**BIOL432 Project Proposal**

Group 1 - Sunny Days : Amelia Walsh, Dale Moskoff, Micah Grubert Van Iderstine, Sydney Berman, Yifan Duan

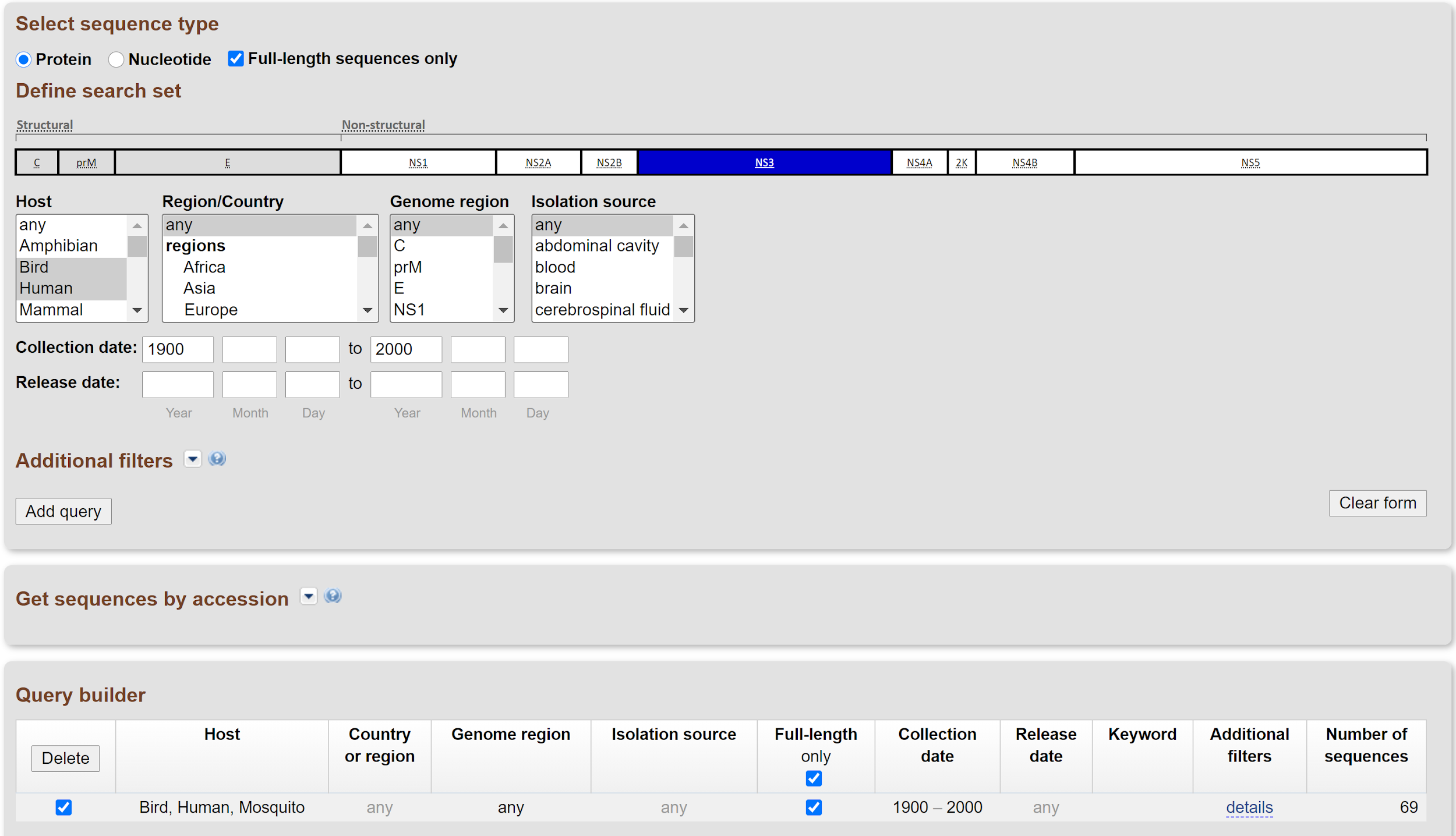
1. **Pick a single GitHub repo from your group to house your group project files**

https://github.com/15aw57/Sunny-Days

1. **Identify a dataset to analyze (e.g. DNA sequences, gene expression profiles, large ecological dataset) using tools that you’ve learned in the course. You can search ncbi.org, datadryad.org, or look for a 'data availability' or similar section in a recently published journal article. You can also ask your lab members for unpublished data to analyze, but please do not use your own data (e.g 537 thesis students).**

**West Nile Virus Database - Strains, Dates, Species, Origins**

https://www.ncbi.nlm.nih.gov/genomes/VirusVariation/Database/nph-select.cgi?taxid=11082



**Database Search Criteria:** Bird, Human, Mosquito, Full Length CDS Only, Unique Sequences Only, 1900-2000 (**69 sequences**)

To access our data, we used NCBI’s West Nile Virus database (link above) and filtered the data to include Bird, Mosquito, and Human strains of the virus. To access the genomes we will be using in our analysis, use the link above and input the search filters displayed in the screenshot.

1. **Think of 1 to 3 biological questions you can address with the data. NOTE: you may have to iterate between #1 and #2 before you settle on something you are comfortable with.**

“*What is the phylogeny of the West Nile Virus across geographic locations?*”

“*What is the phylogenetic relationship of West Nile Virus between mosquito (vector) and avian (reservoir)?*”

These biological questions are relevant because of the current negative impact of the West Nile Virus on North American bird populations. The virus has been decimating North American bird populations since its introduction to the area in 1999 (LaDeau *et al.* 2007, May *et al.* 2010), and thus represents a key conservation concern (LaDeau *et al.* 2007). In addition, avian hosts of the virus act as reservoirs, creating reserves of the virus for transmission to humans (Taieb *et al.* 2020). Identifying the strain of the disease currently impacting North American bird populations thus has implications for human health in addition to conservation efforts. Determining the phylogeny of the strain across geographical locations and animal groups will provide an understanding of the history of the disease and its rate of adaptation.

1. **Briefly (<500 words) describe your dataset and what questions you think you can address using the computational techniques you have learned in this course.**

The dataset we will be working with is from the West Nile virus database of the NCBI Virus Variation. The data contains complete viral sequences (defined as from UTR5->NS5) from Human, Mosquito, and Bird hosts from the years 1967-2000. In our initial query we were able to identify 69 unique sequences. The data will provide us with information relating to the variation in the virus’ sequence across geographic locations and across species. Our goal is to produce a phylogenetic tree, along with a map indicating the origins/locations of various strains. We chose to include older data in our analysis as this provided us with a more manageable number of sequences and also represented a more variable span of time when the virus jumped across multiple species and geographic locations. The data in this data set will allow us to analyze how the West Nile Virus phylogeny differs across species and geographic locations. We can further use this information to determine how often the West Nile Virus jumped species in its first 30 years, and how that contributed to its global geographic spread.

The types of data points in our data set include Accession Numbers, Host Species, Geographic Locations, Amino Acid Sequences, Dates of Collection, and Isolation/Tissue Sources. We will be using genbankr to pull the 69 complete sequences and mosquitos for alignment. Multiple Sequence Comparison by Log-Expectation (MUSCLE) will then be used to align all of the sequences and to identify any conserved motif across the different strains. This alignment will be visualized using ape package from R. We will also construct a distance matrix with the 69 sequences using the ggplot package and the matrix will be used to generate a phylogenetic tree using ggtree. Most of the techniques will be from week 6 and week 7 that deal with alignment and phylogeny. Some additional techniques we will include bash scripts to reorganize and relocate the fasta files and the construction of an average amino acid identity matrix (ANI matrix) using heatmap from gplots package. We will aim to utilize the ggmap package to visualize location data from the sequences.

**References**

LaDeau, S. L., Kilpatrick, A. M., and Marra, P. P. 2007. West Nile virus emergence and large-scale declines of North American bird populations. *Nature, 447*(7145) 710-713.

May, F. J., Davis, C. T., Tesh, R. B., and Barrett, A. D. T. 2011. Phylogeography of West Nile Virus: from the cradle of evolution in Africa to Eurasia, Australia, and the Americas. *Journal of Virology.*