IRD Highly Pathogenic H5N1 Clade Classification Tool

November 15, 2012

The IRD H5N1 Clade Classification Tool was designed by Dr. Catherine Macken of the Los Alamos National Laboratory, a member of the Influenza Research Database (IRD) Team. The tool assigns an H5 clade annotation to an HA/H5 sequence (the “query” sequence). The tool works for highly pathogenic H5 sequences and low-pathogenic H5 sequences. This tool does not propose new clades. Instead, it is designed to find the clade in the WHO full tree (http://www.who.int/influenza/gisrs_laboratory/201101_h5fulltree.pdf) to which the query sequence is most closely related. When the query sequence is from a virus outside of the A/goose/Guangdong/1/96 (GsGua)-lineage, the classification tool will return a result that indicates this fact.

Algorithm
We first created a classification tree designed to represent all clades in the WHO full tree. For each of the currently defined WHO clades, we attempted to include in the classification tree viruses from the earliest emergence through to the latest observations of the clade. To accomplish these ends, our classification tree was derived by reference to the WHO full tree and to our in-house phylogenetic tree of all available H5 sequences. The final classification tree includes a total of approximately 300 H5 sequences representing all of the clades of the A/goose/Guangdong/1/96 (GsGua)-lineage, together with representatives of Eurasian H5 sequences from outside the GsGua lineage, and American low-pathogenic H5 viruses. The early high-pathogenic virus A/chicken/Scotland/1959 was included as the outgroup.

The query sequence is first aligned, using ClustalW [1], against a curated alignment of the sequences comprising the classification tree. Then placer [2] is used to attach the query sequence to our classification tree. Placer uses the nucleotide alignment of the sequences in the classification tree and estimates of parameters of evolution under the GTR model of evolution to find the maximum likelihood location and branch length for attachment of the query sequence to the classification tree. Key to our procedure for clade assignment is that the classification tree does not change. Thus, the classification tree acts as a “scaffold” upon which the query sequence is hung.

When the WHO defines new clades, we will update our classification tree and recompute clade assignments for all HA/H5 records in the database. Updates are likely to occur on an annual or bi-annual basis.

Output
When a query sequence is placed unequivocally within the bounds of a single clade, the classification assigned is that of the clade. If a query sequence is attached to a branch that separates two clades, such as the branch connecting clade 2.3.4 sequences and clade 2.3.4.3 sequences, the query is assigned the clade “2.3.4-like.” The notation “like” indicates that further clade resolution is not possible. If a query sequence lies on a branch that separates two clades, such as 2.3.1 and 2.3.2, then the query is assigned the classification “2.3-like.” Again, the terminology “like” indicates uncertainty in classification to a higher level of granularity. Sequences may also be classified as “LP-Amer” (when the HA/H5 virus is more similar to the low pathogenic viruses of American origin), “EA-nonGsGD” (when the HA/H5 virus is more similar to the Eurasian viruses of non-GsGua origin), and “LP-China” (when the HA/H5 virus is more similar to the low pathogenic viruses of Chinese origin).
virus is more similar to the Eurasian origin, low-pathogenic viruses), “outgroup” (when the virus is a sibling to A/chicken/Scotland/1959), or “unknown” (when a clade cannot be definitively assigned). Note that HA/H5 sequences assigned a classification of “outlier” in the WHO large tree are often given a numerical clade by our classification tool.

The accuracy of clade assignments made by the IRD clade annotation tool has been evaluated by running it against all publicly available members of the WHO large tree. The tool has been found to be highly accurate: for a query sequence not in the classification tree, but in the WHO large tree, the IRD tool predicts the correct clade better than 99% of the time. Details of this evaluation, as well as the current make-up of the classification tree, are described in a publication currently in preparation.

**Limitations**
When applied to non-HA sequences, the IRD H5N1 Clade Classification Tool returns unpredictable and potentially erroneous results.

HA sequences of less than 300 nucleotides may also produce erroneous results.
