Phylogenetics for Predicting Virus Evolution

Paul Wiesemeyer

October 10, 2020

1 Introduction

1.1 Problem

Viruses and the diseases they provoke put a large burden on human life. These comparably simple structures are capable of modifying and possibly destroying vital functions in the human body, by infiltrating their genetic material into the reproductory apparatus of cells.

All viruses that persist in the human population over longer times share some common features:

- They need some form of protection from their surroundings, a hull
- They need to find a way to get into human cells
- They aim at reproducing quickly
- If there is no constant infection source, they need to find a path to get from one individual to another
- They will have to deal with human immune system response
- They will underlie some kind of evolutionary pressure

(This rough characterisation is of course incomplete, and there may also be exceptions.)

To give another rough image of what makes up a virus, here are the essential constituents:

- A piece of genetic code (RNA or DNA, positive or negative sense, usually some few kilobase(pairs) long)
- A hull: either only a capsid (of proteines) or an additional envelope (bilipid layer) with surface proteines

For enveloped viruses such as in the *Orthomyxoviridae* family (including influenza) or *Coronaviridae* (e.g. SARS-CoV-2), there is usually a variety of different surface proteins that populate the bilipid layer.

These surface proteins take up functions such as binding to a cell to infiltrate the viruses genetic sequence into it, or releasing a freshly assembled virus from the host cell's surface into the surrounding body fluids. This

is usually done via interactions with specific host cell receptors, therefore defining a *receptor binding domain (RBD)* as the part of the surface protein that fits onto the receptor as a key fits into a lock.

At the same time, the human immune system will also get in contact with mostly surface proteins. The immune system will eventually develop antibodies that are targeted to bind to a specific region of virus surface proteins, thereby rendering them innocuous. The surface protein region tageted by the immune system is called *epitope site* and may have a large overlap with the RBD.

These two mechanisms put an evolutionary pressure on the virus, especially on the surface proteins and its epitope sites. While having to maintain essentail funcionings as binding to receptors, the virus will draw large advatage from an epitope site modified to such an extent that the antibodies cannot bind to it any longer. This way of disguising itself—by animo acid mutations in crucial places—will allow the virus to reinfect previously immune individuals.

An accelerated evolution can be very beneficial to the virus, *but*—since the capability to sequence and read out the viral genome—also provides a large record to trace a virus' history.

In this write-up, we will look at how the sequencing of the viral genome, along with phylogenetic tree inference, can provide useful insights on (i) the route that the virus takes to spread in the human population, (ii) the existence and prevalence of strains witin a virus (sub)type, and (iii) the projected evasion of the virus from immune system response and vaccines.

2 Influenza

2.1 Basics

History diagram from Alberts, 2015
- HA: epl
- NA: expl
- Phylo tree of A/H3N2 HA Volz+2013

2.2 Hemagglutination Inhibition Assay

Short description
- Hirst 1943
- mention cartography from Smith+2004 (?)