Spatiotemporal and Phylogenetic Analysis of Influenza A Virus

Fangye Tang
B00612172

Introduction

Influenza A viruses are very common respiratory pathogens and infect many animal species. There are several times of flu outbreaking in history. The recent one is H3N2 flu outbreaking in 2017; it caused more than 100 people dying a week in America (Mortimer, 2018). Since there are lots of subtype of influenza A virus and each subtype has mutated into a variety of strains, the gene sequences analysis becomes important. There are lots of literature about this area. Kim phylogenetic analyze H3N2 virus in Korea in 2012. He found some segments are similar with same segment but in different region, for example he pointed out the segment PB2 in Ontario is similar (97.5%) with that in Minnesota and different from that in Korea (Kim, 2014). Another report is about phylogenetic analysis and pathogenicity of H3 subtype in China by Cui. He found the revolution route within H3 subtype (Cui, 2016). All these researches can help scientist upgrade new vaccines for an unknown virus based on previous similar virus or its ancestor virus.

From these literatures, the first one proof region does affect the revolution of H3N2; and second one points out homology of same subtype. However, there are lots of types of influenza A virus and they only analysis H3 subtype. I'm curious about is there any spatiotemporal effect on other types. In this report, I will analysis the spatiotemporal effect on three different types of influenza: H1N1, H3N2, and H5N1 and the homology of them (these three subtypes are typical). To do that, I use RNA sequences from NCBI Influenza Virus Resource, the data collects around world from 1970s to 2017. I first cluster analysis sequences, that will help me to find the spatiotemporal effect. And then running multiple sequences alignment to feed into phylogenetic tree builder. By comparing these phylogenetic tress, I can figure out the homology of these influenzas.

Methods

The datasets are from NCBI Influenza Virus Resource (https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database). This is a very strong power searching engine for influenza virus. It provides the full sequences for any segment and any subtypes. I can select sequence type, like protein or nucleotide. I can choose the type of influenza, for my project I will choose influenza A and for human. And I can select country or region that I need to analyze, also the year. I can also choose which segment and subtype. Since my project can be divided into two parts, I will talk about them separately.

1. The spatiotemporal effect on three different types virus.
   Each influenza A virus has 8 independent segments, and each of them works separately. But since I wish to analyze human influenza virus, and only segment PB2 and HA takes genetic factors. I will just use these two segments to analyze. I first need to download these segments separately. I will have 6 sets of sequences.

For each set, I will run greedy algorithm cluster analysis provided by USEARCH (http://www.drive5.com/usearch/). It will help me to check whether spatiotemporal effect on this type or not. I will set the minimum sequence identity to 99.9%, that means only the identity between two sequences is 100% will be considered as a cluster. After that, I can find the effect by checking centroid node and its group sequences. And repeat previous step for each set. Because I need to analyze both time effect and regional effect, I need to do them separately. By running regular expression, I can find all sequences in a specific year, this will help me to analyze the regional effect. With same method, I can find all sequences in a specific region, this is useful for checking time effect.

2. The phylogenetic analysis across these types virus.
   This part is a little bit hard and cost time. I first need to build datasets manually. For each segment, I need to copy the corresponding segment in H1N1, H3N2, and H5N1, and paste it into a new file. I also need to label it before pasting. I will reuse the previous datasets. I will have 2 bigger sets of sequences.

   The key idea is to analysis phylogenetic tree. To build the tree, it needs three steps:
   1. Cluster:
      Because these datasets are very large, I cannot feed it to multiple sequences alignment. I still need to cluster them first by using exactly same method as before. After that, the number of sequences reduce from 6000+ to 1600.
   2. Multiple Sequences Alignment:
      I use MUSCLE here (http://www.drive5.com/muscle/), set everything default. MUSCLE is a very good tool for multiple sequences alignment. It includes fast distance estimation called Kmer, progressive alignment using log-expectation score, and refinement using tree dependent restricted partitioning. The speed and accuracy of MUSCLE are higher than T-Coffee, MAFFT and CLUSTALW by running four test sets. Even without refinement, MUSCLE gets average accuracy from T-Coffee and MAFFT (Edgar, 2004).
   3. Build Phylogenetic tree:
      I use IQTREE here (http://www.iqtree.org/#download), just set everting default and let IQTREE test substitution model. The basic model of IQTREE is constructing parsimony trees. IQTREE use maximum likelihood to compute distance based on best-fit substitution model. The program also tests the reliability of tree by using bootstrapping procedure, so the final phylogenetic tree should be good. This step will take very long time.

   Repeat above three steps for each dataset. By comparing and combining two result trees, I can do phylogenetic analysis for these three subtypes.

Results

1. The spatiotemporal effect on three different types virus.
   As mentioned in Methods, I will analyze time and regional effect separately because one factor may affect another. For better visualization, I choose two periods: 2009-2010, 2016- 2017 and compare sequences inside clusters and their centroid nodes. After a quick view of cluster sets, I find most clusters are within same continent and same period, like that:

```
S   69  759 *   .   *   *   *   ADF27388 A/H1N1/Texas/46193632/2009 2009/10/19 PB2  *
H   69  759 100.0   .   0   759 =   ADJ80783 A/H1N1/California/VRDL115/2009 2009/12/04 PB2  ADF27388 A/H1N1/Texas/46193632/2009 2009/10/19 PB2
H   69  759 100.0   .   0   759 =   ADJ80823 A/H1N1/California/VRDL119/2009 2009/12/15 PB2  ADF27388 A/H1N1/Texas/46193632/2009 2009/10/19 PB2
H   69  759 100.0   .   0   759 =   AGI53536 A/H1N1/Kansas/21/2009 2009/09/16 PB2   ADF27388 A/H1N1/Texas/46193632/2009 2009/10/19 PB2
```

   We can see the centroid is in Texas, and its group members are all in USA. And we can also see the collect data are very close.

```
S   124 759 *   .   *   *   *   AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2 *
H   124 759 100.0   .   0   759 =   AEJ82821 A/H1N1/Taiwan/552/2011 2011/01/18 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEL89690 A/H1N1/Bangkok/INS491/2010 2010/08/20 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEL89710 A/H1N1/Bangkok/INS493/2010 2010/08/23 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEL89720 A/H1N1/Bangkok/INS494/2010 2010/08/23 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEL89740 A/H1N1/Bangkok/INS497/2010 2010/08/24 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEL89790 A/H1N1/Bangkok/INS503/2010 2010/08/25 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEL89821 A/H1N1/Bangkok/INS506/2010 2010/08/27 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEN55502 A/H1N1/Bangkok/INS490/2010 2010/08/20 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AEO37549 A/H1N1/Bangkok/INS516/2010 2010/09/02 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AFP35804 A/H1N1/Singapore/GP4183/2010 2010/09/28 PB2    AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AFP35814 A/H1N1/Singapore/GP3956/2010 2010/09/02 PB2    AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AFP35908 A/H1N1/Singapore/GP4266/2010 2010/10/09 PB2    AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AGB13409 A/H1N1/Bangkok/INS581/2010 2010/09/08 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AGB13439 A/H1N1/Bangkok/INS584/2010 2010/09/30 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AGB13449 A/H1N1/Khon Kaen/INS585/2010 2010/09/01 PB2    AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AHM98448 A/H1N1/Khon Kaen/INS3_647/2010 2010/09/23 PB2  AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AIE52342 A/H1N1/Bangkok/SIMI506/2010 2010/09/28 PB2 AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AIE52347 A/H1N1/Bangkok/SIMI511/2010 2010/09/28 PB2 AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AJJ99737 A/H1N1/Thailand/SN10445/2010 2010/09/28 PB2    AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AJK00052 A/H1N1/Thailand/KS08356/2010 2010/07/06 PB2    AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
H   124 759 100.0   .   0   759 =   AKQ11543 A/H1N1/Singapore/DMS136/2010 2010/08/12 PB2    AEJ82817 A/H1N1/Taiwan/1018/2011 2011/01/08 PB2
```

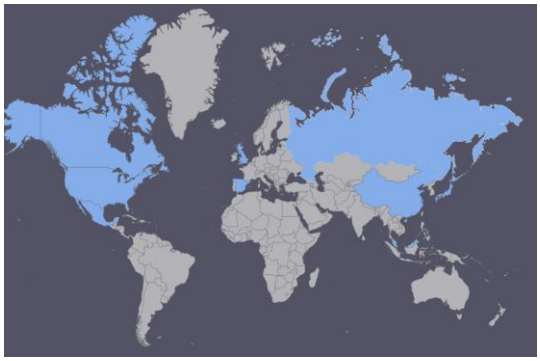
   Another example shows the centroid is in Taiwan, and its group members are all around Southeast Asia. The collect data are all around the end of 2010.

   However, there are some clusters are very strange:

```
S   364 759 *   .   *   *   *   ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2 *
H   364 759 100.0   .   0   759 =   ACS92578 A/H1N1/Fukuoka-C/2/2009 2009/06/07 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ACS92588 A/H1N1/Fukuoka-C/3/2009 2009/06/07 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ACT21978 A/H1N1/Kagoshima/1/2009 2009/06/13 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ACZ16059 A/H1N1/Mexico City/009/2009 2009/05/12 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ADF28094 A/H1N1/Brussels/INS106/2009 2009/10/29 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ADK21932 A/H1N1/Hvidovre/INS290/2009 2009/11/27 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ADK32399 A/H1N1/Brussels/INS206/2009 2009/10/30 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ADK32409 A/H1N1/Brussels/INS208/2009 2009/11/03 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ADK32419 A/H1N1/Brussels/INS209/2009 2009/11/04 PB2 ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ADK32429 A/H1N1/Pensacola/INS210/2009 2009/11/10 PB2    ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   ADK33890 A/H1N1/Athens/INS263/2009 2009/12/13 PB2   ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   BAM34356 A/H1N1/Yamagata/473/2009 2009/11/09 PB2    ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
H   364 759 100.0   .   0   759 =   CRI06090 A/H1N1/England/01220736/2010 2010/02/04 PB2    ACS69020 A/H1N1/Fukuoka-C/1/2009 2009/06/07 PB2
```

   The centroid is in Fukuoka in Japan. But we can find the group members are scattered around the world. This makes sense, the collecting data is in the middle of 2009 flu pandemic. Mexico, Japan, Belgium, Denmark, and England are the confirmed community outbreaks (Wikipedia, 2018). The flu may travel from one country to another.

   From above section, it seems time and region can affect virus. The centroid and its group member are close to each other. But how about the centroid with centroid? I use a world map tool to label each centroid country, it shows like that:




The left one represents the centroid country in period 2009-2010 for H1N1; and right one is for period 2015-2016. It is clearly to see each centroid are regional distribution, not randomly scattered around the world.

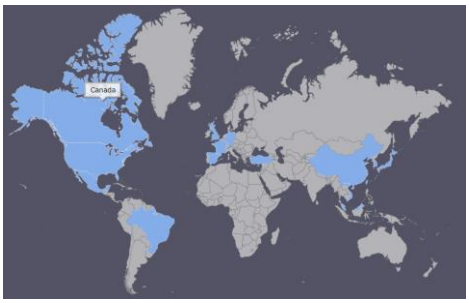
For now, we can say there is a spatiotemporal effect on H1N1. But how about other subtypes? Since other researchers has proved effect on H3N2, I just analyze H5N1 here. Because there are very few data for H5N1, this is quite easy. All centroid and its corresponding group member are in exactly same country and period, and centroids are only in China and Indonesia. However, this result is just for segment PB2, but how about segment HA?

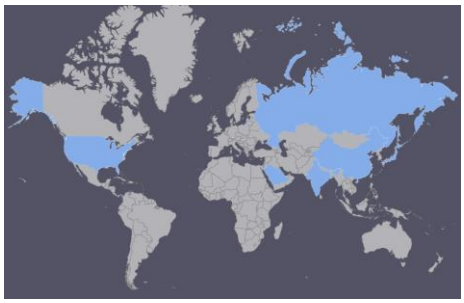
I use exactly same process as we did for segment PB2. First, we still focus on centroids and their group members.

```
S   132 566 *   .   *   *   *   BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA   *
H   132 566 100.0   .   0   566 =   BAI94581 A/H1N1/Nagasaki/HA-10-11/2010 2010/01/12 HA    BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAJ06677 A/H1N1/Nagasaki/HA-10-14/2010 2010/02/01 HA    BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAJ39986 A/H1N1/Sendai/TU433/2009 2009/10/05 HA BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAJ40006 A/H1N1/Sendai/TU460/2009 2009/10/20 HA BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAJ40019 A/H1N1/Sendai/TU489/2009 2009/10/22 HA BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAJ40020 A/H1N1/Sendai/TU490/2009 2009/10/23 HA BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAJ40045 A/H1N1/Sendai/TU588/2009 2009/11/17 HA BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAM34269 A/H1N1/Gunma/262/2009 2009/11/19 HA    BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAM34349 A/H1N1/Tochigi/10/2010 2010/01/12 HA   BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
H   132 566 100.0   .   0   566 =   BAM34369 A/H1N1/Yamagata/674/2009 2009/11/27 HA BAI59782 A/H1N1/Nagasaki/HA-56/2009 2009/10/22 HA
```

We can see the basic idea is same as segment PB2. The centroid and its group members are all located in Japan. And it also has some special case as before, the flu may travel around the world because of outbreaking.

For analyzing centroids and centroids, I also label to world map:

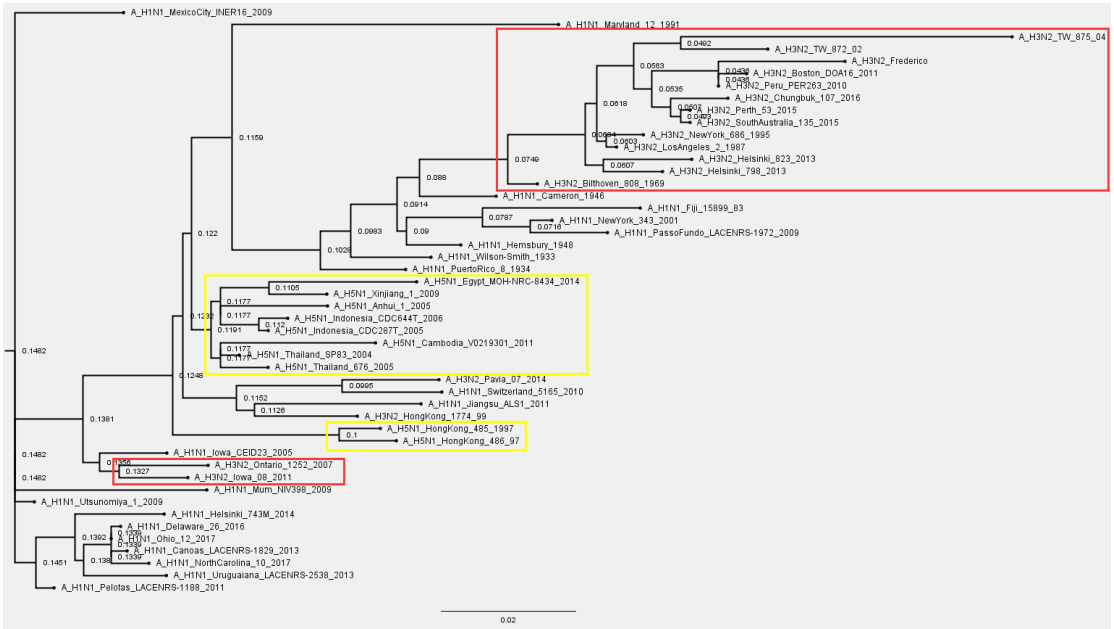




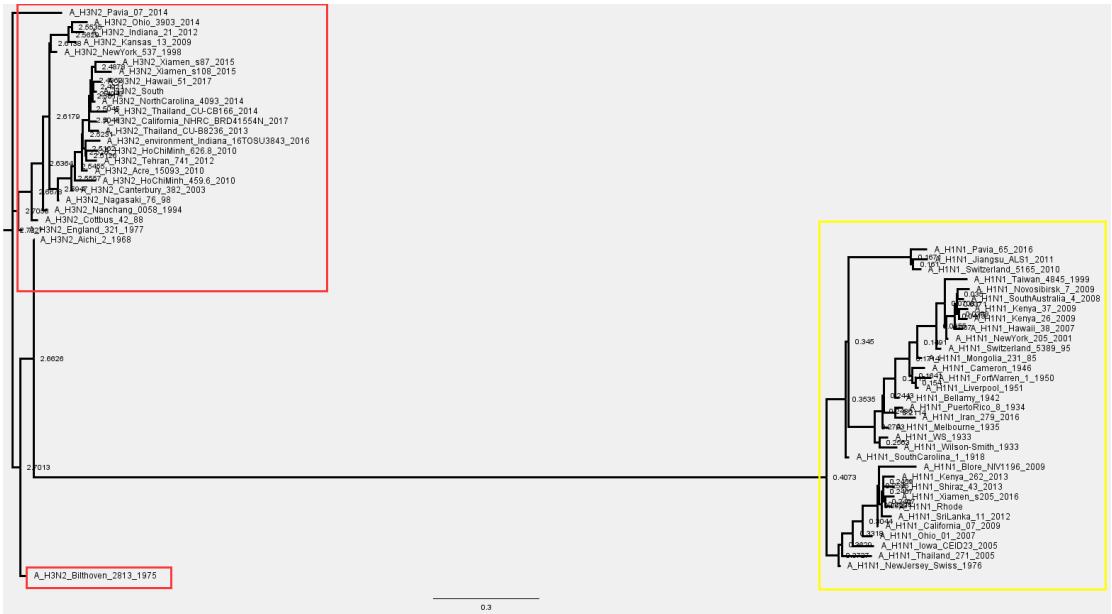
The left one represents the centroid country in period 2009-2010 for H1N1; and right one is for period 2015-2016. We can see the centroids are similar with segment PB2. The centroids are located regionally not randomly scattered around the world. The fact are also same as H3N2 and H5N1.

Now, We can conclude there is a spatiotemporal effect on Influenza A virus, it doesn't matter witch subtype of virus and doesn't matter witch segment we used.

2. The phylogenetic analysis across these types virus.
   As mentioned before, I need to cluster datasets before alignment. Because I copy all three subtypes sequences in one file, I cannot still set minimum sequence identity to 99.9%, this will give me more than 6000 centroid nodes, and each node will become a leaf of phylogenetic tree. I set the identity to 99%, which is still a good tag. After clustering, I get 66 groups and normally distributed each subtype. Here is the phylogenetic tree looks like:



It is easy to see that the red zone is for H3N2 and the yellow zone is for H5N1. Almost all subtypes are in one branch, which is perfect. These two subtypes are not randomly distributed in the tree. They basically lie in same branch. That means the subtypes of influenza A virus are evolved from H1N1, which the first found in 1918 when Spanish Flu Pandemic. This also match the result from Morens' research. He pointed out all influenza A virus nowadays are evolved from 1918 H1N1 (Morens, 2009). More specifically, we can see the virus evolution also follow the time effect, for instance: The second yellow zone, which contains two H5N1_HongKong, are collected in 1997. On the other hand, the time collected are after 2004 in the first yellow zone. It seems time also affect the evolution of influenza A virus. As result, all subtypes can be thought as the descendants of H1N1, and H1N1 we found nowadays is the descendants of H1N1 found in 1918 Spanish Flu Pandemic. The tree's log-likelihood is -5657, which is good, the tree is reliable. Next, let's look at segment HA.



This tree is very interesting, all H3N2 are located together (red zone), and all H1N1 are located together (yellow zone). Because the data for H5N1 with segment HA is lack, it be clustered in some clusters, not represent as centroids. However, this tree seems showing H1N1 is evolved from H3N2. There is a very long branch between H1N1 subtypes and root. This makes sense, why? We can see the neighbor of this branch is “A_H3N2_..._1975” and “A_H3N2_..._1988”. The collect year of these two nodes are long time ago and closer to 1918. It seems showing the H1N1 nowadays are evolved from some different older subtypes. The tree’s log-likelihood is -10999, which is very good. So this tree is more reliable than previous one.

Now, we can conclude H1N1 in 1918 should be the ancestors of all influenza A viruses we found after. But the evolutionary route is tricky. H1N1 we found nowadays might not be evolved from H1N1 directly found older, but from some different subtypes much older.

Conclusion

In conclusion, the result of first part doesn’t surprise me. It basically supports my main hypothesis: the time and region do effect on the virus sequences. But the phylogenetic analysis is very interesting: H3N2 and H5N1 are the descendants of H1N1. By further researching, it shows all influenza A virus subtypes are the descendants of H1N1 in 1918 Spain Flu. The evolutionary route is not a straight line. The ancestor of one subtype might not be same subtype found before, but may be another subtype found much older. The weakness of my report is not related to statistical methods. The result might not be statistical significant; the IQTREE doesn’t provide the confidence level of phylogenetic tree. It just provides the log-likelihood. It is not enough to tell the accuracy of phylogenetic tree.

The study of phylogenetic for influenza A virus is very useful. If there is a new subtypes outbreaks, the scientists can upgrade new vaccines based on the unknown subtypes’ ancestor. For further study, I wish to extend my project by adding more subtypes and more segments. I wish the model can be used for every subtypes of influenza A, not

just for H1N1, H3N2, and H5N1. In addition, I wish to use another method to build phylogenetic tree, since IQTREE just provide maximum likelihood to computer distance. There are more methods can be used. By comparing these trees to find a best model. The result should be more significant.

References

Cui, H. et al. (2016, 2 2). *Phylogenetic analysis and pathogenicity of H3 subtype avian influenza viruses isolated from live poultry markets in China*. Retrieved from https://www.nature.com/articles/srep27360

Edgar, R. C. (2004, 3 19). *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC390337/

Kim, J. et al. (2014, 2 11). *Phylogenetic Analysis of a Swine Influenza A(H3N2) Virus Isolated in Korea in 2012*. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0088782

Morens, D. M. et al. (2009, 7 16). *The Persistent Legacy of the 1918 Influenza Virus*. Retrieved from http://www.nejm.org/doi/full/10.1056/NEJMp0904819

Mortimer, C. (2018, 1 20). *Flu outbreak: 100 people a week dying in US as virus continues to spread*. Retrieved from https://www.independent.co.uk/news/world/americas/flu-outbreak-aussie-flu-japanese-flu-america-cdc-100-dying-a-week-a8169896.html

Wikipedia. (2018). *2009 flu pandemic by country*. Retrieved from https://en.wikipedia.org/wiki/2009_flu_pandemic_by_country

.