

Bioinformatics: Life Sciences on Your Computer

Viral Databases

This is supplemental reading for the video lecture on Viruses.

Discussion Topic: Viruses have small genomes and are easy to sequence. The focus of sequencing is on subtypes and not on new viruses. Why do you suppose that might be (there are many reasons)?

ICTV

[ICTVdb](#) is a database of viral taxonomy. Viruses are grouped into order, family, genus, species. The interface can be a bit difficult to navigate.

Discussion Topic: What are metagenomic sequencing projects?

NCBI Viral Genomes

[Entrez Viral Genomes](#) has an extensive catalog of sequences, as of March 10, 2014, there were 5,233 reference sequences for 3,843 viral genomes. There are links for specific types of virus (e.g. dsRNA) and a pull-down menu to access families of viruses. Note that on the front page, there is a prominent question shown that reads, "Did viruses invent DNA?"

VIZIER Project

VIZIER is a European Union-sponsored project that began in 2004. It's a large scale project that goes beyond basic bioinformatics and extends to drug design. The VIZIER Targets Database provides tools for target prediction. Some viral proteins are more suitable as antigens or as targets for antivirals than others, thus target prediction is an essential part of the database and project. The [Vazymolo interface](#) allows for BLAST searches and transmembrane protein prediction. A [brief tutorial](#) is now available, but it is short and mostly screenshots. The Browser tab allows for browsing viruses by taxon and by protein within each taxon, and also has text search.

Hepatitis virus

The hepatitis virus "group" really consists of five unrelated viruses. They are primarily classified due to the effect on the liver. Hepatitis B can cause long-term chronic issues, but a vaccine is available. Information is available at [HBVdb](#).

Hepatitis C is an RNA virus that evolves VERY rapidly. Genotypes can be divided into seven groups. Vaccines are difficult to develop because of the rapid evolution--there can be greater than 30% divergence in nucleotide sequence between genotypes! Los Alamos hosts the [HCV sequence database](#).

Influenza

The influenza virus belongs to the viral family, [Orthomyxoviridae](#), which included Influenzas A, B and C, Isavirus and Thogotovirus. It's a single-stranded RNA genome that is "negative sense." The Influenza A virus is the one that most concerns public health officials. The genome has eight segments, PB1, PB2, PA, HA, NP, NA, M and NS. The strains are named for the HA and NA segments. The H1N1 strain, has type 1 hemagglutinin and type 1 neuraminidase. In 1918, H1N1 wiped out a good chunk of the European population that were able to survive World War I. That subtype caused concerning 2009, but H1N1 was

not nearly as deadly this time around. H2N2 was responsible for a 1957 pandemic and H3N2 hit in 1968.

The H5N1 strain has been mostly limited to birds, but prior to the recent H1N1 scare, it was H5N1 that had the health community worried. So far, the virus has jumped to humans, but it is believed that human-to-human spread has not (yet) happened, at least not in outbreak mode. As of December 17, 2012, there have been 610 confirmed human cases of H5N1, killing 360, mostly in Asia. It is likely that the number of actual cases is higher, because genotyping is rarely done in the absence of very severe symptoms.

NCBI maintains the [Influenza Virus Resource](#), which includes sequences from very recent outbreaks, as well as other tools such as BLAST for genotyping and clustering for phylogenetic tree building.

Discussion Topic: Please bring your perspective, from your research or from what you know, to influenza and potential dangers of emerging strains & subtypes.

Herpesvirus

These are dsDNA viruses, often with very large genomes for viruses. The study of herpesviruses can be useful because a good chunk of their genomes code for orthologs of proteins from the host organism that the virus infects. These viruses proved useful in confirming phylogenetics studies in mammals, as the branching of the viruses strongly correlates with the branching of species. For more details, check out a review by McGeogh et al. (PMID: [16490275](#)).

Because many herpesviruses code for 80+ proteins, they can be useful in the study of viral protein clusters.

Human Immunodeficiency Virus

The virus we now know as HIV-1 was once called HTLV-3 when discovered in the early 1980s and renamed in 1986. As you well know, this virus leads to devastating symptoms because infection compromises the immune system. It is a retrovirus, different from RNA viruses like influenza in that the life cycle includes a DNA intermediate.

The [NCBI Retrovirus Genomes](#) page includes tools for genotyping, BLAST and browsing in addition to a multiple sequence alignment tool specialized for retroviruses. Like with influenza, there are links to recently added sequences and publications. There is a specialized database for the study of [HIV-1, human protein interactions](#).

The Los Alamos National Laboratory also hosts databases for viral genomes, including the [HIV Sequence Database](#). Many of the same tools available at NCBI are here as well, but LANL also has a geography based tool and a program called SNAP that looks at substitution rates in coding and noncoding context.

Discussion Topic: Herpesviruses and HIV differ greatly in how long ago they first appeared. Discuss some history of these and other viruses. How does the phylogeny differ?

Measles Virus

Long after vaccines have been available, measles continues to be a deadly disease. Measles is part of the Paramyxoviridae family and the reference genome can be found at NCBI with accession NC_001498. To access the genome resources, enter that RefSeq number in the [Genomes](#) database (not the

nucleotide database). The virus codes for seven proteins with six genes: N (nucleocapsid), P (phosphoprotein), M (matrix), F (fusion), H (hemagglutinin), L (large polymerase). P also codes for a nonstructural protein.

Surprisingly few measles-specific bioinformatics resources are widely available. That may be due to a limited number of genotypes.

Discussion Topic: What was the first viral vaccine? How successful was that vaccine and how far have we come since then?