**BIOL 432 Major Project**

**Intro**

Since the initial isolation of the West Nile Virus in 1937 from a human in the West Nile District of Uganda, the West Nile Virus has spread globally and become an increasingly important pathogen for humans and animals. West Nile Virus is a mosquito-borne pathogen of the family *Flaviviridae,* genus *Flavivirus*. The natural transmission cycle of the virus primarily involves *Culex* species mosquitoes and birds, with humans being incidental hosts. While the majority of infected humans are asymptomatic, there have been significant WNV epidemics with multiple reports of encephalitis and death in history. The 1999-2000 epidemic in the USA coincided with an epizootic event in which substantial numbers of wild birds and equines died.

The WNV genome is a single-stranded positive-sense RNA encoding a polyprotein precursor of approximately 3430 amino acids. The polyprotein is cleaved into 3 structural and 7 non-structural proteins. The three structural proteins are the capsid (C) protein, precursor and membrane (prM/M) protein, and the envelope (E) protein. At present, there is no registered human vaccine or established treatment for WNV. Structural proteins represent viable targets for the development of new antiviral agents, as such proteins are critical for many stages of the virus replication cycle. Mutations in these critical cell proteins can disrupt host-virus interactions and the efficiency of viral replication; as such, cell proteins involved in viral replication can be expected to be highly conserved between divergent host species.

**\*next slide\***

Recent phylogenetic analysis on nucleic acid sequence data from a portion of the E protein gene has highlighted two distinct lineages of the WNV. Lineage 1 has a worldwide distribution and is known to exist in such geographically distinct places as Western Africa, the Middle East, Eastern Europe, the USA, and Australia. Lineage 2 exists exclusively on the African continent. Lineage 2 appears to be less virulent to humans, while lineage 1 isolates have been known to be involved in epidemics and epizootics.

**\*next slide\***

We will conduct an amino acid sequence alignment for each of the three proteins. As these proteins are integral for viral replication, we expect to see high levels of conservation across sequences. Identifying conserved sites is important in elucidating candidates for vaccine or treatment targets. We will then use sequence alignment information to construct a phylogenetic tree of each of the three proteins. In our analysis, we expect to see two lineages evident in the E protein phylogeny, with one corresponding to Lineage 1 (global distribution) and the other corresponding with Lineage 2 (African distribution only). We will investigate whether the two known lineages are parallelled in C and prM protein phylogenies.

**Methods**

In our study, we conducted a phylogenetic analysis to confirm previous phylogenetic reportings in the literature, with an additional goal of determining if specific proteins or regions of proteins may be suitable therapeutic targets for WNV infection. Using 104 WNV complete genome amino acid sequences collected from a wide variety of vectors across the globe, we were able to isolate each of the three structural proteins (E, C, pM) encoded by the WNV genome, and create a series of alignments and phylogenetic trees. This was to shed light on the similarities and differences between each of these sequences based on the protein itself, as well as the vectors and geographic regions from which the sequences were obtained.

We initially attempted to do this by using regex to find the open reading frames within a FASTA file of all the nucleotide sequences we obtained, as the structural proteins are always the first three proteins in the genomic sequence. However this strategy resulted in us isolating a single massive open reading frame, which we later figured out was because the WNV genome encodes a polyprotein that is cleaved into its constituent parts. We adjusted our strategy accordingly, and ended up going with the amino acid sequences instead for our analysis.

Protein sequence isolation was done by running a compiled FASTA file of all the 104 amino acid sequences through the through the NCBI CDART (Conserved Domain Architecture Retrieval Tool) which provided the numeric positional ranges of amino acids within the WNV polyprotein which correspond to the structural proteins. These ranges were used to isolate each protein sequence from each genome in our dataset using the strsplit function applied over a dataframe of the sequences.

**\*\*Next Slide\*\***

The quality of the 104 protein sequences was inspected and visualized using ggplot. This was important to ensure we were grabbing the right regions of every genome (i.e. that the selected regions contained the proteins of interest). The majority of the sequences had a genome length of 3433-3434 amino acids meaning we were likely to capture the sequences as intended.

(mention pie chart is usually not good for visualization but this is the exception).

The collection date was limited to the 20th century from 1900 to 2000, although the first west nile virus strain was first isolated in 1937 and the first sequence of the virus was collected in 1953. As evident from the graph, most of the sequences were collected from 2000 and few are scattered around the 1960s and 1970s, so we expect to see the evolutionary history of this virus from the sequences.

Most of the virus were isolated from mosquito and bird species along with some strains isolated from humans. This is expected since the transmission cycle of west nile virus involves the mosquito as the vector and other animals as amplification hosts, and ultimately infecting humans.

The sequences were collected from all over the world, spanning 5 continents with the majority of the sequences coming from the United State and Australia. Some sequences are collected in African countries like Congo, Madagascar, and Nigeria. Few virus strains were isolated from India and Russia.

The percentage identity for the protein is generated as a heatmap with the scale from 90% to 100% at gradient increment of 2%. Whole genome amino acid sequences were also clustered in a bash script using the CD-Hits algorithm to supplement our phylogenetic analysis, and act as a third point that may support either our phylogeny or the 2 lineage phylogeny we found during our literature review.

MUSCLE alignments were created and the Ape package was used to visualize sequence similarity based on identity and the biochemical properties of the amino acid sequences. This used the alview function.

Phylogenetic trees for each of the three structural proteins (E, C, pM) were created, with groupings corresponding to the geographical regions, in addition to a set of trees focusing on the vector species.

An analysis was performed to identify regions of interest within the proteins, and determine if vector or geographic region played a greater role in grouping the sequences.

**Main Findings**

After constructing our phylogenies,we had several initial observations. We did not notice any distinct groupings amongst vectors in our phylogeny, so those phylogenies were not included in our presentation. We did notice some groupings by country and we also identified three distinct lineages in each of the protein phylogenetic trees. In figure 7 you can see our full phylogeny for protein C as well as three distinct groupings outlined in the box. Figures 8, 9, and 10 display these zoomed in groupings for each of the three protein phylogenies (C, E, and prM). We were intrigued by these groupings because we thought that they could be similar to the phylogeny lineages from literature. Interestingly, the two lineages identified by prior literature did not match our constructed phylogenies. Based on literature, we expected to see two lineages; one with only African isolates, and one with globally distributed isolates.

**Switch slide - describe**

Our lineages show that the most ancestral strains of our sample were isolated in the USA. Isolates from Africa were distributed in the second lineage with isolates from other continents.

**Describe next 3 slides**

It is possible that these discrepancies in our findings are because we used a neighbour-joining approach in constructing our phylogenetic tree, while the previously published study used a distance-based algorithm within PAUP. Our findings might also differ because of a drastic difference in which genomes were examined in our analysis vs the literature’s analysis.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

As we mentioned we used MUSCLE to visualize the sequence similarities based on the biochemical properties. These are our 3 visuals for each protein and as it can be seen that all three proteins have high levels of conservation. At the top of each alignment the grey regions indicate sequences that are not conserved whereas the white does indicates a conserved amino acid. This high level of conservation was expected based off prior literature results and based on the fact that these three proteins are structural, and integral for viral replication.

However, there are a couple areas in each protein that are not highly conserved. In figure 11, you can see that protein C has several regions with high variation with concentrations at the end seen by the gray regions. At approximately 20, 40, 60, 100, and 110 base pairs. In figure 12, protein E has variation at ~50 and 100 base pairs, and in figure 13, protein prM has high variation at ~10 and 60 base pairs. This high amount of conservation makes sense given the importance of structural proteins to the virus.

**Conclusion**

In conclusion, based on our analysis these structural proteins are highly conserved and would be good targets for antiviral agents or for use in a vaccine. We would probably recommend the prM (premembrane) or C (capsid) protein as targets because they are likely to be localized on the exterior of the virus making them easily accessible to antibodies or antiviral agents. Of course, extensive additional research would be required to confirm this.

Overall, our analysis of the West Nile Virus protein genomes resulted in different results compared that presented in the literature. While the literature uncovered two distinct lineages in the West Nile Virus genomes, we uncovered three in each protein’s genomes. Additionally, the countries where our lineages originated differed significantly compared to the lineages found in literature. While these results could have been the result of a different genome data set or a different phylogenetic method, it could also indicate a need for further research into the topic. If we are able to target these structural proteins with the antiviral agents we could prevent future epidemics.