Network-based models for identifying reassortant influenza viruses

Eric J. Ma[1], Nichola Hill[2], Jonathan Runstadler[1,2]

[1]Department of Biological Engineering and [2]Division of Comparitive Medicine, MIT

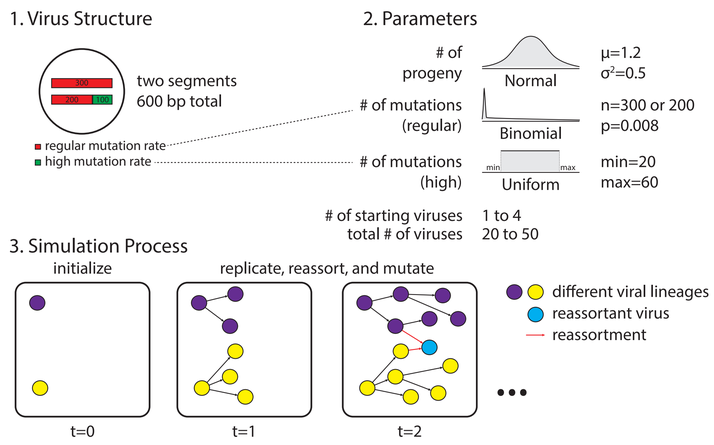
Introduction

- The influenza virus is a segmented RNA virus capable of undergoing reassortment, bringing antigenic shifts that enable host immunity evasion.
- Surveillance efforts have yielded densely sampled sequence datasets.
- Phylogenetic methods are useful for inferring most recent common ancestors and their time of appearance. However, they may not result in accurate genealogies where parent and child viruses are both present in the dataset.
- Previously published methods for reconstructing virus transmissions use densely-sampled sequences with sampling dates attached.
- **We extend this by explictly considering:**
  - **multiple virus seeding events**
  - **genome segmentation with independent substitution rates**
  - **different segment evolutionary rates**
- We show that this method can accurately reconstruct simulated virus transmissions and identify reassortant viruses.
- We apply this method to influenza viruses isolated from wild ducks in Minto Flats, AK, and observe unexpected trends in viral gene flow.

Model Construction

A. Viral Spread Simulations

We have developed a simulator that simulates the spread and reassortment of a segmented virus, such as the influenza virus.



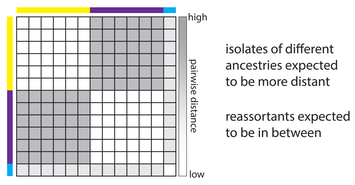
B. Network Reconstruction

We use sequence and date from each segment of the virus to reconstruct the most probable ancestor of each virus.

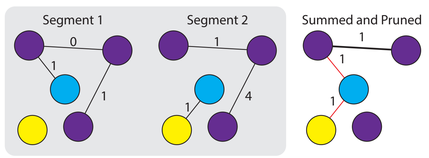
1. Per-segment sequence alignment

```
1  ...ACGGCATTACGAGCGATC...
2  ...ACGGCATAACGAGCGATC...
3  ...ACGGCATCACGAACGATC...
4  ...ACGGCATCACGAGCGACC...
5  ...ACGGCATCACGAACGAAC...
6  ...ACGGCATCACGAACGGTC...
7  ...ACGGCATCACTAACGGTC...
8  ...ACGGCATCACGAACGATC...
⋮              ⋮
```

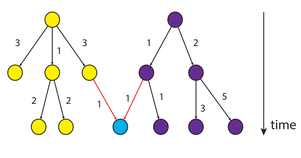
2. Pairwise distance calculation



3. Segment-by-segment reconstruction



4. Reconstructed genealogy

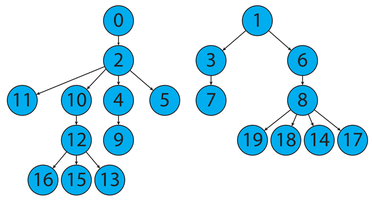


Model Validation

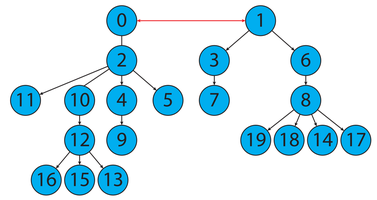
A. We can accurately reconstruct simulated viral transmissions.

- Simulations indicate that we can accurately identify all transmissions, but reconstructions overestimate other transmissions.
- This may be due to seed isolates being connected together.
- We will define a statistically-sound threshold to eliminate such scenarios.

1. Example Ground Truth



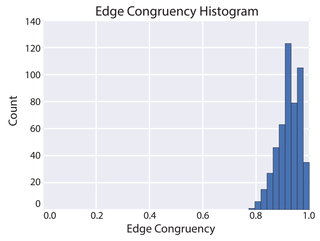
2. Example Reconstruction



1 edge estimated more than the original transmission.

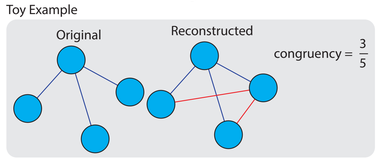
3. Simulation Results

We ran 500 simulated influenza outbreaks of different sizes and starting viruses, and were able to identify all transmissions present in the simulated outbreak, while overestimating other transmissions.



$$\text{Edge congruency} = \frac{\text{\# of true edges found}}{\text{\# of edges found in reconstruction}}$$

Toy Example

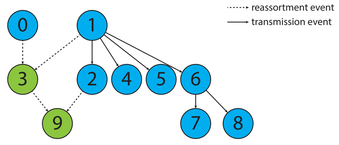


Edge congruency is a measure of overestimation of transmissions due to highly similar sequences. Smaller values indicate higher over estimation

B. We can accurately identify the simulated reassortant viruses present.

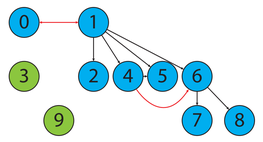
- With reassortants added in, identification of reassortants results in many more runs where we overestimate the transmissions, but average accuracy is still high.
- We are able to identify reassortants at a rate better than pure random guessing.

1. Example Ground Truth



Isolate 3 and 9 are simulated reassortants

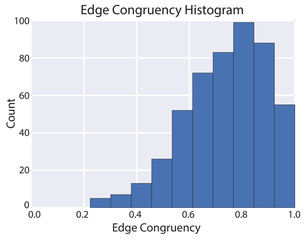
2. Example Reconstruction

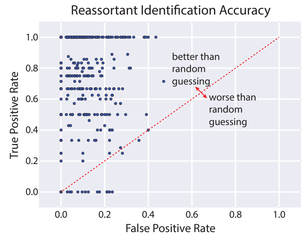


Reassortants are identified by not having any full transmissions into them.

3. Simulation Results

We ran 500 simulated influenza outbreaks of different sizes and starting viruses, including reassortant viruses, and were able to accurately identify reassortants while maintaining relatively good transmission accuracy.





Real-World Application

A. We tested our network reconstruction algorithm flu isolates from Egypt.

- With our reconstruction of flu outbreaks in Egypt, we identified chickens as the source of viruses that jumpt to humans, as previously hypothesized.
- Humans do not transmit the H5N1 virus from one to another, confirming what we know.

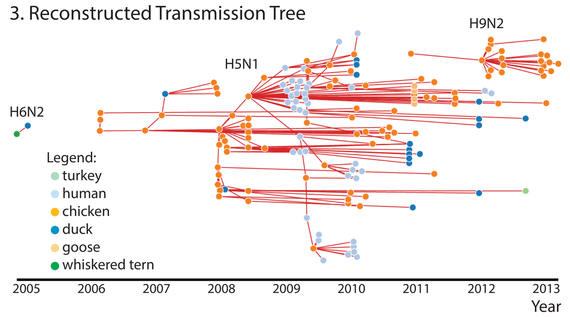
1. Basic Statistics

# of isolates: 157
segments: HA and NA
subtypes: H6N2, H5N1, H9N2
time span: 9 years

2. Results

172 estimated transmissions
3 estimated reassortants

3. Reconstructed Transmission Tree



B. We identify reassortments happening in wild duck reservoirs in Alaska

- Unexpectedly, there is little gene flow from one year to the next. Seasonality is a larger barrier to gene flow than host species.
- Our measured substitution rate of 10^{-3} nt.site^{-1}.yr^{-1} from the dataset implies negligible number of mutations occurring. This parameter can be used in our simulations

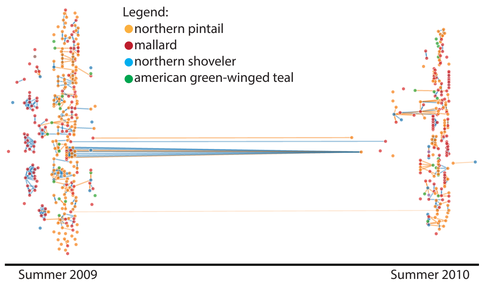
1. Basic Statistics

# of isolates: 516
segments: all 8
subtypes: diverse
time span: 2 summers

2. Results

972 estimated transmissions
56 estimated reassortants

3. Reconstructed Transmission Tree



Conclusions

- We can use isolation dates and genomic sequences to reconstruct viral outbreaks.
- We can accurately identify all reassortant viruses a majority of the time.
- Our reconstructions match known biology (Egypt data), and provide unexpected insight into viral gene flow (Alaska data).

Further Research

- Refine reconstruction algorithm with thresholding for better reconstruction accuracy and identification of reassortants
- Extension to global influenza dataset to measure influenza's global reassortment rate.
- Metadata overlay: what factors are most correlated with influenza reassortment?

Acknowledgments:

MIT BE

CEIRS
NIAID Centers of Excellence for Influenza Research and Surveillance

Massachusetts Institute of Technology

The Division of Comparative Medicine
MIT - DCM Building 16, Room 825
77 Massachusetts Avenue
Cambridge, MA 02139

MIT CEHS