary of the discrepancies in residues between sequences in the alignment across host species. The chart summarizes the Protein ID of the sequence with the different residue, the T-Coffee evaluation of the change (Di Tommaso, 2011), the specific residue change, and the potential general and specific effects of the change (Voet et.al 2016).

Protein ID/Species	T-Coffee Evaluation	Residue Change	Potential Effect: General	Potential Effect: Specific
AHL83687.1 (Horse)	.	Serine instead of proline	Polarity	Serine has an uncharged polar side chain while proline has just an uncharged side chain. This could affect interactions with other residues and protein function. Serine is more flexible than the cyclic proline residue, which could affect function.
AHL83687.1 (Horse)	:	Isoleucine instead of methionine	Size	Isoleucine is a little bit bulkier, but overall they are very similar in properties.
AMT79966.1 (Mosquito)	:	Isoleucine instead of valine	Size	Isoleucine has one more carbon atom, making it a little bulkier, but overall, there is no major difference.
AHL83755.1 (Mosquito)	:	Asparagine instead of aspartatic acid	Charge	Asparagine is not charged, while aspartic acid is negatively charged. This could affect interactions with other residues and protein function.