8-18-2022 1:00 PM

Outbreak Detection From Virus Genetic Sequence Variation By Community Detection

Mo Liu, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Pathology and Laboratory Medicine

Abstract

Detecting risk groups in transmission networks can be difficult due to a virus' high transmission rate. We hypothesize that this problem can be resolved by community detection methods. Community detection is a clustering method based on edge density, which can break a connected component into multiple smaller clusters. My project develops a framework to find more informative clusters of virus sequences by applying community detection methods to transmission networks of HIV-1 sequences from Beijing and Tennessee, and a global dataset of SARS-CoV-2 sequences. We set the sequences with the most recent sample collection date as “new cases” and the remaining as “known cases”. Then, the difference of Akaike information criterion (AIC) between two Poisson regression models is measured. By using this framework, we determine that the HIV-1 database from Beijing favors a higher distance threshold than Tennessee, and in the SARS-CoV-2 transmission network, some pairs of countries (i.e., England and Portugal) are more significantly associated than by chance.

Keywords: Bioinformatics, HIV, SARS-CoV-2, community detection

Lay Summary

Identifying risk groups among infections can be difficult in the study of virus epidemiology. A transmission network is a graph-based method to describe the relations among infections by considering pairs of sequences to be connected if their difference (e.g., genetic pairwise difference) falls below a given threshold. A transmission network can be partitioned into several connected components or clusters. A connected component in a network is a subgraph in which node representing infections are connected to each other. Previous research in transmission networks has focused on HIV-1 due to its rapid evolution. This method can also be applied to other viruses, such as SARS-CoV-2. However, due to the rapid transmission rate of SARS-CoV-2, component based clustering is not able to detect informative clusters from a large number of infections with a small number of mutations. We hypothesize that this problem can be resolved by community detection methods. Community detection is another clustering method based on edge density, such that infections within a community would have more edges and fewer edges between communities. My project develops a framework to find more informative clusters of virus sequences by applying community detection methods to the network given by pairwise distances from three different datasets: Beijing and Tennessee HIV-1 sequence data and global SARS-CoV-2 sequence data. We observe a higher optimal threshold in community detection methods, so that we are able to include more cases in the model than connected component-based clustering methods. By using this framework, we determine that the HIV database from Beijing favors a higher distance threshold than Tennessee. In the SARS-CoV-2 transmission network, some pairs of countries (i.e., England and Portugal) are more significantly associated than by chance.

Acknowledgments

I would like to express my deepest thanks to my supervisor Professor Art Poon, for his guidance along these two years' study. Art is really professional, I am always impressed by his broad knowledge and he makes me enjoy studying bioinformatics so much. Art is always patient and supportive, this journey would have been difficult if I have not meet him. I would also like to thank the other members in the Poonlab, all of them are so sweet and incredible, it is pity that all of our meetings are online.

My committee member, Professor Lindi Wahl and Professor Parisa Shooshtari, thank you for providing vital feedback on my research which have kept my work on track. I would also like to extend my thanks to Professor Grace Yi and Professor Zia Ali Khan for providing fantastic graduate courses.

Loves to my dearest parents and lovely friends (especially Qinbei Li, Haohan Liang and Yufei Xia) for being the love and joy of my life.

Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Background

1.1 HIV

Human Immunodeficiency Virus (HIV) is a type of lentivirus that targets the immune system and can lead to Acquired Immunodeficiency Syndrome (AIDS) if not treated (Centers for Disease Control and Prevention, 2021). Lentiviruses are species of retroviruses that copy on RNA genome that is converted to DNA within the host cell. (Boskey,2022). Retroviruses encode their own reverse transcriptase protein to transform their single-stranded RNA into double-stranded DNA which can become integrated into the host genome (Boskey,2022). This conversion from RNA to DNA manipulates the infected cells into replicating the genes of the virus (Boskey,2022).

There are many lentivirus species that infect other primates and mammals, e.g., cats and rabbits. There have been two species of lentivirus that have been discovered in humans so far, Human Immunodeficiency Virus Type 1 (HIV-1) and Human Immunodeficiency Virus Type 2 (HIV-2), respectively. These are further divided into 4 classes for HIV1 which are M, N, O and P and another 9 subgroups A-I for HIV2 (Robertson et al., 2000). HIV-1 viruses in subtype M are the main group that dominates the HIV pandemic and can be further classified into subtype A to H, J and the newly defined K. The most prevalent subtypes for HIV-2 viruses are A and D. HIV-1 accounts for over 90% of infections worldwide, whereas

HIV-2 is less common and has few infections (Robertson et.al., 2000). AIDS was first discovered by an unusual increase in death rates among young homosexual men in 1981, which was later determined to be caused by HIV-1. Another reporting of a virus similar to HIV-1 was found to cause AIDS in Western Africa, despite having little relationship with HIV-1. This virus was closely related to a simian virus that caused immunodeficiency in macaques. This virus belonged to a single evolutionary lineage of primate lentiviruses and appeared to cause no harm in bodies for both humans and non-human hosts (Sharp & Hahn, 2011).

Over the past 40 years, from the first reported case of HIV-1 infection in the 1980s, to 3.7 million new cases and drug treatments in 1997, the AIDS epidemic has expanded significantly with increased transmissions. In the 2000s, approximately 9.7 million people in low and middle-income countries received antiretroviral drug treatment (Lee 2010). There were still approximately 37.7 million people living with HIV and 1.5 million people acquiring HIV by the end of 2020. Of those, 95.5% of the population were adults and 1.3 million were children aging between 0 – 14. Since the start of the HIV epidemic, a total of 79.3 million people has been infected by the human immunodeficiency virus with a death toll of 36.3 million (World Health Organization, 2021). HIV prevalence rates vary significantly between countries with Africa being the most affected continent on earth. Out of the 37.7 million people living with HIV globally, 69% of them live in sub-Saharan Africa. Furthermore, all the top 5 countries with the highest HIV rates are located in Africa which are Eswatini - 26.8%, Lesotho – 21.1%, Botswana - 19.9%, South Africa-19.1% and Zimbabwe – 11.9%. The most common reasons for cases are poverty and lack of knowledge about HIV. Around 390 million sub-Saharan Africans are living in extreme poverty. These people have a lack of access to basic health care service and medical devices like condoms. Poverty is also related to low

education of means preventing HIV infections (World Health Organization, 2022). In 2019, a total of 2122 HIV diagnoses were reported in Canada, with the highest rate of new HIV diagnoses being 5.6 per 100000 population. Saskatchewan reported the highest provincial diagnosis rate at 16.9 per 100,000 population. The 30-to-39-year age group had the highest HIV diagnosis rate at 12.7 per 100000 population. The number of reported HIV cases are also dramatically expanding among Chinese youth. According to data collected by the China Center for Disease Control and Prevention, the annual number of new HIV diagnoses grew from 2705 cases in 2005 to 42406 cases in 2019 (Xu et al., 2021). There were around 1.045 million Chinese residents living with HIV by October 2020 with an incidence rate of 0.075%. The HIV transmissions were majorly dependent on needle sharing and blood contact back in the 20th century. However, over 50% of new HIV infections were caused by sexual transmission by 2006, with heterosexual sex becoming the main cause step by step. A large number of new cases among the gay community also increased briskly thereafter, representing 34% of all new infections in 2016, up from only 2.5% in 2006 (Xu et al., 2021). In the meantime, some major people groups experience a greater risk of HIV infections compared with the rest of the population. Bisexual men are considered to be the most vulnerable people to the HIV infections. In 2019, men who have sex with men (MSM) took responsibility for 69% of new HIV cases of which Black Americans accounted for around 36% and white MSM accounted for more than 30% in the United States. Heterosexual Americans were 23% infected in 2019, the transgender people made up around 2%, and injection drug users accounted for 7% (U.S. Statistics, 2022).

Although the Single-Genome-Amplification (SGA) test reduced/eliminated certain errors, it is important to note that it was only conducted on the Env major gene. An HIV genome contains nine genes which encode 15 viral proteins in addition to the three major

genes: gag, pol and env. (Li et al., 2015). Subtypes are defined by nucleotide/amino acids divergence. The envelope glycoprotein (env) consists of a complex of gp 41 (transmembrane protein) and gp 120 (surface protein). The Gag reading frames contain p17, p24, p7, p1 and p6 proteins. The Pol gene proteins encoding follows the gag reading frames for the late-phase protease, reverse transcriptase (RT) and integrase (Int) and complex with RNase (German Advisory Committee Blood (Arbeitskreis Blut), 2016). In addition to the 3 major genes, the HIV genome also codes 6 regulatory proteins which are transactivator protein (Tat), RNA splicing-regulator (Rev), negative regulating factor (Nef), viral infectivity factor (Vif), virus protein r (Vpr) and virus protein unique (Vpu) and have the essential impact on viral replication and budding. The genome of HIV-2 codes virus protein x (Vpx) rather than Vpu, which conducts of reducing pathogenicity (German Advisory Committee Blood (Arbeitskreis Blut), 2016). The HIV genome comprises two single-stranded RNA molecules that are inside the core of the virus particles. The RT in the Pol gene proteins is able to transcribe the RNA genome into DNA, degrade the RNA and combine the double-stranded DNA to generate the HIV proviral DNA. The HIV-1 genome comprises of 9700 nucleotides and HIV-2 contains around 9800 nucleotides. (German Advisory Committee Blood (Arbeitskreis Blut), 2016). A study by Shaw and his colleagues (2012), the molecular and biological features of the HIV virus were determined using SGA of endpoint-diluted plasma vRNA / cDNA approach. This approach offered improvement in the analysis, such as eliminating Taq polymerase errors, template switching and template resampling from viral and single genomes respectively. This method also reduced errors related to the misidentification of target frequencies caused by unequal cloning. The method was used to test the composition of HIV-1 subtypes A, B, C, D, CRF01_AE and others with env major gene with full-length sequence of gp160 genes. It was found that all Envs were biologically

functional and dependent on CD4 type cells. Of the 55 Envs used in the test, only one was found to be CCR5/CXCR4 dual tropic. All other Envs tested were of the CCR-5 tropic type (Shaw & Hunter, 2012).

The transmission of HIV requires intimate contact, such as the exchanging of body fluid. The transmission of HIV can differ greatly between acute transmission, chronic transmission and AIDS. Acute transmission belongs to the early-stage infection and the symptoms will develop between 2 to 4 weeks. During the early stage, HIV viruses usually replicate and spread throughout the body to launch attacks on the CD4 T Lymphocyte (CD4 cells). Chronic HIV transmission (asymptomatic HIV infection) is the second stage of infection. During this time, HIV viruses will continue to replicate but relatively slower and patients usually will not experience any HIV related symptoms. The final stage of infection is AIDS, and the viruses will cause severe damage to the immune system. During the final stage, the viral load reaches the peak and the CD4 counts drop to the minimum. The immune system within a patient's body is too weak to fight off opportunistic infections and typically they won't be able to survive about three years without any treatment (NIH, 2021). Acute transmission rates can be much higher than transmission from chronic hosts/infections in both animals and humans due to the high viral load. From the Indian Rhesus Macaque model of SIV transmission, it was found that the acute stage of infection had a specific transmissivity which was approximately 750 times greater than a chronic stage (NIH, 2021). Factors such as other sexually transmitted diseases (STDs) and pregnancy can increase infection susceptibility by approximately 2 to 11 times more (NIH, 2021). The risk of transmission of HIV-1 increases exponentially going from the eclipse phase to when it is detectable in blood plasma. The HIV eclipse phase is an interval following HIV acquisition in which HIV cannot be tested. From a laboratory staging experiment by Shaw 2012, it was

found that the plasma virus RNA copies increased exponentially from an order of a 10^4 to 10^6 after the initial eclipse phase, which ranges from 7 to 21 days post infections (Shaw & Hunter, 2012).

1.2 SARS-CoV-2

2019 December 31st, the first reports for a novel coronavirus, SARS-CoV-2 were reported by the Wuhan Municipal Health Commission of China. Until July 26th, the disease has caused around 567million cases confirmed cases and 6.3million deaths (World Health Organization, 2022). The first genome sequence was named WH Human 1 coronavirus (WHCV), also known as '2019-nCoV'. The whole genome sequence (29903 nt) has been assigned GenBank accession number MN908947. The viral gene organization of WHCV is determined by a human-associated coronavirus and a bat-related coronavirus (bat SL-CoVZC45, GenBank access No. MG772933) (Wu et al., 2020). There are many guesses about the origin of SARS-CoV-2, Andersen et al. observed all notable SARS-CoV-2 features and they don't believe that there is any type of reasonable laboratory-based scenarios. Andersen et al. suggest a further observation of animals will be the most definitive way to find the origin of SARS-CoV-2(Andersen et al., 2020). However, even though there is evidence suggesting SARS-CoV-2 is not a purposefully manipulated virus, it is still impossible to support this hypothesis over other theories to date.

Infectivity and transmissibility of SARS-CoV-2 in humans can be detected from the first group of genome sequences. The high similarity of genomes implies fast human-to-human transmission. On the other hand, the rate of evolution is slower than the rate of transmission, while the mutation rate remains similar, many genomes are identical to each

other. The rate of evolution of the SARS-CoV-2 genome is 7.3 × 10−4 (5.95 × 10−4–8.68 × 10−4) nucleotide substitutions per site per year (Bukin et al., 2021). The transmission rate of SARS-CoV-2 is between 0.19-0.29/day (Romero-Severson et al., 2020). Due to the high transmission rate, enormous amounts of SARS-CoV-2 sequence data are being collected in a relatively short period of time. As of 27th July 2022, there were over 12.1 million SARS-CoV-2 genomes shared on the Global Initiative on Sharing All Influenza Data (GISAID) database. Massive and identical SARS-CoV-2 sequences make the SARS-CoV-2 transmission network harder to cluster than the HIV transmission network.

An accurate understanding of the global spread of emerging viruses is critical for public health responses and for predicting and preventing future outbreaks. There are some studies that analyzed the early spread of the SARS-CoV-2 epidemic. The subsequent spread of SARS-CoV-2 around the world was reconstructed from genome sequences. Worobey et al. found that the SARS-CoV-2 virus arrived in Europe and North America in late January or early February. The first virus genome detected was similar to the mutation found in the Chinese sample, it spread rapidly and caused a widely undetected community transmission. More precisely, the viruses first infect Italy around the end of January, then reach Washington state around the beginning of February, and get to New York city later that month (Worobey et al. , 2020). Nadeau et al. found that SARS-CoV-2 was widely spread out in France, Germany, Italy, and other European countries from China approximately two to four times each before 8 March 2020 (Nadeau et al., 2021). Genome sequencing of SARS-CoV-2 is used to reconstruct the spread of SARS-CoV-2, and this process can highly depend on sampling. Bedford et al. build a Maximum-likelihood phylogeny from SARS-CoV-2 viruses. Clusters of closely related viruses suggest an independent introduction event followed by

local transmission. A high-density comb-like structure cluster indicates rapid exponential growth. (Bedford et al., 2020)

A high transmission rate and evolution rate result in an unexpected massive number of SARS-CoV-2 sequences. To overcome this problem, Rambaut et al. present a virus nomenclature. They build a maximum likelihood tree and then find the most contributed lineages. Besides the phylogenetic framework, another important part of virus nomenclature is the naming system. The naming system of this study involving

a dynamic nomenclature system proposal and Lineage naming rules. The valid standard for terminology in the naming system needs to capture coherent global patterns of viral genetic diversity in time, be flexible enough to adapt to the new viral diversity, and be dynamic (i.e. contain births and deaths).

While the virus spreads, it constantly replicates itself and makes numerous copies of it. During the replicating process, there might be slight differences between copies. In other words, virus sequences can differ slightly over time. A mutation is defined as the changes in sequences during this process, and variants are defined as virus sequences with mutations. Note that variants can differ by more than one mutation. Among all variants, a variant can be called a Variant of Concern (VOC) when is significant enough to affect one or more of the following: Transmissibility (spread), Virulence (severity of disease), Vaccine Effectiveness, and Diagnostic tests. A lineage means the closely related virus variants come from a common ancestor. The table below lists all the lineages for SARS-CoV-2(Public Health Ontario, 2022).

World Health Organization label	PANGO lineage
Alpha	B.1.1.7
Beta	B.1.351
Gamma	P.1
Delta	B.1617.2
Omicron	B.1.1.529

Table 1.1 table listed variant of concerns (VOCs) identified globally and in Ontariolineages forSARS-CoV-2(Public Health Ontario, 2022).

The high mutation and replication rates of RNA viruses have been proven for more than half a century. These fast mutation frequencies compared with the host allow them to change in genomic evolutionary space, speeding their variability process, and in some cases may allow them to acquire suitable phenotypes to survive in the stressful environment, for example in antiviral therapy, the lineages of viruses can accumulate changes in certain ways. The mechanism to accumulate the favorable genomic change and clear the bad mutations is exchanging function for mutually exclusive types of a gene. The process includes at first re-classification for viruses with segmented viral genomes. then followed by recombination that happens for both segmented or non-segmented virus; currently, instead of the fast mutation speed of other RNA viruses, the genetic diversity of SARS-CoV-2 has mutated quite slowly: in public databases, there are tons of genomes worldwide, but only 7- 8 major circulating clades were found. Due to the relatively stable genomic evolvement form, the development of effective vaccines was fast and supports the interpretation of SARS-CoV-2 pathology(Kozlakidis, 2022).

1.3 Clustering

1.3.1 Network and Clustering

the mathematical theory of networks can be traced back to the Euler's solution of Konigsberg's bridges puzzle which asked to devise a walk through seven bridges once and once only that span a river flowing past the city (Euler, 1736). The connections between groups of individuals with an infectious disease can defined as a network. An edge is a connection that extends from one vertex to another, and vertex represents infections which are connected by an edge if they are closely related. Clusters in the network can represent the transmission risk structure of a population if the rate of evolution is sufficient high. In general, understanding the transmission network's structure allows us to improve predictions of the likely distribution of infection and the early growth of infection. A transmission network is a graphic-based method to describe the relations among infections by considering pairs of sequences to be connected if their distance (e.g., pairwise genetic difference) falls below a given threshold. A transmission network is often partitioned into several connected components or clusters.

Any method to identify similar data groups in a collection of data points can be called a clustering method. Community detection is one way of clustering Like other branches of network science, clustering in networks has made great progress and is widely used in various areas in the past years, such as social network analysis and medical imaging. Biological, mechanical, and social networks can be represented as graphs, and cluster analysis has become pivotal to comprehend the elements of these frameworks. Image segmentation in machine intelligence studies is a method to break down a digital image into different subgroups. This method can be treated as a graph partitioning problem in the

image data (Shi & Malik, 2000); in urban construction or the performance of water distribution systems. Which can be visualized by dividing the system into clusters and demonstrates their connections according to the flow directions (Perelman and Ostfeld, 2011).

Clustering and the relation between the number of intra-cluster and inter-cluster edges is important and meaningful for a network. More precisely, identifying clusters and their borders gives the classification of vertices. Vertices that have numerous edges to other vertices, or in another word, vertices with a high degree size may have a significant role of importance and control. For instance, a recent study estimated which individuals may have responsible for a disproportionate number of infections by reconstructing a graph in which hosts were represented by vertices and find the ones with a high degree size (Liu et al, 2020).

Clustering algorithms differ in what criteria establishes a cluster, these algorithms can be categorized into two groups, Connectivity-based clustering and Centroid-based clustering. Connectivity-based clustering, also known as hierarchical clustering, is a type of clustering method that groups the closer or more similar vertices by distance measurement into clusters such as Euclidean distance. The formular to calculate the distance between p and q is:

$$d(p,q) = \sqrt{(q1-p1)^2 + (q2-p2)^2 + \cdots + (qn-pn)^2}$$

Parameters (q1, q2,… ,qn) are the coordinates of point q.

Parameters (p1, p2,….,pn) are the coordinates of point p.

A dendrogram is a tree diagram which is often being used to illustrate the output and of a hierarchy cluster. However, Connectivity-based clustering is not proficient in dealing with outliers because it would either create additional clusters or cause unwanted merging of other clusters. In Centroid-based clustering, clusters are represented by a central vertex, and vertices are assigned to the closest clusters which have the minimized squared distance. One limitation is that the initial setting of the medoids would affect both the shape and effectiveness of the clustering result (Lloyd, 1982). The k-means clustering is the most commonly used algorithm in this clustering type (Lloyd, 1982). The goal is to sort unlabeled data into groups with the nearest mean and the number of groups are represented by the variable K. Distribution-based clustering and Density-based clustering are the most popular among all clustering methods. Distribution-based clustering consider graph as a composition of distributions where the type of distribution of data is known, such as Gaussian distribution. Previous studies have used this method to merge sequences from the same organism (Preheim, 2013) and to detecting earthquakes (Xu 1998). Apart from its high scalability, this type of clustering method is computationally expensive, and overfitting requires a large volume of data (Xu, 2015). Overfitting usually happens when a statistical model performs worse impacts on the test data in contrast of good performance on training datas when feature increase. Density-based clustering refers to a method to distinguish the data clusters based on its concentration and density by contagious region. This is suitable for data with arbitrary shapes and outliers, but the clustering results can be highly biased and affected by parameters (Xu, 2015).

1.3.2 Random walk and Clustering

In 1828, Brown described an irregular motion of pollen particles under the telescope, now known as Brownian motion (Brown, 1828). Around one hundred years later, Einstein further introduced this idea into one of the three fundamental advances of physics (Einstein, 1905). Random walks are simplified models of Brownian motion. Nowadays, random walk theory describes an unbiased stochastic process consisting of a sequence of steps where a walker is able to move along every possible path with some non-zero probability. This can be used to represented erratic changes, like a random path formed by a person walking after drinking.

Random walks can be helpful for finding clusters. A random walk on a graph would spend a longer time within clusters due to the larger number of intra-cluster edges. Zhou (2003a) defined the distance between vertices by the average number of edges traversed in a random walk from one vertex to another, such that vertices having smaller distance are more likely to belong to the same cluster. Latapy and Pons (2005) introduced a different distance measure also based on random walks as a graph. The distance is calculated by the probabilities that one vertex can connect to another vertex in a certain random walking step. Finally, Weinan et al.Finally (2008) used random walk by applying the Markov chain on the metagraph to get the best k-clusters result.

The Markov Cluster Algorithm (MCL) is a robust clustering method based on random walks. Generally speaking, it simulates the general flow of diffusion in a graph (Dongen, 2000a). MCL calculates the probabilities of random walks through the graph to detect clustered structures by a mathematical bootstrapping procedure. Bootstrap, also known as bootstrap method, is a resampling technique in statistical learning, used to estimate

standard errors, confidence intervals and deviations. Bootstrapping statistics use random sampling with replacement to estimate the sampling distributions based on a given sample. At present, the MCL is one of the most popular clustering algorithms in large-scale biological cluster detection. For instance, Enright's study has successfully applied MCL to detect and categorize protein families within the draft human genome in 2002. Unlike many other algorithms, MCL doesn't require the use to specify the expected number of clusters manually.

1.3.3 Modularity and Clustering

Community detection is one way of clustering, and its community structure is correlated to density. For instance, a cluster in connected component method can be further break down to multiple communities (Figure 1.1). Modularity is one way of measuring community density. Modularity was first introduced by Girvan and Newman's algorithm as a stopping criterion to determine network division (Newman, 2004). It is a numerical method to determine if a vertex should be decoupled into other clusters. This is done by calculating the density of edges inside the current cluster relative to edges outside the cluster. Modularity can be either positive, negative or zero. Positive modularity stands for a powerful community structure. Zero modularity is less powerful and has the same performance as random grouping. Negative modularity has worse performance than random grouping. High values of modularity indicate dense bonding between the vertices within clusters. Danon suggested to normalize the modularity changes in order to lead better modularity optima (Danon et al., 2006).

`

The Louvain algorithm is another well-known method based on modularity. The method starts by treating each vertex as its own cluster. The algorithm first moves a vertex from one cluster to another to find a partition when a local maximum of modularity is obtained, then creates an aggregate network based on the results. These two steps are repeated iteratively until the cluster quality cannot be increased further. For example, Sanchez used the Louvain algorithm to detect users with similar political preferences and to track their activity on social networks on Twitter (Sánchez, 2016).

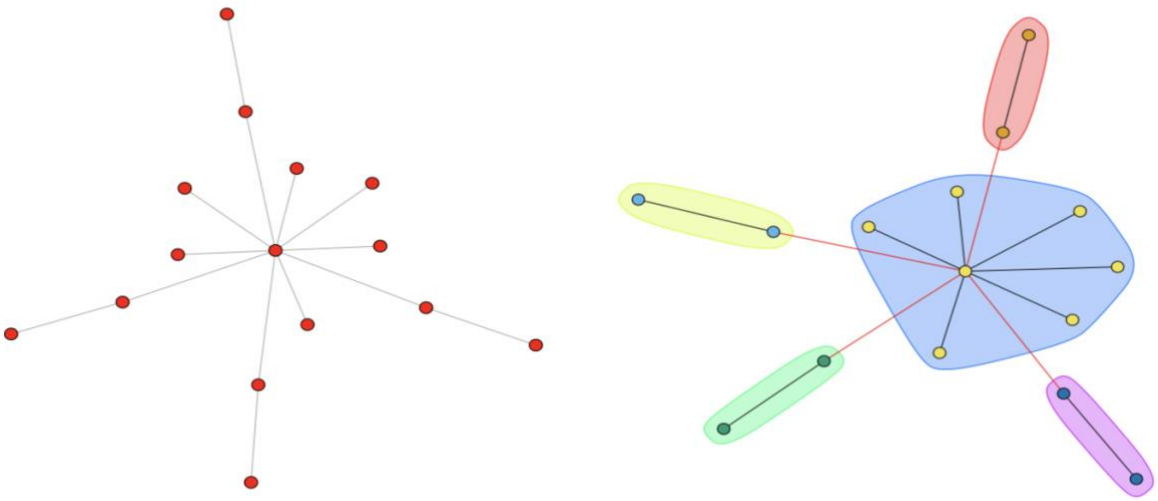


Figure 1.1: A connected component cluster(**left**) can be partitioned into five communities(**right**) by using modularity-based community detection clustering method. Communities are shown in different colors.

Modularity can also be applied to weighted graphs (Newman, 2004) and directed graphs (Arenas, 2007; Leicht, 2008). A directed graph is a kind of graph in which a set of vertices are connected by directed edges. The direction edge meaning the edge with a direction. A weighted graph has edges labeled by numbers. One crucial concern is that

detecting large maximum modularity on a graph does not necessarily mean that it contains a solid or meaningful community structure (Santa, 2010). A community structure exists in a network if the nodes can be easily divided into each group and have internal connection. For instance, cluster structure should not be observed in random graphs, and yet a previous study showed that significant large modularity values can be obtained in random graph partitions (Guimer`a et al., 2004; Reichardt and Bornholdt, 2006a). Also, modularity may not be able to detect clusters with a smaller size relative to the whole structure, even if they have a distinct cluster structure like cliques (Fortunato and Barth´elemy, 2007).

1.3.4 Genetic Clustering and Tamura-Nei (1993) Model

Genetic clustering adopts clustering method to genetic sequence, which is high-dimensional structured data due to to contains thousands of discrete immutable variables. extensive computational technique that is being used to divide a large population of sequences into smaller groups. Typically, two closely related sequences tend to form a group instead of joining other sequences with larger genetic distance.

Genetic clustering can revel patterns in the network transitivity. It has been widely used in characterizing virus diversity, since it could determine if infections are related by a common source/site (Fisher et.2020). There is growing interest in public health with the applications of genetic clusters, where we would predict the disease outbreaks on the basis of genetic variation and potentially inform the pandemic prevention if genetic clusters define meaningful groups with higher rates of transmission quickly.

To date, there are mainly two ways to construct clusters from genetic sequence data, distance-based and sub-tree-based methods. A genetic distance is a non-negative number calculated from the number of differences between the sequences, i.e., a genetic distance of zero would mean that the sequences are identical. Pairwise genetic distance comparisons have played an importance role on virus classification (Bao 2008, Van Regenmortel 2007) and molecular evolution (Real LA, 2005). Clusters are generated by specifying a threshold for distance from a phylogenetic tree or pairwise distance matrix, where individuals below that distance are assigned to the same cluster (Poon et al. 2015, Aldous et al. 2012). Sub-tree-based methods produce clusters from evolutionary distances which is the sum of the branches' length and sequence relationships Clusters also can be characterized by a number of distinct subjects, sub-tree reliability or geographical constraints, i.e., sequences share same age categories, country, or collection date (Prosperi et al. 2011, Billock 2020).

The Tamura-Nei (1993) model is used to compute pairwise distances between aligned nucleotide sequences and is the most general nucleotide substitution model (Tamura & Nei, 1993). It attempts to account for the difference between transversion mutations and transition mutations of two different transition categories (purine, pyrimidine, pyrimidine & purine). The TN93 model also has four sequence parameters A, C, G and T (Salamat et al, 2021).

The TN93 model of nucleotide evolution can be used to estimate the pairwise evolutionary distances and sequence relatedness for cluster analysis. Evolutionary distance under the TN93 model can be estimated directly from Hamming distance of a single pair of sequences. The TN93 distance corrects for unequal base composition, and it allows rapid

comparisons of 10^4 to 10^5 aligned sequences (Aldous et al., 2012). The graph of pairwise TN93 distances is formed with the computation of all individual pairs. Nodes of individuals and pairs of individuals are connected by the edges. Visually, the connected components will show as transmission clusters (Salamat et al., 2021).

The general time reversible model (GTR) is derived from a reversible nucleotide rate matrix Q. It is more efficient to reduce the number of free parameters, especially for unknown parameters. Substitutions are named transversions (Tv), where a purine is exchanged for a pyrimidine and the rest of the substitutions are transitions (Ts). Furthermore, purine transitions (A to G) T_{SR} and pyrimidine transitions T_{SY} are used to distinguish the substitutions between purine and pyrimidines (Strimmer & Haeseler, 2003). The equations to define TN93 model can be expressed as:

$R_{ij}^{TN} = k(\frac{2y}{y+1})$ (Equation for T_{SY})

$R_{ij}^{TN} = k(\frac{2}{y+1})$ (Equation for T_{SR})

$R_{ij}^{TN} = 1$ (Equation for Tv)

Parameter k is the ratio of Ts and Tv.

Parameter y is the ratio of the two different classes of transition rates.

1.3.5 Clustering and Outbreaks

One he purpose of clustering is about to detect the occurrence of the outbreak. Cluster detection is able to help identify environmental factors and spread patterns related to the disease and find the cause of the disease. It allows the public health organization to focus on preventing these groups and maximize their efforts (National Collaborating, 2012).

Outbreak investigation usually begins with identification. When several illness cases in a cluster shown by investigation have high similarities with clear associations and result to common exposures is an outbreak. Outbreak identification requires the ability to detect the illness rate when a higher-than-expected number of new cases are reported in a particular location(Wertheim et al., 2018). Therefore, it is a priority to define the expected prevalence in a certain region over a certain amount of time. According to a study of HIV from the Centers for Disease Control and Prevention, HIV transmission is around 10 to 11 times higher in a rapidly growing cluster than in the general population (National Collaborating, 2012).

HIV cluster detection and response (CDR) helps public health organizations identify the need for HIV prevention, medical treatments and HIV testing in order to against HIV transmission. Some communities have greatly succeeded in reducing HIV transmission and improving HIV care (Centers for Disease Control and Prevention, 2022). The presence of HIV clusters indicates that this community is experiencing HIV transmission, and a gap exists in HIV prevention. If the community is experiencing a rapid increase in HIV diagnoses among a specific type of group, it means that the HIV cluster is formed. Molecular data analysis is also able to quickly identify the HIV cluster by generating genetic sequences from the virus. This allows the health department to analyze the sequences to match the corresponding

clusters more comprehensively due to the high mutation rate of HIV (Centers for Disease Control and Prevention, 2022).

The characteristics or medical conditions may increase the risk of people having severe illnesses than other is named risk factors. Knowing risk factors helps people take precautions in daily living to reduce the risk of getting infected by diseases (Porta et al., 2008). Quarantine strategies are always associated with the transmission dynamics of contagious diseases like Covid -19. Clustering coronavirus disease also effectively detect unknown characteristics of clusters that appear with rapid transmission (Hong et al., 2021). After analyzing 539 clusters with a mean size of 19.21and a mean duration of 9.24 days, Korean researchers realized that the clusters with high transmission rates were in companies, factories, healthcare facilities and nursing homes. Furthermore, clusters related to markets, business and religious facilities such as churches also showed rapid growth (Hong et al., 2021). Therefore, a more efficient quarantine policy should be applied by studying these high-risk clusters. It is more reasonable and logical for the government to target the health screening test with the regional approach instead of focusing only on individual risk factors (NC Department of Health and Human Services, 2022).

1.4 Model Selection

The model selection process is vital for both academic and industry-based fields. Model selection, generally speaking, is estimating the performance of different models in order to choose the best one. Model selection strategies usually mean finding the model selection *optimize (minimize or maximize) some predetermined criterion*, often based on an estimator of generalization performance, such as k-fold cross-validation. The validation error for k fold can be divided into bias and variance components. Bias is often defined as the difference between the expected or averaged prediction of the model and the true value which we are trying to predict. The variance describes how much the predictions for a given point varies differently with each iteration of the model (Cawley, 2010). With the increasing of the model complexity, the number of model parameters increases which tends to the overfit of the model. This results in the increasing of the variance and the decreasing of the bias, and this trade-off is described as U-shaped error curve. We want to find the estimator leading to a minimum value of the test error curve.

There are many model selection criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC) and Colin Lingwood Mallows (Mallows' Cp). Among them, the AIC and BIC are most commonly used in many statistical fields (Bozdogan, 1987). The Akaike information criterion was formulated by the statistician Hirotugu Akaike. The formula can be expressed as:

$$AIC = 2k - 2\,ln(\hat{L})$$

$whose\ k$ stands for the number of estimated parameters in the model, and

$\hat{L}$ is the maximum value of the likelihood function for the model.

The basic idea about the AIC is that we can treat model performance as the sum of two parts, the first part being the goodness of fit to of the training data and the other part being the complexity of the model. The model scoring is represented by the negative of the log maximum likelihood estimate. Model complexity can be quantified by the number of parameters in the model (Brownlee, 2019). The AIC basically illustrates the trade-off between the bias and the variance. We can use it to find a "sweet spot" where we can expect optimal model performance while avoiding overly complex models. We should select the model with the minimum AIC value given a set of models. The Bayesian information criterion was formulated by statistician Gideon E. Schwarz. It is derived from a Bayesian perspective. The BIC is quite similar to the AIC but adds a stricter penalty for the number of parameters. The formula for BIC can be expressed as:

$$BIC = kln(n) - 2\ln\left(\hat{L}\right)$$

$where\ n$ stands for the number of observations or sample size.

By comparing the two formulas we can see that the penalty for AIC is 2k, whereas the penalty for BIC is $kln(n)$. This means BIC penalizes the model more for its for larger sample size, so more complex models will get a worse score and will, in turn, be less likely to be selected (Brownlee, 2019). AIC and BIC are two similar methods in model selection, however, both classes of criteria perform asymptotically well in different situations. BIC is consistent in selection when the true model is parametric; AIC performs well in an asymptotic efficiency when the true model is nonparametric scenario (Liu, 2011). If the true model is finite dimensional (parametric scenario), BIC (as a representative) performs well in selection. If the true model is high dimensional (nonparametric scenario), AIC performs well in an asymptotic efficiency. Delta-AIC or delta-BIC, i.e., the difference between the two

model's AIC or BIC values, is one of the most commonly used measurements on model selection. Previous study has shown that if the value of delta-AIC or deta-BIC is larger than 2 then we should consider one model is significantly better than the model it is being compared to. In contact, if the value of delta-AIC or deta-BIC is around 0 then there is no enough evidence to choose one model than other(Burnham & Anderson 2004).

Chapter 2

Method

2.1 Study Population and Data Processing

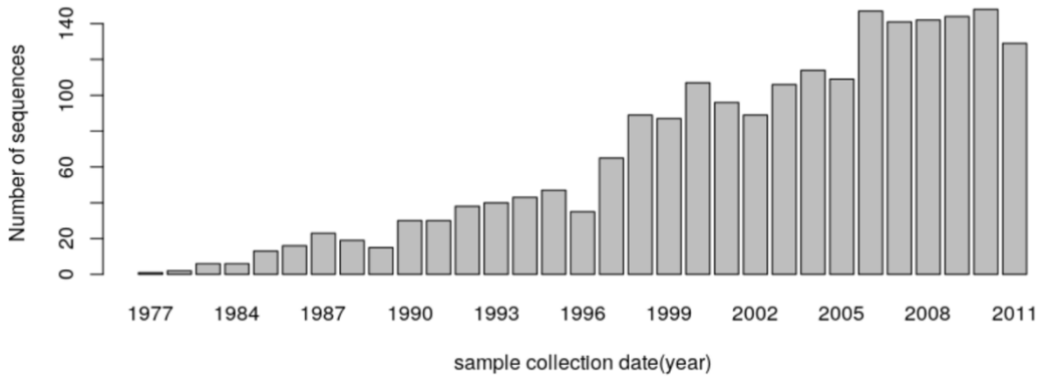
We applied our framework to three virus sequence data sets, including two HIV data sets and a SARS-CoV-2 data set. For each dataset, we collected each sequence's accession number and collection date. Additionally, we collected sequence's collection location(country) for SARS-CoV-2 dataset.

2.1.1 Study Population and Data Processing of HIV datasets

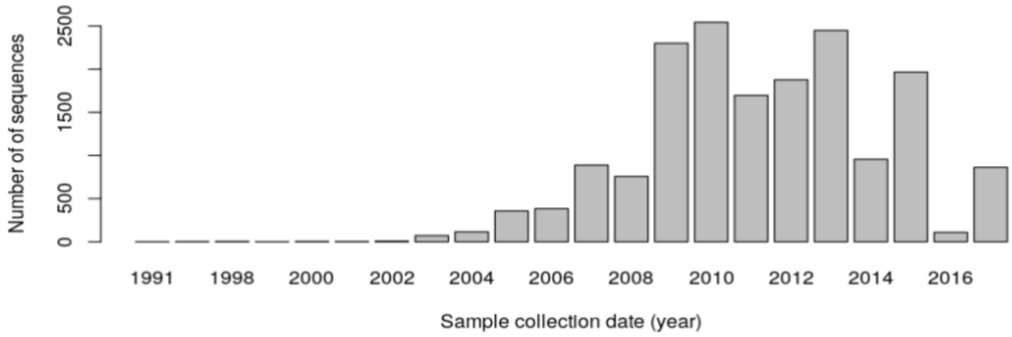
For the Tennessee HIV dataset, we obtained n= 2915 HIV-1 *pol* sequences that were sampled in middle Tennessee (US) by the Vanderbilt Comprehensive Care Clinic (VCCC) (GenBank accessions MH352627–MH355541. People were included in that study cohort if they aged 18 years or older and had more than one HIV-1 pol sequence sampled from 1977 to 2011 (Dennis et al., 2018) (Figure 2.1a).

The Beijing HIV dataset contains n = 25,648 HIV-1 *pol* sequences from the Beijing HIV laboratory network (BHLN) in China (accession numbers can be find in Appendix A). People were included if they were aged 18 years or older and had more than 1 HIV-1 *pol* sequence

(a)



(b)



(c)

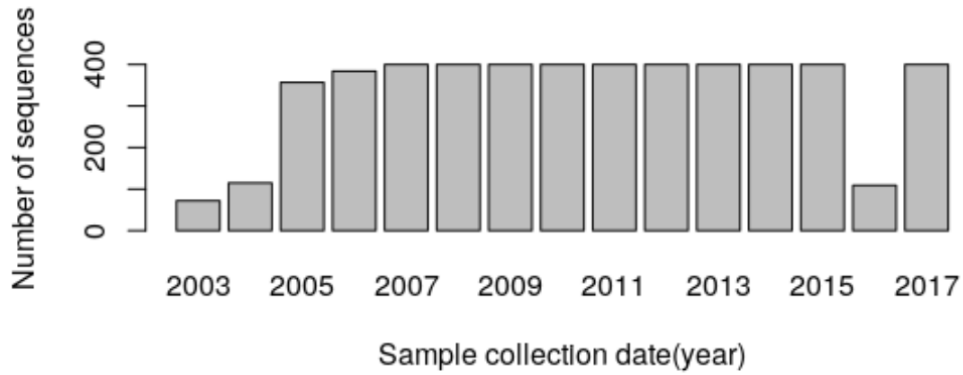


Figure 2.1: Bar plot of HIV-1 *pol* sequences data sets, (**a**) representing the distribution of sequence collection years for Tennessee HIV data from 1977 to 2011. (**b**) representing the distribution of sequence collection years for Beijing HIV data from 1991 to 2017 (**c**) representing the distribution of sequence collection years for Beijing HIV data from 2003 to 2017 after down sampling.

sampled from 1991 to 2020 (Ye et al., 2020). Data with missing dates, 8,314 sequences, were excluded from our analysis. There are fewer samples before 2003 among the resting 17,334 sequences (Figure2.1b), so in order to get a more uniform distribution of samples per year and reduce the size of the dataset, we down sample the data by taking a random subset from and reduce the size of the dataset, we down sample the data to 5,037 sequences by taking a random subset from year 2003 to the most recent year with maximum 400 samples per year (Figure 2.1c).

For the above HIV-1 pol sequences datasets, we first used an open-source program Multiple Alignment using Fast Fourier Transform, in short MAFFT (version v7.310; Katoh, 2017) to align the sequences. We then applied the Tamura and Nei (1993) genetic distance (https://github.com/veg/tn93) to compute the pairwise distances between all aligned nucleotide sequences. All options for MAFFT and TN93 analyses were set to the default values. TN93 result usually present in the form of pairwise distance list in a text file. For instance, "*KF267642-2010 KF267641-2010 0.0217375*" is one line in the TN93 text file, this means the sequence in column 1(*KF267642-2010*) have a pairwise distance of 0.0217375 with the sequence in column2(*KF267641-2010*). Furthermore, we wanted to get the sample dates from the original accession numbers. In R, we spited the sequences' information between "-" or "_" using the *strsplit()* function, and store all the new information separately into a new data frame as following:

$ID1	*$t1*	*$ID2*	*$t2*	*$Distance*
KF267642	*2010*	*KF267641*	*2010*	*0.0217375*

where $ID1 and $ID2 represent the sequences name, $t1 and $t2 represent the collection dates corresponding to ID1 and ID2 respectively, and $Distance indicate the pairwise

distance among these sequences. We can also treat this data frame as an edge list of a graph. For example, IDs ($ID1 and $ID2) are the nodes. Under a threshold *d*, for instance *d* = 0.0.03, an edge will be considered between these two nodes if the pairwise distance value ($Distance) is below *d*. An edge list can be further transformed into an adjacency matrix. An adjacency matrix A for a graph with n sequences can be defined as a square n*n matrix such that,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

the value a_{ij} represents presence of an edge from sequence i to j. With an adjacency matrix, we can generate a graph which will later be used for clustering analysis.

Additionally, with the knowledge of collection dates, sequences are separated into two subsets, “Known cases” and “New cases”. "New cases" are defined as the sequence collected in the most recent time interval (i.e., month or year), which will later be used to train regression models to predict the distribution of new cases among genetic clusters of known cases. Sequences in the most recent time slot are belonged to “New cases” subset, and “Known cases” subset contains all the remainder sequences.

2.1.2 Study Population and Data Processing of SARS-CoV-2 dataset

For global SARS-CoV-2 dataset, we obtained n= 64,143 genome sequences from GISAID (Global Initiative on Sharing All Influenza Data http://gidaid.org) in year 2020(before September 27th). The first few steps are similar to the CoVizu project (Ferreira et al., 2021) We aligned each genome to the WH1 reference genome (GenBank accession NC 045512; Wu et al., 2020a) using the program minimap2 (version 2.17; Li, 2018). Minimap2 is a fast sequence mapping and pairwise alignment tool for nucleotide sequences. Next, we used a Python script to find all mutations (insertions, deletions and nucleotide substitutions) from the WH1 reference genome and set them as "features". All features result were stored into a JSON file. And then, using a Python script on the JSON file, all genomes with identical sets of features were grouped into a single "variant". We labeled the variant and store all sequences result in a CSV file. For instance, here is is one line of the CSV file:

"*hCoV-19/Australia/NSW2608/2020| EPI_ISL_500717|2020-07-25,32515"*.

"*hCoV-19/Australia/NSW2608/2020|EPI_ISL_500717|2020-07-25*" indicate the sequences information (accession number, sample date and sample collection date) and "*32515*" indicate the variant label. As many features were compressed into one variant, we use a Python script to select only one sequence with the earliest sample date for each variant. As we have discussed in the introduction chapter, SARS-CoV-2 compared to HIV-1 is very easily transmittable due to it being an airborne infection, and people may get infected within 2–14 days after exposure to the virus. Due to the much higher pre-exposure transmission rate relative to HIV, genetic clustering analysis of SARS-CoV-2 data tend to connect every sample to one large component, even if we use the lowest pairwise TN93 threshold that corresponds to a single nucleotide difference between aligned sequences. To avoid this, we

computed the Hamming distance regarding to aligned sequences as a genetic distance. Hamming distance counts a set of places are different, and which are the same. For instance, if phenom1 has features “1,2,3”, and phenom2 has features “1,4”. Both phenomes contain feature 1, and they have differences on feature “2,3,4”. The Hamming distance between phenom1 and phenom2 is calculated by counting the total number of differences on feature. In our case, they have a Hamming distance of 3.

Computing Hamming distances returns a result that be satanized in the form of an edge list. From here, we repeated our step on processing HIV data set: transformed edge list into an adjacency matrix, then with an adjacency matrix, we generated a graph which will later be used for clustering analysis.

2.2 Markov Cluster Algorithm

The Markov Cluster Algorithm (MCL) is an unsupervised algorithm based on the probabilities of random walks through the network, and it can simulate the general flow of a network (Stijn, 2000). “Flow” is a pattern simulated by realizations of a stochastic process, for example, the transmission rate between nodes within network. Mathematically, flows are modeled by performing algebraic operations on probability matrices associated with a graph. In addition to requiring a graph G, the algorithm takes two matrix operations called expansion *e* and inflation *i*. Expansion simulates the flow within a cluster, while inflation eliminates flow between different clusters.

Let M1 be the matrix of random walks on G. Expansion *e* represents taking the e^{th} column wise product power of M1 as M2. Inflation i represents taking the i^{th} entry wise

product power of M2 as M1. While M1 is not equal to M2, we repeat above step. If there is no difference between M1 and M2, we then have converged to an equilibrium state, which we then apply towards cluster extraction. This process can be written as pseudocode as follows:

```
G is a graph
set M_1 to be the matrix of random walks on G
while (change) {
M_2 = M_1 · M_1                # expansion
M_1 = M_2 ∘ M_2                # inflation
change = difference(M_1, M_2)
}
```

MCL has been applied in many different areas, mostly in bioinformatics. For example, it has been used in protein-protein interaction networks as an effective clustering approach (Rani et al., 2019; Shih & Parthasarathy, 2012). To date, there are more than ten-thousand papers citing MCL as their core method. MCL's source code in implemented in the C programming language and can be found at GitHub (https://github.com/micans/mcl); and there are also R packages contain MCL algorithm, for instance package mcl (https://cran.r-project.org/web/packages/MCL/MCL.pdf)

2.3 Louvain Clustering

The Louvain method is a hierarchical algorithm based on the optimization of modularity (Blondel et al., 2008). Modularity is a numerical measurement that represents the density of connections within a cluster for a given arrangement of edges in a network. It calculates the number of edges falling within groups minus the expected number of edges placed by chance. Modularity can be positive or negative, and its value usually falls in the range [-0.05, 1] for unweighted and undirected graphs. Having a higher positive value indicates that edges are more abundant within the cluster than expected by chance and the graph is more likely forming a community structure.

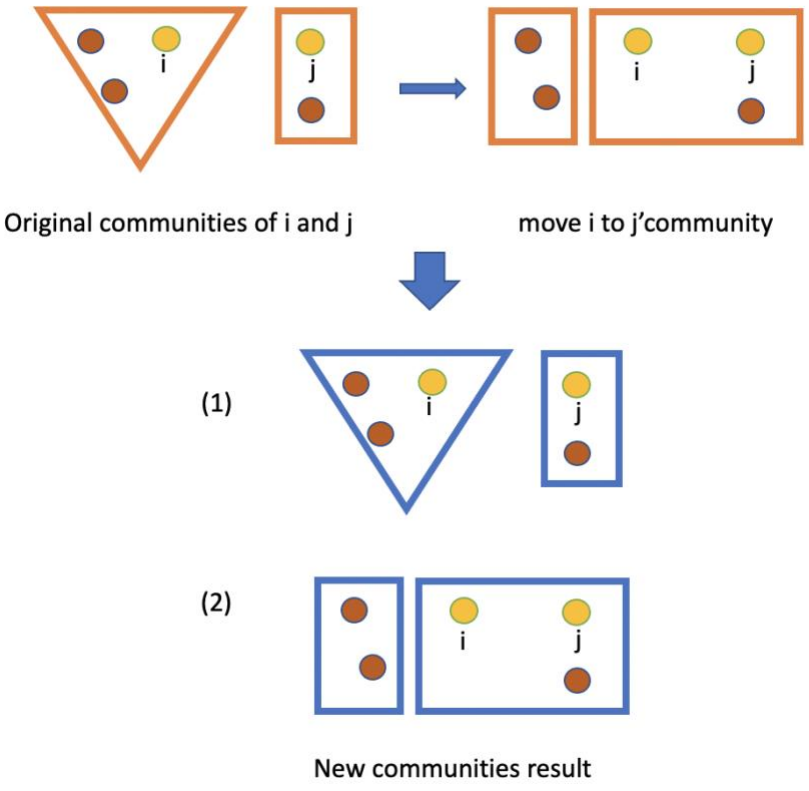


Figure 2.2: A graph representing the progress of evaluating the gain of modularity for node i and j(labelled in yellow). From the original community structure, we remove i from its community and by placing it in the community of j. Case(1) If this gain is negative, then i stays in the original community. Case (2) If the gain is positive, then node i is placed in the j community as the new community result.

This algorithm can be treats as two part that are repeated iteratively. Assume that we a network with N nodes. First of all, we assign an independent community to each node of the network. In other words, we start with an initial network with N communities. Then, for each node I, we find a neighbor node(have an edge connection to i) of i, say node j. We evaluate the gain of modularity that would take place by removing i from its community and by placing it in the community of j. If this gain is positive, then node i will be placed in the community for which this gain is maximized. If the gain is negative or zero, i stays in its original community (Figure 2.2). The second phase of the algorithm consists in building a new network by considering communities found in the first phase as nodes. Any connections between nodes within the same community are now represented as self-loops on the new community and connections between nodes from the different community are represented by weighted edges between new communities. Once the second phase has ended, the first phase will be re-applied to this new network.

2.4 Akaike Information Criterion

Akaike information criterion (AIC) is a mathematical measurement used to evaluate the quality of how well a model fits the data (Kiado, 1973). Suppose we have a statistical model of some data, then AIC can be defined as:

$$AIC = 2k - 2ln(L)$$

where k is the number of estimated parameters in the model and L is the estimated maximum value of the likelihood function for the model. Likelihood is defined as the probability of the data given the hypothesis model. It represents the objective function for estimating parameters of the model.

While a statistical model is used to implementing a data, there are always be some information lost during this process. AIC quantifies the information loss of a model and reducing information loss leads to better performance of a model. Compared to other information criteria, AIC tends to select a model that has higher dimensionality. Delta-AIC, i.e., the difference between the two model's AIC values, is one of the most popular measurements on model selection. Previous study has shown that if the delta-AIC is larger than 2 then we should consider one model is significantly better than the model it is being compared to (Burnham & Anderson 2004).

2.5 Framework Overview

Our framework's structure can be written as pseudocode as follows:

Input: TN93 edge list
Output: delta-AIC value respectively to each algorithm
(Connected component/MCL/Louvain)

*Generated a 3*N Latin hypercube sampling data set, with threshold d, inflation i, expansion e.*

For each run in N do:
1. *Filter edge list according to t. Separated nodes into "known cases" and "new cases".*
2. *Graph the known cases and cluster by connected components, MCL(i,j) and Louvain method.*
3. *For each cluster result, add new cases to their closest cluster.*
4. *Fit Poisson regressions and compute delta-AIC value*

We have multiple tuning parameters in the framework. To test result in the given parameter space, we apply Latin hypercube sampling (LHS) to get a series of parameter combinations. LHS is an algorithm for generating a sample of N points that are uniformly distributed in an N dimensional space. More precisely, LHS partitions each variable's range into N non-overlapping intervals based on equal probability 1/N. Every value for each interval is randomly chosen based on the probability density in that interval. Applying LHS to each parameter, threshold on TN93, expansion and inflation on MCL, we generated a parameter set with a size of 3* N as following (one sample row):

$TN93	*$expansion*	*$inflation*
0.03269472	*3*	*2*

We using *maximinLHS()* and set boundaries for each parameter using *lhs()* function in R. Then we use this parameter sets respectively on above steps. Further information for LHS parameter setting is summarized in the following table for reference (table 2.1).

Data Set	Threshold Range	Inflation Range	Expansion Range	Run Time(N)
Tennessee	0 – 0.6	2-25	2-25	500
Beijing	0 – 0.35	2-5	2-15	300

Table 2.1 A table representing parameters (threshold, inflation and expansion) bounders setting, and total run time for Tennessee HIV sequence data set and Beijing HIV sequence data set.

In order to further reduce the run time, we used parallel computing and cluster computing. Poon Lab operates a computing cluster called BEVi (Bioinformatics and Evolution of Viruses). BEVi combines a set of computers: there is a head computer node that distributes tasks to multiple children computer nodes, and there are children computer nodes that are able to handle independent task. More precisely, for each run time N, we separated every 100 runs to a child computer node. Furthermore, computations in R can be done faster using parallel computation. Parallel computation is the execution of breaking a larger computation to multiple computing cores. We use parallel computing to process our code with a usage of 16 cores in each computer node by applying *mclapply()* function in R.

TN93 gives an edge list as the input for our framework. We then can create a filtered edge list by using an optimal threshold d. Any TN93 pairwise distance below *d* would be marked as connected in the filtered edge list, and likewise, any pairwise distance above or equal to *d* would be excluded. Sequences will be separated into “known cases” and “new cases” based on sequences’ collection time. Next, by only using the sequences in the “known cases” subset, we generated an adjacency matrix from the filtered edge list to produce a graph. Two community detection methods and connected component method will be used to partition each graph into a set of clusters.

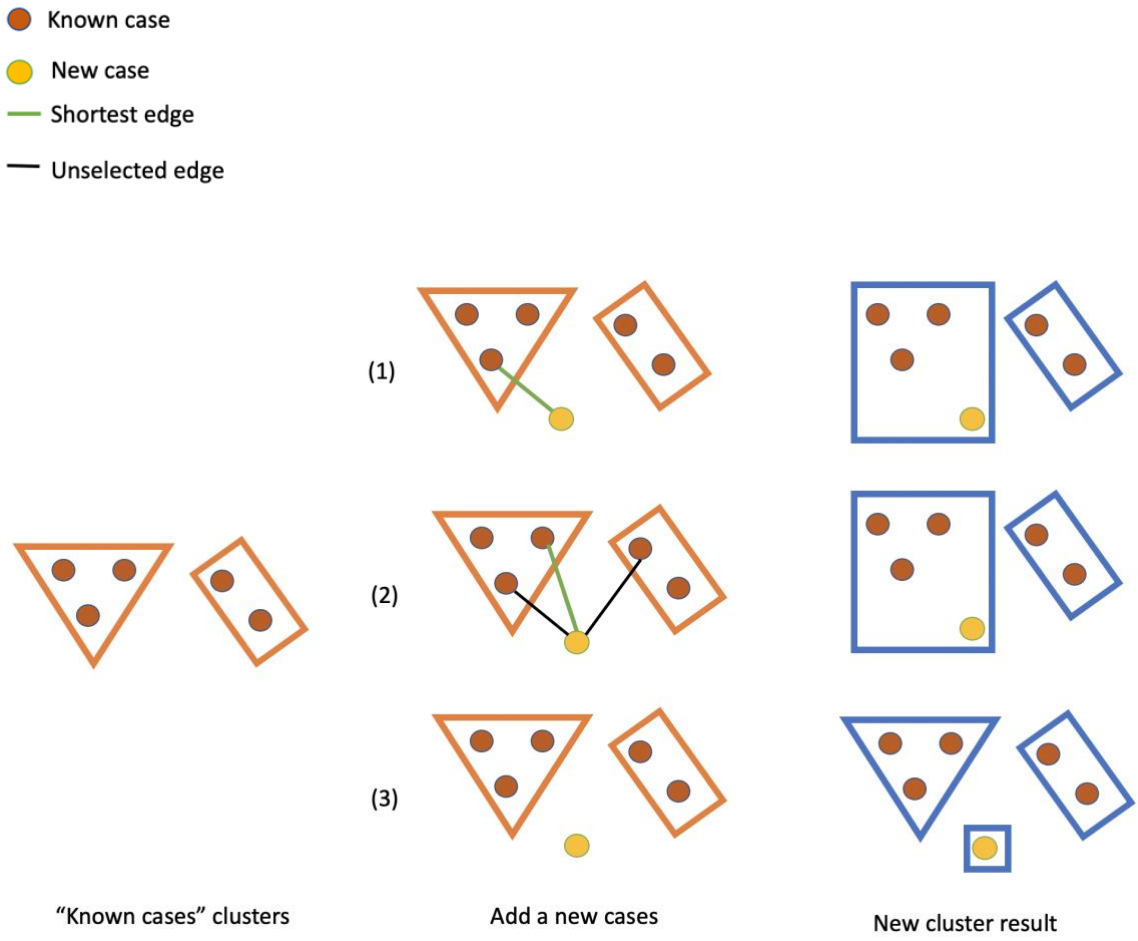


Figure 2.3: Three scenarios while inserting new cases to clusters formed by "known cases". Edges are unselected due to having larger pairwise distance than the selected edge. Case (1) the new node is only connect to one known case, thus the new node joins its cluster as the new cluster result. Case (2) There are multiple edges connected to the new case, we select one known case that have the shortest edge with the new case, and insert the new case to its cluster as the new cluster result. Case (3) If the distance is greater than current threshold for a given node, than the new node will not be attached to any clusters. In this case, the new case forms an induvial cluster in the new cluster result.

New cases are added to clusters by connecting each new case to the node which has the shortest TN93 pairwise distance (Figure 2.3). Some any new cases are not considered because they are not connected to any other nodes; specifically, any new case for which all its pairwise distances to known cases are above the current threshold *d*. To clarify, we only insert each new distances to known cases are above the current threshold *d*. To clarify, we only insert each new case to one cluster. There is the possibility that a new case has connections to multiple nodes and considering all these new edges might cause cluster to merge. To prevent this type of edge cases, we only consider one edge which represent the shortest distance to the new node. If a node doesn't connect to any other nodes below current threshold, we consider this node as a new cluster.

To evaluate the performance of three methods on charactering the transmission risk structure of virus epidemics, we want to estimate if one set of clusters is more informative than other set of clusters by examine at how adding information on the recency of known cases in each cluster affects the predictor number of new cases. We use Poisson regression model defined as:

$$log(y) = \alpha + \beta_1 X_1 + \beta_2 X_2 \,,$$

where y is the number of new cases in each cluster and X_1, X_2 are covariates used to predict outcomes. By using *glm()* function in R, we fit two Poisson regression models whose both outcome is total number of new cases in each cluster. For the null model L_{null}, we only take the number of known cases in a cluster as the only independent variable; and for the proposed model $L_{proposed}$, we add one additional independent variable that is how recent the known cases are. For each known case in the cluster, we take the difference of the most recent collection date with its sample collection date and sum them up to find the recency

for this cluster. To quantify the model information given by of a specific set of clusters we use AIC on previous Poisson regressions. We find the maximum log-likelihood estimation by sample mean of the n observations in the sample. Then we measure the difference of delta-AIC value (ΔAIC) between L_{null} and $L_{proposed}$:

$$\Delta AIC = AIC(L_{proposed}) - AIC(L_{null})$$

by using the $aic property in *glm()* function in R. Lastly, we find the value of d that minimize delta-AIC.

Chapter 3

Result

3.1 HIV result on Tennessee and Beijing date set

3.1.1 TN93 result and selected threshold range

The pairwise genetic distances of all sequences in the Beijing and Tennessee HIV data set were calculated by the TN93 method. The means of the distance among those two locations were 0.054, 0.056 respectively. Besides that, the medians of the pairwise distance were 0.053, 0.056, and the standard deviation among those two locations were 0.020 and 0.011. From the histograms of the pairwise distance of these two locations, we can see that the Tennessee data tend to be distributed symmetrically (Figure 3.1 top), however the Beijing data showed a right-skewed curve (Figure 3.2 top).

We ran a Shapiro test to a random sample of 5000 sequences for Beijing and Tennessee data sets, and both normality tests fail (p-value less than 0.0001). hence, we derived that both data are not sampled from a normal distribution. Then we applied the non-parametric pairwise ranked-sum Wilcoxon test to determine if they are from the same distribution. For pairwise ranked-sum Wilcoxon test, we sampled 100000 observations from each of dataset due to make sure the same length. The result shows us that the TN93 distribution for the Beijing data was significantly different ($p < 2 \times 10^{-16}$).

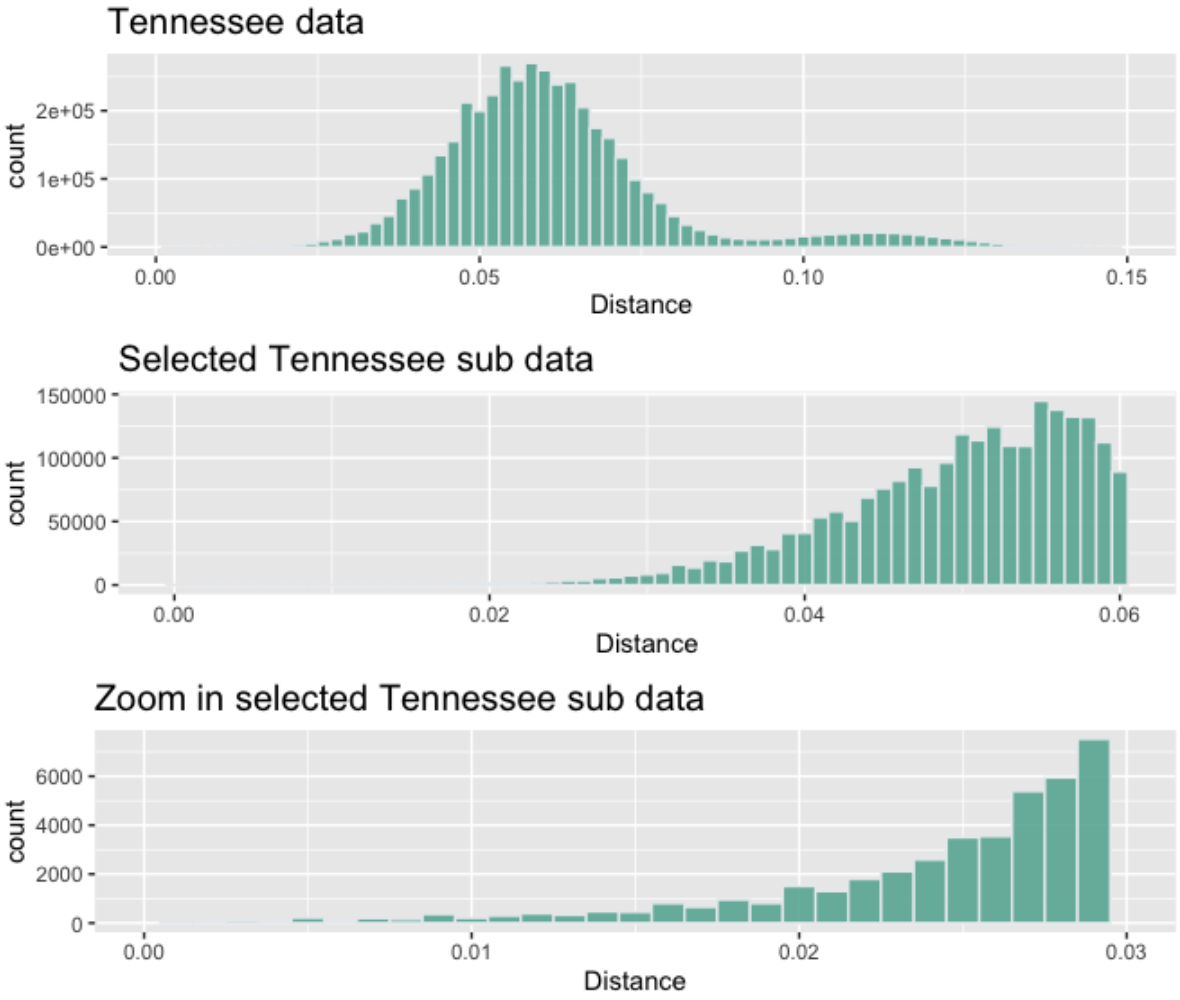


Figure 3.1: (**top**) Histogram, representing the distribution of pairwise TN93 distances for all sequences in Tennessee HIV data set. (**middle**) Histogram, representing the distribution of pairwise TN93 distances for the selected sub sequences (below 0.06) in Tennessee HIV data set. (**bottom**) zoom in on selected sub sequences (below 0.03).

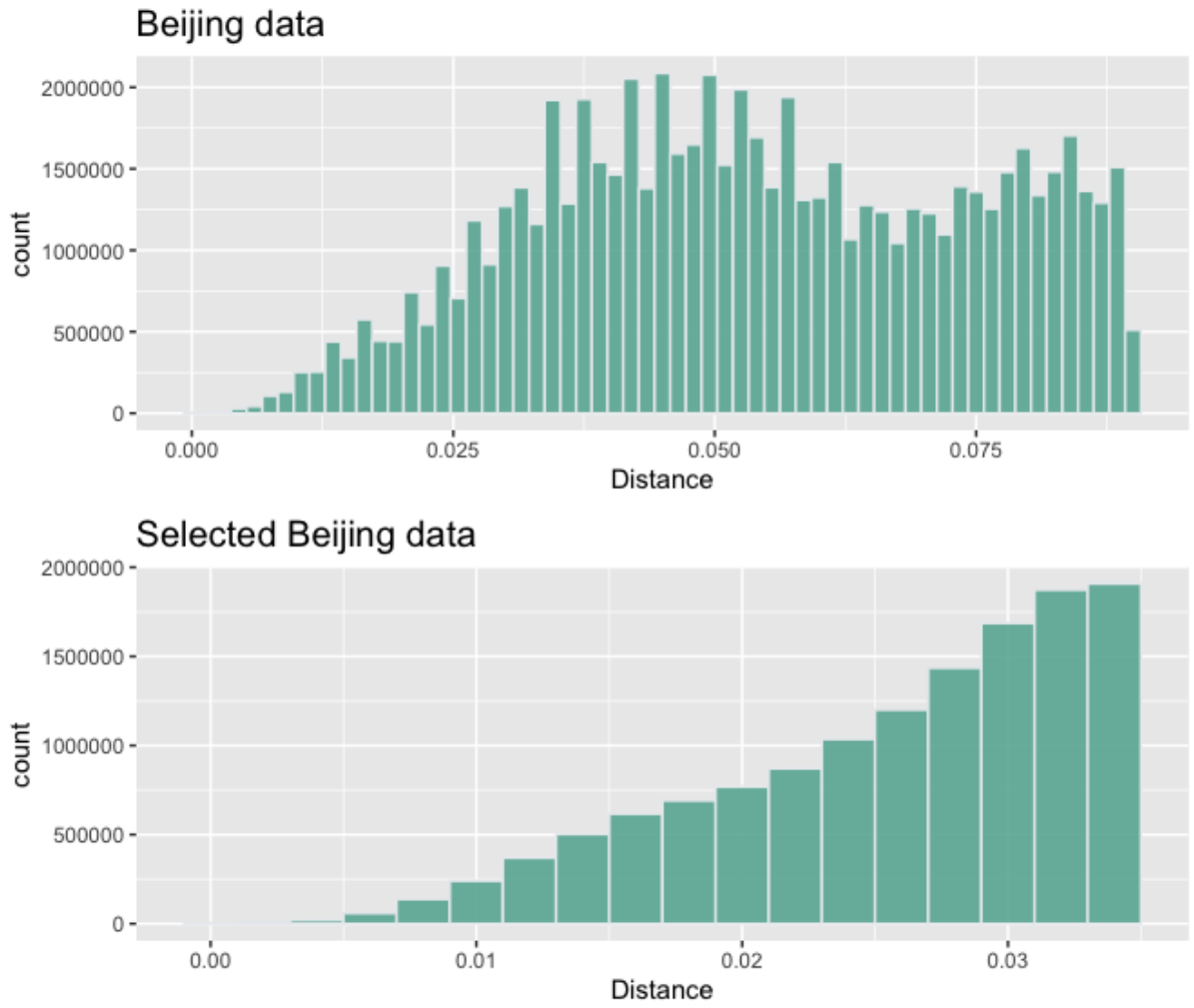


Figure 3.2: (**top**) Histogram, representing the distribution of pairwise TN93 distances for all sequences in Beijing HIV data set. (**bottom**) Histogram, representing the distribution of pairwise TN93 distances for the selected sequences (below 0.035) in Beijing HIV data set.

The TN93 distance can be computed quickly with a low memory. However, constructing the network from these distances can be computationally heavy since the number of edges grows rapidly with the number of nodes. To reduce the memory requirements, we excluded pairwise distance exceeding a threshold of 0.06. Furthermore, we randomly sub-sampled the data sets to reduce the computing time and decease the input of sample size. For Tennessee HIV data set, we selected the sequences pairs which have a pairwise distance below 0.06(Figure 3.1 middle and bottom). For Beijing HIV data set, a much bigger sequence data set, we tried multiple strategies to reduce the running time. First of all, as I have stated in method chapter, we excluded sequence samples before 2003 and randomly selected 400 sequences for each year. Secondly, we tried varying threshold range with AIC and finally narrowed the threshold range to pairwise distance below 0.035(Figure 3.2 bottom)

3.1.2 Connected Component result at 1.5% and 3% threshold

In most of the previous HIV studies, the standard pairwise distance thresholds used for connected components-based clustering methods is 0.015. At this threshold, we obtain 253 connected components in Tennessee HIV data set, of which 150 (59.3%) clusters are in pairs, 87 (34.4%) clusters contain more than two but less than ten sequences, and 16 (6.3%) clusters have more than ten sequences. Generally speaking, beside a large connected component with a cluster size of 299, smaller connected components are formed at lower distance threshold (Figure 3.3 left).

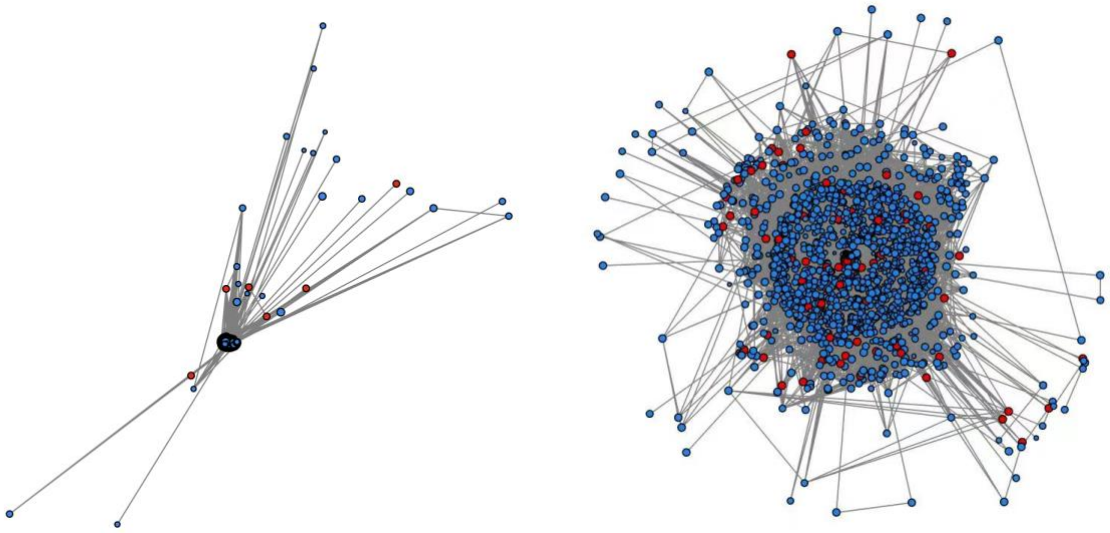


Figure 3.3: Graphs created from Tennessee HIV data set at threshold 1.5%(**left**) and 3%(**right**). All sequences are represented by nodes and colored differently in known cases(blue) and new cases(red). Grey lines indicate theTN93 pairwise distance between two nodes is lower than the current threshold. Graph excludes all unconnected nodes.

Furthermore, at threshold of 0.015, fewer cases are included to the graph, especially new cases (Figure 3.4). When the threshold is very small, the new cases are unlikely to be connected to known clusters (all nodes almost isolated) so the new cases detection rate is almost 0. As the threshold increases, the new cases more likely to be connected to a known cluster. For the traditional selection of threshold of 1.5%, we can see that only around 60 of new cases are connected to known cases. When the threshold increases to 3%, around 120 of new cases are connected, which is double the size. Thus, if we want to include more sequences to the model, we could increase the threshold from 1.5% to 3%. At this threshold, among all 73 components, 42(57.5%) clusters are in pairs, 28 (38.4%) clusters contain more than two but less than ten sequences, 3(4.1%) clusters have more than ten sequences. Among the clusters with >10 known cases, there is one much larger connected

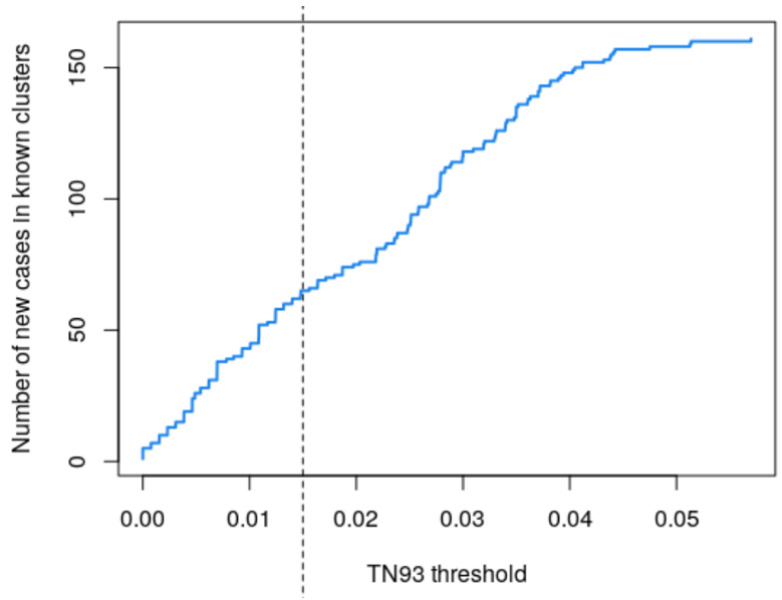


Figure 3.4: Step plot, representing the number of new cases corresponding to TN93 threshold. The vertical line marks the 0.015 value in TN93 threshold.

component with a cluster size of 2051, nodes which is larger than the giant component we obtain at the 1.5% thresholds (299 nodes). Generally speaking, fewer larger components are formed at higher threshold because nodes tend to span their neighbor at this threshold. Hence, more and more nodes are connected with each other, and the proportion of unconnected nodes decreases. As a result, small clusters merge with each other, and most cases collapse into a single giant component (Figure 3.3 right).

3.1.2 Connected Component and Community Detection Clustering Result on Varying Thresholds

Using our framework, we test the performance of the connected component clustering method with two different community detection methods. For each clustering method, we calculate the AIC difference of the null model and the proposed model. We fit two Poisson regression models to the observed variable, which is the outcome is number of new cases in each cluster. The null model only takes the number of knows cases in the

cluster as the independent variable. The proposed model incorporates an additional independent variable, namely how recently the known cases were sampled from the population. The MCL algorithm did not converge for all parameter settings and there are no patterns to parameter setting that MCL failed to converge for. Hence, here we only compare the runs that all methods have valid clustering results. The resulting AIC loss was calculated by the difference of AIC in proposed model and AIC in null model is shown in Figure 3.5. As we have mentioned in previous chapter, if the delta-AIC is larger than -2 then we should consider proposed mode is significantly better than the null model. If the delta-AIC is near zero, then there is weak evidence for choosing proposed model than null model.

Among all 500 runs in the Tennessee HIV data set, there are 265 runs converge under MCL algorithm (Figure 3.5 left), each curve representing a method's AIC loss as a function of the TN93 threshold. For the connected components-based clustering, which is represented by the red line, the delta-AIC at first decrease with increasing threshold, reaching the lowest delta- AIC at -32.23 with the threshold equal to 0.02. Then it starts to increase with higher thresholds, the result of connected component base clustering eventually approaching delta-AIC = 0 as the threshold approaches 0.38. For the MCL clustering method, represented by the blue line, the delta-AIC has a similar pattern as the connect components method. However, the tread in delta-AIC is steeper for MCL than the connected component around the minimum delta-AIC point. It decreases when the threshold approaches 0.03 and increases after this point, eventually converging to 0. For the Louvain method, which result is represented by the yellow line, shows a similar pattern as MCL. At first, the delta-AIC shows a decreasing trend until 0.22, and then it starts to increase at 0.29. Its increasing tendency stops when threshold approach 0.3. Then the delta-AIC value tends to be stable

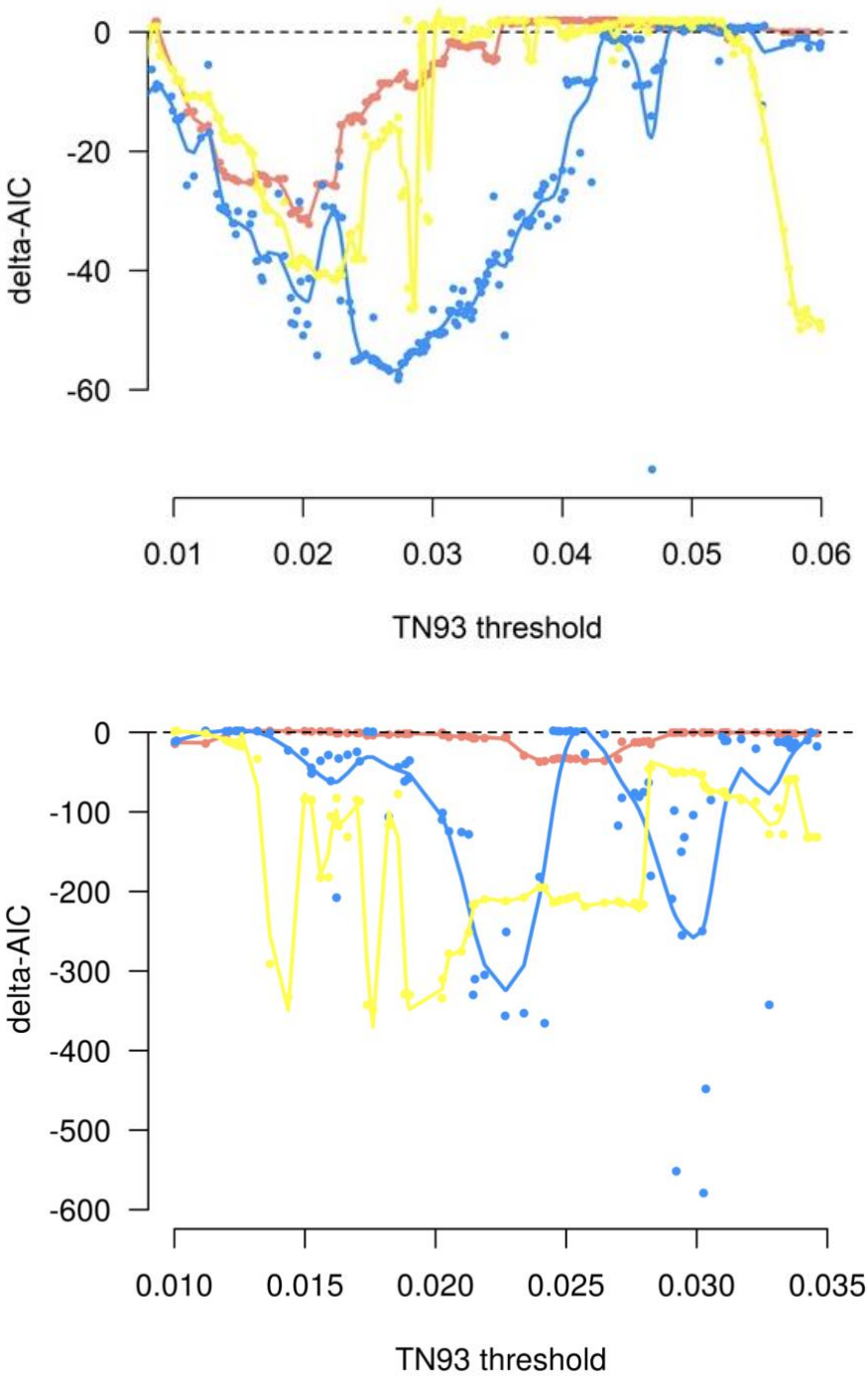


Figure 3.5: The AIC loss for predictive growth models corresponding to the TN93 thresholds for Tennessee HIV data set(**top**) and Beijing HIV data set(**bottom**). The AIC loss is calculated between a proposed model and null model.

around 0.3 to 0.5. There is a huge drop of delta-AIC value to -50 at its minimum threshold at 0.06.

For the Beijing HIV data set, the MCL algorithm was less likely to converge, with only 83 valid runs over 300 attempts (Figure 3.5 right). For the connected components base clustering, which result is represented by the red line, all delta-AIC values are close to 0. It has a sudden drop begin from 0.0225 and reaches its lowest point at 0.025 with a delta-AIC about -40. Both the result of MCL and Louvain clustering, representing in blue and yellow lines, are not continuous because we only look at the runs with valid MCL runs. By observing the existing points, both MCL and Louvain change rapidly with huge raise and drop. Additional, both lines basically lay below the line of connected components. MCL contain a few extreme delta-AIC values, and some of them almost reach -600. This supports the observations we made in Tennessee HIV data set, that the MCL and Louvain clustering method have larger maximum delta-AIC value, and wider AIC loss range.

There are two parameters for the MCL algorithm that are repetitively called expansion and inflation. To visualize on how these parameters affect AIC loss with varying distance thresholds, we used the contour plot to display trends in delta-AIC as a function of distance threshold and ether expansion or inflation. A contour plot is used to represent a three-dimensional surface. For a given value of z, lines are drawn for connecting the (x,y) coordinates where that z value occurs. In detail, we apply the thin plate spline method implemented in the R package fields, *Tps()* function, to smooth the observed distribution of z as a function of x and y. In our case, we use threshold as the x coordinate, either inflation or expansion as the y coordinate, and the delta- AIC as the z value. We plot the contour.

plots result of thin plate spline by using *surface()* function in R. A bluer area indicates a more negative delta-AIC value, and in contrast, redder area indicates delta-AIC is close to 0.

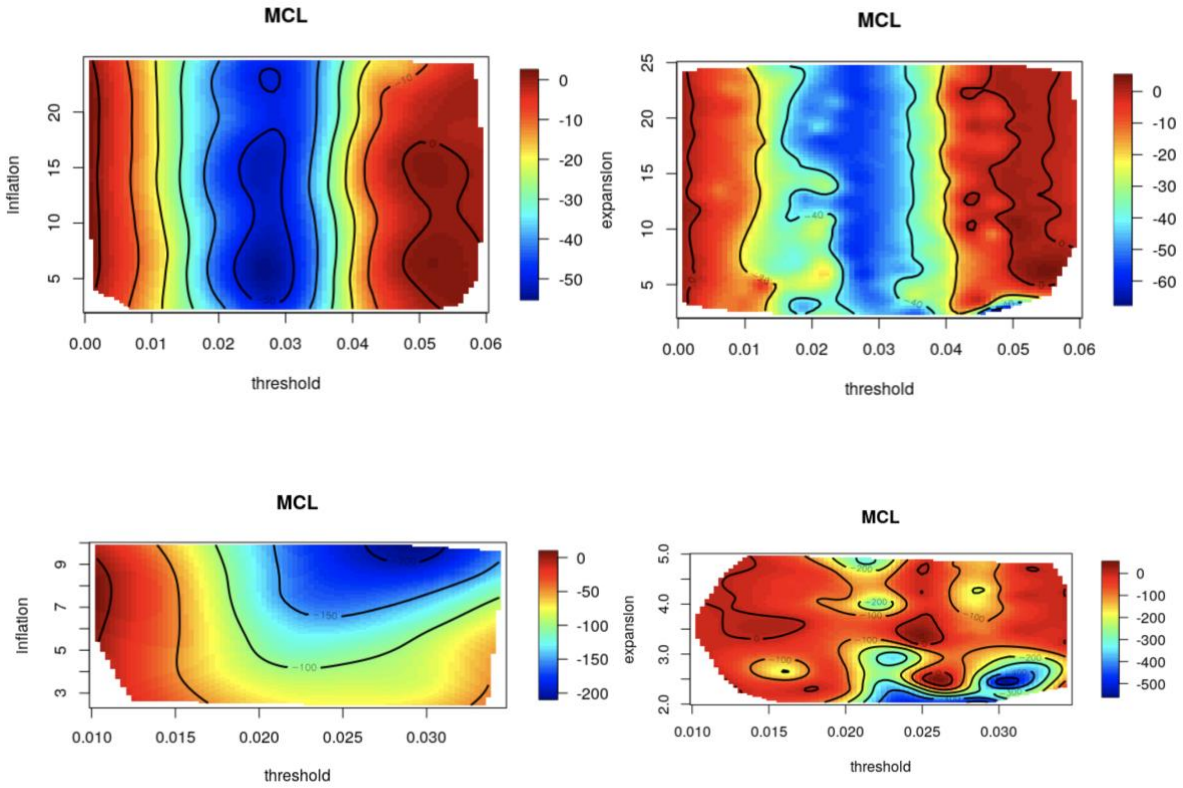


Figure 3.6: Contour plot, representing the AIC loss corresponding to expansion or inflation with varying threshold for Tennessee HIV data set (**top**) and Beijing HIV data set(**bottom**). A bluer area indicates a more negative delta-AIC value, and in contrast, redder area indicates delta-AIC is close to 0.

In Tennessee HIV data set, the delta-AIC surface shows almost a vertical pattern for the contour plot when y represents inflation, which illustrates that the inflation parameter has no effect on the value of delta-AIC (Figure 3.6 top left). Compared to inflation, we observe that expansion has a more measurable effect on the response surface. We can see that the valley of the surface skews to the right when expansion is less

than 15, such that there is a small blue area when expansion is around 2 with the threshold range between 0.04 to 0.05 (Figure 3.6 top bottom). This suggests that lower expansion values enlarge the difference in AICs under a certain level. Since most of the MCL results did not converge, the delta-AIC surface obtained for the Beijing HIV data set was not as continuous as the Tennessee HIV data set. As we observed in Tennessee HIV data set, delta-AIC was insensitive to inflation (Figure 3.6 bottom left). And again, we observe more negative delta-AIC value when expansion is around 2, with the threshold range between 0.02 to 0.035 (Figure 3.6 bottom right). This also supports the result we have from the Tennessee HIV data set that delta-AIC is slightly responsive to smaller value of expansion. Additionally, even though most of the delta-AIC is between -2 and 2, the delta-AIC range for these two data sets performed quite differently. The Tennessee HIV data set has a narrower AIC loss range, from 10 to -80. In contract, the Beijing HIV data set has a wider range, from 10 to -600. Since lots of points are missing, contour plot is hard to compare the relation of expansion and inflation (Figure 3.6 bottom left), it would be clearer to compare it by fitting smooth spline on both parameters with delta-AIC value (Figure 3.7). We observe that the tread on delta-AIC value is overall increasing while the expansion value goes up. On the contrary, tread on delta-AIC value is mostly decreasing while raising the inflation value.

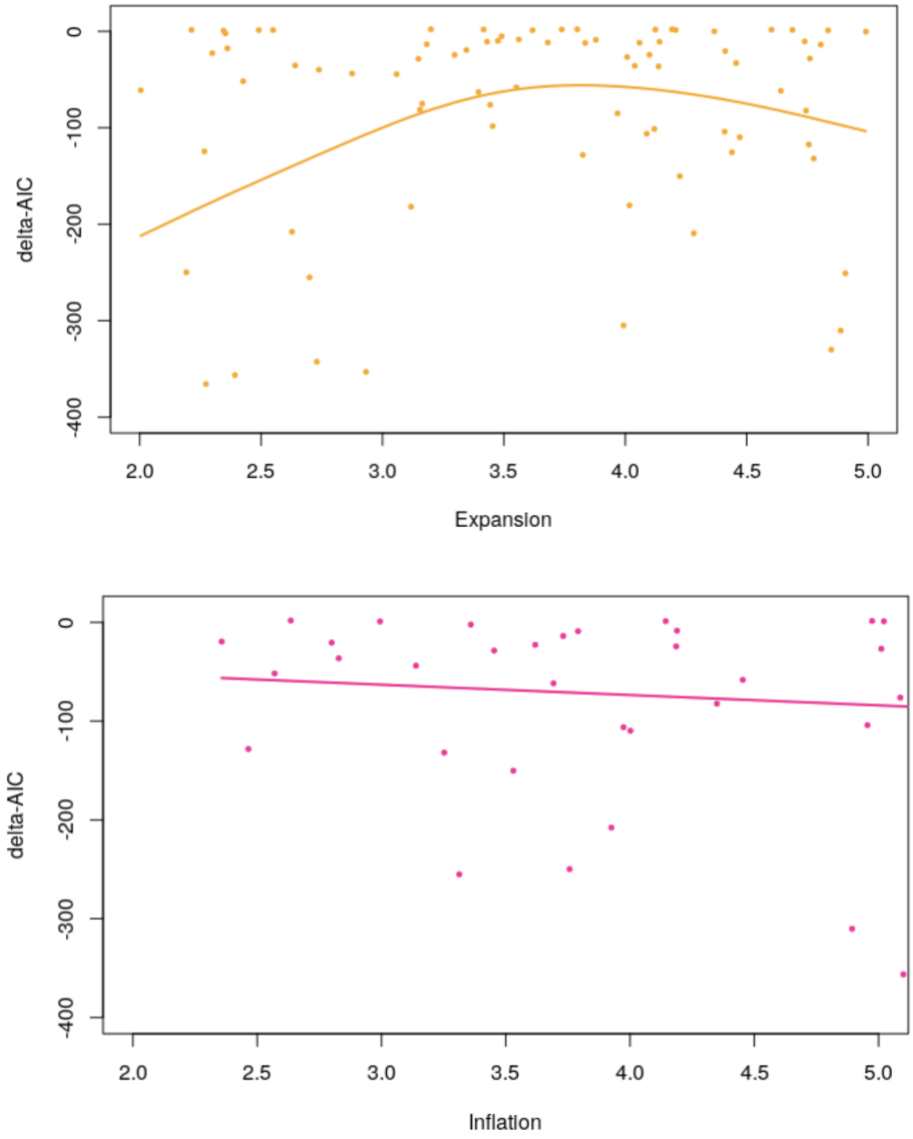


Figure 3.7: Fitting smooth plane on MCL parameters (expansion, inflation) with delta-AIC value in Beijing HIV- 1 *pol* sequences data set.

3.2 SARS-CoV-2 Result

3.2.1 Hamming distance result

The number of mutation difference between all pairs of sequences(n=64143) in global SARS-CoV-2 are calculated by Hamming distance which has been introduced in the previous chapter. After selecting only one sequence from each variant by the earliest collection date, we obtain 32515 sequences. In terms of edge list, there are over 520 million edges in total. This number is too huge to visualize as a graph. The mean of the Hamming distance is 12.5 mutations, and the mode difference between sequences is 11 mutations. While taking 1 mutation as the threshold, the graph contains 19,741 edges. If we raise the number to 2 mutations, this number rapidly increases to 1,472,042 edges. Among all these 520 million possible edges, around 157 (30%) million edges have a mutation difference less than 10. The minimum difference is 1 mutation, and the maximum difference is 311 mutations. For the consideration of the computational running time, we only consider the sequences with 1 mutation difference in our result chapter. Study have showed that maximum spanning tree could demonstrate the skeleton of the graph (Nguyen & Do, 2015). We have tried to use maximum spanning tree to further narrow the data, but only small number of sequences are affected. For example, when we applied a maximum spanning tree to the graph, 960(4.9%) edges out of 19,740 edges are excluded by a Hamming distance of 1 mutation.

3.2.2 Clustering result

The MCL algorithm did not converge for the graph when Hamming distance is equal to 1, thus we only apply the Louvain clustering method to compare with connected component method. In total, 19845 nodes are counted into the framework. There are 1520 clusters produced by the connected component clustering method, 896 (58.9%) clusters are pairs, 550 (36.1%) clusters contain more than two but less than ten sequences, 68 (4.5%) clusters have more than ten sequences but less than hundred sequences, and 6 (0.4%) clusters have over hundred sequences. There are two large connected components with a cluster sizes of 4129 and 9803 respectively. Compared to the connected component method, Louvain clustering results have more clusters and with smaller cluster size (Figure 3.8). There are 1610 clusters in Louvain clustering result, 896 (55.7%) clusters are in pairs, 560 (36.1%) clusters contain more than two but less than ten sequences, 115 (7.1%) clusters have more than ten sequences but less than hundred sequences, 39 (2.4%) clusters have over hundred sequences. There are only two clusters exceeding one thousand sequences in size, with a cluster size of 1392 and 1728 respectively. In general, the clustering results of smaller components, which have less than ten sequences, are not affected by community detection method. However, Louvain community detection method breaks large components into multiple smaller clusters.

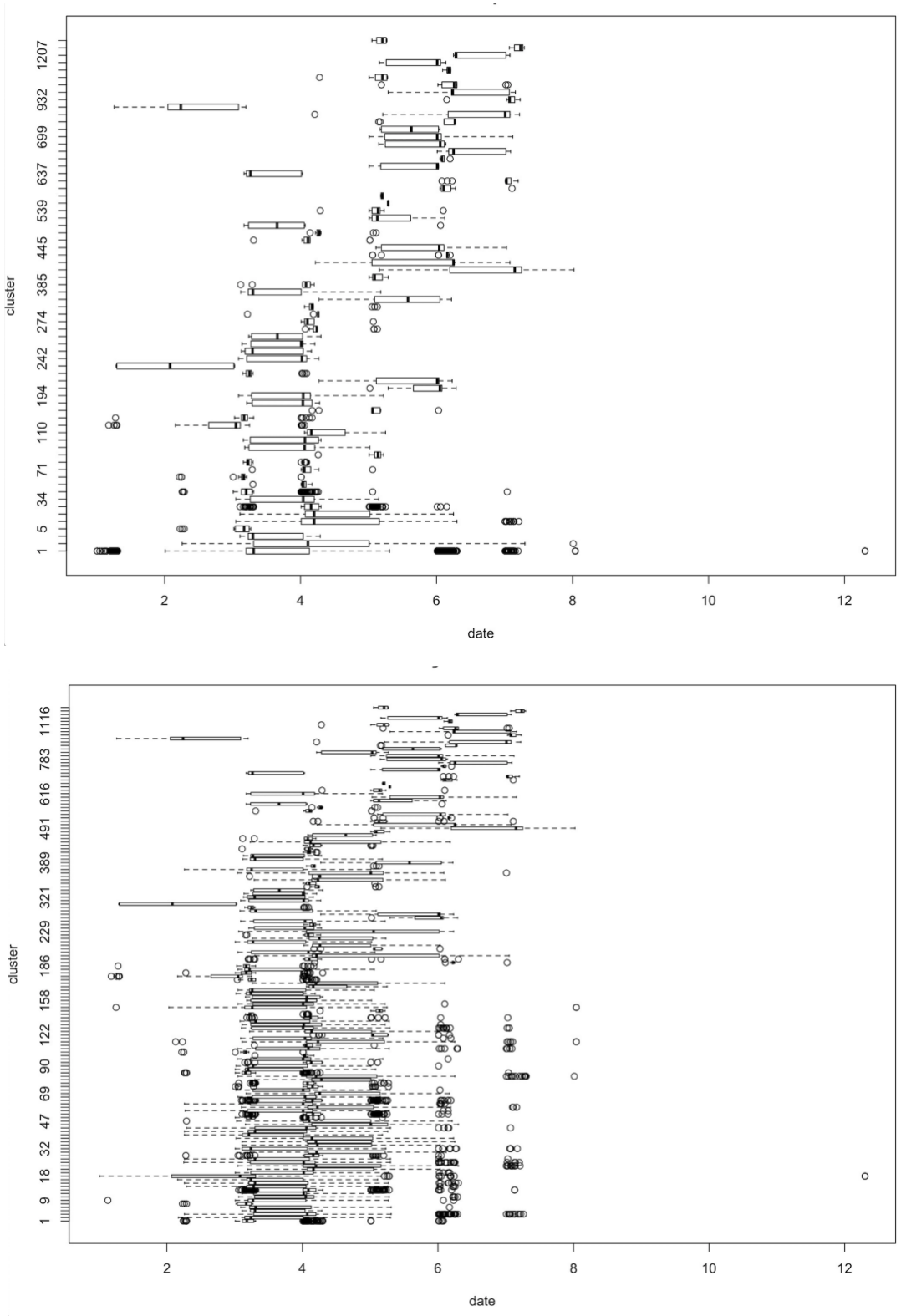


Figure 3.8: Box plot, representing the cluster result of connected component clustering method(**top**) and Louvain community detection clustering method(**bottom**). Only clusters with more than 10 sequences are shown here.

3.2.3 Clusters and Time

For better understanding, we converted the sample collection date to a numeric value. For instance, if the collection dates for a sequence is March 4th, we convert the date as 3.04. The distribution of sampling dates for clusters of size ≥ 10 are summarized in Figure 7. And as we have concluded, Louvain community detection have more medium size clusters.

Furthermore, with the information of sample date, we want to determine if nodes within clusters have significantly more earliest collection dates than expected by chance. We test this by using permutation test. We first calculated the average time for each cluster by summing the total time of every node and dividing it by total number of nodes in that cluster. Secondly, we randomly shuffled the nodes among clusters and calculated the new average time. The principle of this permutation test is keeping the total cluster number and each cluster's cluster size unchanged, and then randomizing the cluster membership of each node. Additionally, instead of doing one permutation test, we throwed 500 permutation test for each clustering method and took the average number of it. Then we used ANOVA test on each clustering method and its permutation test. Before the permutation, the F value for connected component is 3608, and 4316 for Louvain method. After the permutation, the average F value for connected component is 0.86, and 0.95 for Louvain. Among all 500 permutation runs for both clustering method, most of the F values fall into the range from 0 to 5 (Figure 3.9). This result suggests that collection dates are significantly correlated within clusters.

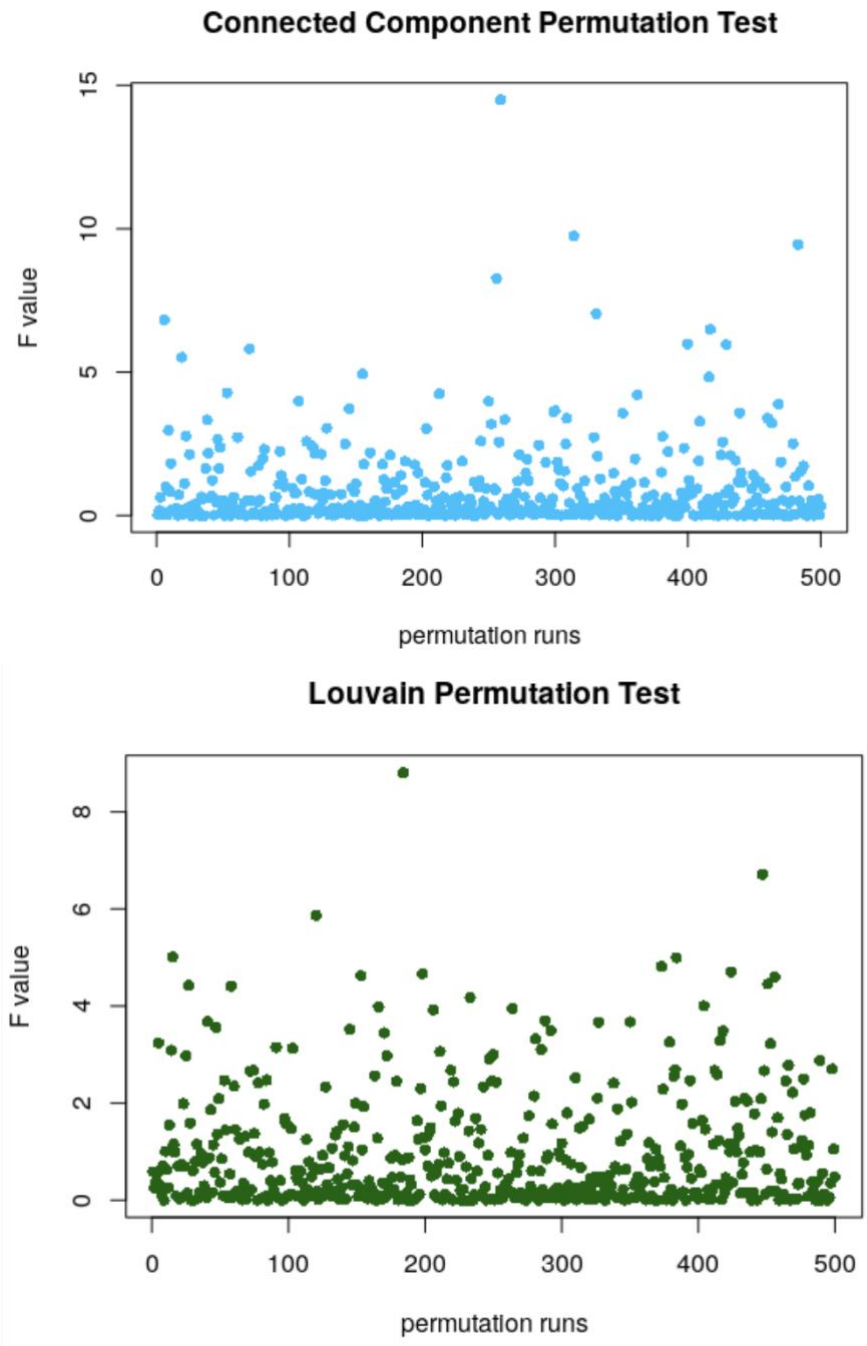


Figure 3.9: Scatterplot, representing 500 permutation tests' F value for connected component clustering method(**top**) and Louvain community detection clustering method(**bottom**) by using ANONA test.

3.2.4 Countries Correlation

Beside the relationship between cluster and time in the SARS-CoV-2 transmission network, we also want to determine if some pairs of countries are more significantly associated than by chance. We used Spearman coefficient to calculate the correlation. The Spearman correlation coefficient, named after the British statistician Charles Spearman, is defined as the Pearson correlation coefficient between the rank variables. The Pearson correlation is a statistical measure of the strength of the linear relationship between two random variables, while the Spearman correlation examines the strength of the monotonic relationship between the two. The Pearson correlation coefficient is calculated using the data sample value itself, while the Spearman correlation coefficient is calculated using the data sample rank value. The correlation coefficient ranges between 1 and -1; if a correlation coefficient is greater than 0, that indicates that there is positive correlation between two observation and vice versa; if the correlation coefficient is equal to 0, that indicates there is no correlation between them.

There are 124 countries represented in our global SARS-CoV-2 data set. Hence here we are only showing the correlation result of top twenty counties with the largest sample size (Figure 3.10). From the connected components correlation plot, we observe that most of the countries barely have any correlation with each other. There are some countries show a strong negative correlation, for instance, England with USA. For the Louvain method, we can observe that the pairwise correlation plot tends to be more positive in general, which suggest that more counties under this clustering result tend to appear together. We expect this result due the ability of community detection methods to partition large

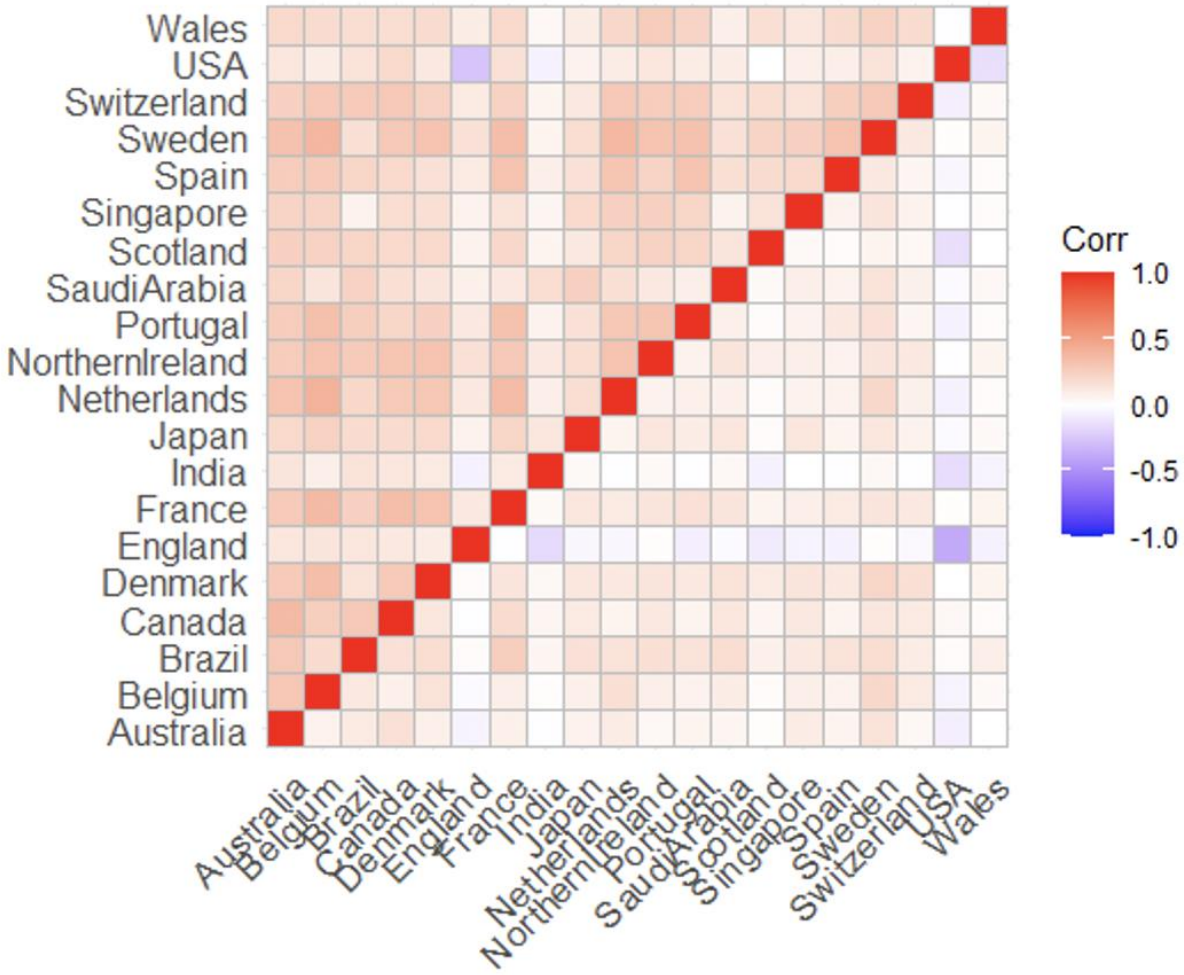


Figure 3.10: Correlation plot, representing the association between clusters and 20 countries with the largest sample size. Deeper red or blue indicate stronger positive or negative Spearman's correlations. Spearman's correlations for connected components are shown in lower-right, and the Spearman's correlations for Louvain method is shown in upper-left.

components into smaller clusters that may reduce novel patterns. The severely reduced size for the majority of clusters limits our ability to detect correlations with respect to countries of sampling. Connected component methods have fewer clusters, and 95% of the clusters have a cluster size less than 10. If most of the clusters only contain a few sequences, we are less likely to combine sequences from different countries, and this led to less information we can use while looking for correlation between countries. Compared to connected component method, the Louvain clustering method yields more clusters. Since Louvain method can break a large component into multiple communities, Louvain cluster results contain more clusters with intermediate cluster sizes (cluster of cluster size between 10 to 100 sequences), and this increases the frequency of countries appears together.

Chapter 4

Discussion

4.1 HIV Result Comparison

HIV virus have fast evolving genomes and the evolution is determined by its transmission. Therefore, phylogenetic reconstruction is commonly used to retrace the transmission events. The Tennessee dataset I have mentioned in previous chapter have been used in two studies (Dennis et al. 2018, Connor 2020). Both studies used clustering methods to evaluate HIV transmission patterns. Under this shared aim, the studies differ in how they transmission clusters using different parameter settings during the process. Both previous studies used a tree-based algorithm called maximum-likelihood (ML) phylogenetic tree to construct a tree, however the methods they used to build the maximum likelihood tree is a bit different. Dennis used software called FastTree, and the clusters in the tree are defined as patristic distance differences ≤1.5% with at least 2 individuals. Patristic distances describe the amount of branches length that that between two nodes in a tree. Our result in 1.5% threshold in previous chapter agreed with its result, major clusters were in pairs and only few of the clusters contained more than 10 persons. In contrast, the method used in Connor's master thesis is called IqTree, and the clusters in the tree are defined as the same as the setting in my study, which TN93 pairwise difference less or equal than varying threshold and with at least 2 individuals. Notes that if the pairwise distance of genomic

sequences increases, the individuals are further apart from the epidemic (Salamat et al, 2021). Specifically, for those sequences obtained from a group of people that experienced the same outbreak, the evolutionary distance was low. On the other hand, if the group of people were in different epidemics, the evolutionary distance was high. Interestingly, all these previous studies pattern clusters only by connected component method.

Not only different in using other clustering method than connected component method, but my work also applying predictive growth models on clusters as I have demonstrated in method chapter. Most of the previous studies focused on whether if element like age, sex, age, race/ethnicity, and country of origin are significantly associated with sequence clustering. Furthermore, a cluster that contains more known cases does not necessarily have a higher possibility of having more new cases. Dennis' work treated the connections between all individuals as potential routes of transmission, so we cannot compare our predict result with it. The Beijing dataset was used in testing for transmitted HIV drug resistance in china's province such as Beijing and Hubei province (Ye et al. 2020), a maximum likelihood phylogenetic tree was reconstructed in order to define clusters. The study used a subset of patients who had been recently infected with HIV within one year, and instead of predictive the growth, they simply repeated their analysis with this subset. Fortunately, Chato's thesis work calculated the AIC loss by predictive growth models, so it is easier to compare our AIC result, for instance the place where threshold corresponding with the lowest delta-AIC and the trend of delta-AIC with changing in threshold. Our clustering result in Tennessee HIV data set while using connected component support what Connor previous work have.

As I have analyzed in previous result chapter, Beijing HIV data set had a higher optimal threshold than Tennessee HIV data set. There are several possible speculations about this phenomenon. For instance, this may indicate that the sampling fractions, which is the sample size divide by the total number of HIV infections, is lower in the Beijing data set. Another possible explanation is that the majority of people might find to be diagnosed sooner after HIV infection in Tennessee data set. On the other hand, there are many possible consequences of having a higher threshold. Generally, a higher threshold would lead to a larger cluster size since more cases are joined to the network. A larger size on cluster size could lead a harder prediction. Consider two clusters that both have their own feature, are formed into a new cluster while more cases come into the graph. The features used to outstand in old clusters are likely to be averaged out in this new cluster while all the old features competing with each other.

4.2 SARS-CoV-2 Result Comparison

SARS-CoV-2 as an acute infection which is known by rapid onset of disease, is quite different from HIV-1 being a chronic infection. Even through there is a declines in HIV transmission rates between 1992 to 2005, from 20.3% to 2.9% (Park et al., 2010). Compared with HIV-1, SARS-CoV-2 has a much higher transmission rate. The transmission rate of SARS-CoV-2 is between 0.19-0.29/day (Romero-Severson et al., 2020). With a higher transmission rate, more people get infected, and this led to more sequences joined in the data set. However, while the mutation rate remains the similar, the sequences tend to remain identical as the result. Another difference between HIV-1 and SARS-CoV-2 data set is that enormous amounts of SARS-CoV-2 sequence data being collected in relatively short period of time. Massive and identical SARS-CoV-2 sequences make the SARS-CoV-2 transmission network even harder to contrast by using component clustering method.

To better understand the mutational trends of SARS-CoV-2, two types of popular methods have been used to clustering SARS-CoV-2 genetic transmission network. One type is similar to the method that I mentioned in HIV result comparison section. For instance, one study build a pipeline which involving sequence alignment with MAFFT and constructing a maximum-likelihood phylogenetic tree (Yang et al. 2020). However, tree-based methods can be computationally heavy as the size of sequence data continues to increase with time. For this reason, researchers have started to consider faster clustering method form the field of unsupervised machine learning. For example, K-means is a commonly used unsupervised machine learning clustering method, with its performance is not sensitive to sample size unless the sample size is too small. Taking advantage of this property, Hozumi et al. applied k-means clustering, combining it with a few dimension reduction algorithms to large-scale

SARS-CoV-2 datasets (Hozumi et al. 2021). Compared to our community detection cluster results, k-means clustering tends to form fewer large clusters with higher cluster size. The number of clusters(k) must be manually set be users, there are mostly four to six clusters are formed (Hozumi et al. 2021, Fidan et al.2022). These values of k were substantially smaller than the number of clusters I obtained from my analysis.

Sex, total number of infections, population density of a location, average age, mortality, and environmental variables are commonly considered as risk factors while clustering SARS-CoV-2 sequence. There are some studies consider the correlation between countries, for example, Nunes et al. extract a large Maximum Likelihood phylogenetic tree of the SARS-CoV-2 variants circulating in South America, China, India, and the USA. This study partitioned infections in above countries into two sets, infections in South America countries, and infections in Brazil, China, India and the USA (Nunes et al. 2022). In contrast, we are using a global data set representing samples from 124 countries. One of my results underline the odds while some pairs of countries are more likely to appeared in the same cluster by chance.

Collection dates of sequence data are often used while reconstructing the phylogenetic tree relating common ancestors to present-day species. On the other hand, some studies have also used collection dates to observe the change of clusters over time. For example, Alm et al. build a phylogenetic tree and extracted the frequency trajectories of SARS-CoV-2 clades and lineages, based on the samples collection dates (Alm et al. 2020). The framework in my project built a network instead of a phylogenetic tree, so beside using samples' collection date as a tool to map sequences to an evolutionary timeline, I labeled nodes with collection dates and determined if collection dates were significantly correlated

within clusters. More precisely, I am interested in determining if there are clusters that contain significantly more nodes with more early collection dates than expected by chance. As not mentioned frequently in other studies, our permutation test analyses in previous result chapter showed that community detection clusters in SARS-CoV-2 transmission network is highly correlated with collection dates.

4.3 Parameters Affect

In the framework as I demonstrate in previous method chapter, three parameters are used among the three clustering methods. More precisely, there is a threshold parameter in TN93 pairwise distances, and expansion and inflation parameters are associated with MCL clustering method. Neither the connected component nor Louvain clustering method have additional tuning parameters.

Among all these parameters, the TN93 threshold is the only variable which was applied in all methods. When the threshold was set to a relatively high value, such as 0.05, sequences were more likely to form larger clusters. An extremely high threshold results in one enormous cluster. This kind of clustering result is uninformative, and it would result in poor performance when predicting cluster growth since there will be no variation in predictor variable for training the model and all new cases will belong to the same cluster. It should be noted that there is a bound range of thresholds where both community detection methods have a better performance than connected components as measured by delta-AIC. We expected this to occur because community detection methods can break large components into multiple communities. For a given distribution of edge among nodes, community detection methods have more variation in edge densities. Thus, community detection method would still be informative at high threshold. When the threshold is set to a relatively low value, such as 0.005, only a smaller number of cases, both known cases and new cases, are available to the model as training and testing data. As we analyzed in result chapter, a large proportion of cluster result is formed by paired cluster. In edge cases, extremely low threshold could result as having paired cluster or small size cluster only, and since fewer new cases are considered under such extremely low threshold, this kind of

cluster result can lead to poor prediction outcome of cluster growth. In contract, all three-clustering methods performance indifferently while taking relatively high thresholds. All of them provide informative clustering result since there are not enough sequence information are given from the graph to process with.

Expansion and inflation are two parameters in MCL clustering method, they result together in algebraic matrix of a transmission network. As I mentioned in result chapter, MCL doesn't converge all the time, there are combination of expansion and inflation values make MCL clustering method fail to give any cluster result. In the perspective of convergency, MCL did a worse job in Beijing data set than in Tennessee dataset. From the runs have analysis in result chapter that returns cluster result, we observe expansion affect AIC loss more than inflation. Furthermore, our result agrees with the Gibbons' work on inflation (Gibbons et al., 2015). As we have discussed in result chapter, tread on delta-AIC value is mostly decreasing while raising the inflation value. MCL inflation parameter can affect the granularity of the clusters, which a larger value of inflation leads to smaller clusters, and this could lead a less robust clustering result.

4.4 Location of Maximum AIC Loss

As I demonstrated in method section, AIC loss indicate the comparison between a null model and a proposed logistic regression model. On the top of that, one of the most important results of my framework is to avoid the selection of extreme thresholds and find an ideal threshold for individual dataset. The threshold value that corresponding with the maximum AIC loss is the optional threshold for each data set. As we have observed community detection clustering method have a higher ideal threshold than connected component clustering method. For example, there is a threshold shift from 0.015 to 0.03 between connected component clustering method and MCL clustering method in the Tennessee data set. Furthermore, a higher threshold is more adaptable to pattern clusters since population is highly correlated to threshold. By using higher threshold in the framework, we are able to include more cases, especially new cases to the clustering model. If more new cases join to one cluster than the expected rate, this can imply a detection on an outbreak. As the AIC loss measure the difference between two Poisson regression models that predict cluster growth, a more negative number indicate that the proposed model is preferred over the null model. More precisely, the proposed model will be more informative because it is based on recency, and the null model will be informative due to it only depends on cluster size.

Beside the threshold, other parameters associated with the location of maximum AIC loss could also improve the clusters' performance. Noticed that there are several parameters involved in the framework, the combination of parameters setting corresponding with the location of maximum AIC loss, would give us a best set of clusters for current clustering method.

4.5 Conclusions

1. Our framework is able to find the optimal threshold for induvial data set by locating the threshold that have the maximum AIC loss between a proposed and null model. An optimal threshold can avoid our framework's clustering result from getting uninformative clusters. For instance, extreme threshold value will not be selected.

2. For the clustering methods associate with multiple tuning parameters, our framework is able to find the most suitable parameters combination that can provide the most informative clusters. For the MCL clustering method, we find out the inflation parameter barley have influence on delta-AIC value and there usually is a deeper delta-AIC area when expansion is close to 2.

3. Hamming distance is a better way to compute genetic differences than TN93 pairwise distance in SARS-CoV-2 transmission network. And with the usage of Hamming distance, our framework is capable of handling massive and identical sequences in global SARS-CoV-2 sequence data.

4. Community detection method not only have a higher optimal threshold than connected component-based clustering method, but also extract more informative clusters for both HIV and SARS-CoV-2 data than connected component-based clustering method (i.e, have a more positive countries' correlation result in SARS-CoV-2 transmission network)

4.6 Future Directions

One possible direction is a further usage of the AIC plot. Beside the location of threshold, it would be interesting to find out how every AIC difference in the plot can be used to measure if one clustering method has an overall better performance than another. For instance, Akaike weights are usually used in model averaging and represents the relative likelihood of a model (Posada et al. 2004). For each model, we first calculate the relative likelihood of the model, which is exp(-0.5 * ΔAIC score for that model). By using Akaike weight for each model, we obtain the evidence ratio of w of model i / w of model j. This evidence ratio quantifies the strength of evidence in favour of model i over model j. However, the key problem is, comparing AIC values require that the models are being fit to the same data. In our case, it needs to be fit into the same cluster. Even if using the same sequence dataset, the composition of clusters will be different between methods. Therefore, while compare the result under different observations, this comparison cannot be done by simply compare their AIC difference.

Another possible direction is to apply overlapping methods into the framework. An overlapping method means nodes are allowed to be part of multiple communities. Both community detection method, MCL and Louvain clustering method, are non-overlapping approaches. Compared to non-overlapping clustering method, overlapping clustering method have substantially improved on the performance of the identification for disease-relevant clusters in gene-gene networks (Tripathi et al. 2019). The overlapping clustering method can be divided into two phases: (1) find the nodes that are most likely be the “seed node” among all nodes, (2) expand the seed node. In public health, there are only a few

present studies discussed the overlapping nodes. For instance, Villandre use community detection methods on HIV transmission network and measure the overlapping between transmission clusters (Villandre et al., 2016).

Bibliography

About HIV/AIDS | HIV Basics | HIV/AIDS | CDC. (n.d.). Retrieved July 27, 2022, from https://www.cdc.gov/hiv/basics/whatishiv.html

Aldous, J. L., Pond, S. K., Poon, A., Jain, S., Qin, H., Kahn, J. S., Kitahata, M., Rodriguez, B., Dennis, A. M., Boswell, S. L., Haubrich, R., & Smith, D. M. (2012). Characterizing HIV Transmission Networks Across the United States. Clinical Infectious Diseases, 55(8), 1135–1143. https://doi.org/10.1093/cid/cis612

Alm, E., Broberg, E. K., Connor, T., Hodcroft, E. B., Komissarov, A. B., Maurer-Stroh, S., Melidou, A., Neher, R. A., O'Toole, Á., Pereyaslov, D., & Group, T. W. E. R. sequencing laboratories and G. E. (2020). Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance, 25(32), 2001410. https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. Nature Medicine, 26(4), 450–452. https://doi.org/10.1038/s41591-020-0820-9

Ann M. Dennis, Erik Volz, A.S. Md. Simon D.W. Frost, Mukarram Hossain, Art F.Y. Poon, Peter F. Rebeiro, Sten H. Vermund, Timothy R. Sterling, and Marcia L. Kalish.HIV-1 Transmission Clustering and Phylodynamics Highlight the Important Role of Young Men Who Have Sex with Men.AIDS Research and Human Retroviruses.Oct 2018.879-888.http://doi.org/10.1089/aid.2018.0039

Arenas, A., Duch, J., Fernández, A., & Gómez, S. (2007). Size reduction of complex networks preserving modularity. New Journal of Physics, 9(6), 176–176. https://doi.org/10.1088/1367-2630/9/6/176

Balaban, M., Moshiri, N., Mai, U., Jia, X., & Mirarab, S. (2019). TreeCluster: Clustering biological sequences using phylogenetic trees. PLoS ONE, 14(8), e0221068. https://doi.org/10.1371/journal.pone.0221068

Bao, Y., Kapustin, Y., & Tatusova, T. (2008). Virus Classification by Pairwise Sequence Comparison (PASC). Encyclopedia of Virology, 342–348. https://doi.org/10.1016/B978-012374410-4.00710-X

Bedford, T., Greninger, A. L., Roychoudhury, P., Starita, L. M., Famulare, M., Huang, M.-L., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., Shrestha, L., Nguyen, T. N., Adler, A., Brandstetter, E., Cho, S., Giroux, D., Han, P. D., Fay, K., Frazar, C. D., … Jerome, K. R. (2020). Cryptic transmission of SARS-CoV-2 in Washington state. Science (New York, N.y.), 370(6516), 571–575. https://doi.org/10.1126/science.abc0523

Billock, R. M., Powers, K. A., Pasquale, D. K., Samoff, E., Mobley, V. L., Miller, W. C., Eron, J. J., & Dennis, A. M. (2019). Prediction of HIV Transmission Cluster Growth with Statewide Surveillance Data. Journal of Acquired Immune Deficiency Syndromes (1999), 80(2), 152–159. https://doi.org/10.1097/QAI.0000000000001905

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345–370. https://doi.org/10.1007/BF02294361

Brown, R. (1828). On the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. Edinburgh New Philosophical Journal, 5, 358-371.

Bukin, Yu. S., Bondaryuk, A. N., Kulakova, N. V., Balakhonov, S. V., Dzhioev, Y. P., & Zlobin, V. I. (2021). Phylogenetic reconstruction of the initial stages of the spread of the SARS-CoV-2 virus in the Eurasian and American continents by analyzing genomic data. Virus Research, 305, 198551. https://doi.org/10.1016/j.virusres.2021.198551

Cawley, G. C., & Talbot, N. L. C. (n.d.). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. 29.

Danon, L., Díaz-Guilera, A., & Arenas, A. (2006). The effect of size heterogeneity on community identification in complex networks. Journal of Statistical Mechanics: Theory and Experiment, 2006(11), P11010–P11010. https://doi.org/10.1088/1742-5468/2006/11/P11010

Dongen, S. (2000). Performance criteria for graph clustering and Markov cluster experiments. Undefined. https://www.semanticscholar.org/paper/Performance-criteria-for-graph-clustering-and-Dongen/3e0ee8bf437cdc69438470b8f4ddfe22e9745c89

E, W., Li, T., & Vanden-Eijnden, E. (2008). Optimal partition and effective dynamics of complex networks. Proceedings of the National Academy of Sciences, 105(23), 7907–7912. https://doi.org/10.1073/pnas.0707563105

Einstein, A. (n.d.). -INVESTIGATIONS O N THE THEORY .OF ,THE BROWNIAN MOVEMENT. 11.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research, 30(7), 1575–1584.

Fidan, H., & Erkan Yuksel, M. (2022). A comparative study for determining Covid-19 risk levels by unsupervised machine learning methods. Expert Systems with Applications, 190, 116243. https://doi.org/10.1016/j.eswa.2021.116243

Fisher, K. A., Tenforde, M. W., Feldstein, L. R., Lindsell, C. J., Shapiro, N. I., Files, D. C., Gibbs, K. W., Erickson, H. L., Prekker, M. E., Steingrub, J. S., Exline, M. C., Henning, D. J., Wilson, J. G., Brown, S. M., Peltan, I. D., Rice, T. W., Hager, D. N., Ginde, A. A., Talbot, H. K., … Marcet, P. L. (2020). Community and Close Contact Exposures Associated with COVID-19 Among Symptomatic Adults ≥18 Years in 11 Outpatient Health Care Facilities—United States, July 2020. Morbidity and Mortality Weekly Report, 69(36), 1258–1264. https://doi.org/10.15585/mmwr.mm6936a5

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. Proceedings of the National Academy of Sciences, 104(1), 36–41. https://doi.org/10.1073/pnas.0605965104

Gibbons, T. R., Mount, S. M., Cooper, E. D., & Delwiche, C. F. (2015). Evaluation of BLAST-based edge-weighting metrics used for homology inference with the Markov Clustering algorithm. BMC Bioinformatics, 16(1), 218. https://doi.org/10.1186/s12859-015-0625-x

Guimerà, Roger & Sales-Pardo, Marta & Amaral, Luís. (2004). Modularity from Fluctuations in Random Graphs and Complex Networks. Physical review. E, Statistical, nonlinear, and soft matter physics. 70. 025101. 10.1103/PhysRevE.70.025101.

HIV Cluster and Outbreak Detection and Response | Guidance | Program Resources | HIV | CDC. (2022, April 21). https://www.cdc.gov/hiv/programresources/guidance/cluster-outbreak/index.html

HIV Rates by Country 2022. (n.d.). Retrieved July 27, 2022, from https://worldpopulationreview.com/country-rankings/hiv-rates-by-country

HIV/AIDS. (n.d.). Retrieved July 27, 2022, from https://www.who.int/data/gho/data/themes/hiv-aids

Hong, K., Yum, S., Kim, J., Yoo, D., & Chun, B. C. (2021). Epidemiology and Regional Predictors of COVID-19 Clusters: A Bayesian Spatial Analysis Through a Nationwide Contact Tracing Data. Frontiers in Medicine, 8. https://www.frontiersin.org/articles/10.3389/fmed.2021.753428

Hozumi, Y., Wang, R., Yin, C., & Wei, G.-W. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. Computers in Biology and Medicine, 131, 104264. https://doi.org/10.1016/j.compbiomed.2021.104264

Human Immunodeficiency Virus (HIV). (2016). Transfusion Medicine and Hemotherapy, 43(3), 203–222. https://doi.org/10.1159/000445852

Kozlakidis, Z. (2022). Evidence for Recombination as an Evolutionary Mechanism in Coronaviruses: Is SARS-CoV-2 an Exception? Frontiers in Public Health, 10, 859900. https://doi.org/10.3389/fpubh.2022.859900

K Tamura, M Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees., Molecular Biology and Evolution, Volume 10, Issue 3, May 1993, Pages 512–526, https://doi.org/10.1093/oxfordjournals.molbev.a040023

Leicht, E. A., & Newman, M. E. J. (2008). Community Structure in Directed Networks. Physical Review Letters, 100(11), 118703. https://doi.org/10.1103/PhysRevLett.100.118703

Li, G., Piampongsant, S., Faria, N. R., Voet, A., Pineda-Peña, A.-C., Khouri, R., Lemey, P., Vandamme, A.-M., & Theys, K. (2015). An integrated map of HIV genome-wide variation from a population perspective. Retrovirology, 12(1), 18. https://doi.org/10.1186/s12977-015-0148-6

Liu, T., Gong, D., Xiao, J., Hu, J., He, G., Rong, Z., & Ma, W. (2020). Cluster infections play important roles in the rapid evolution of COVID-19 transmission: A systematic review. International Journal of Infectious Diseases, 99, 374–380. https://doi.org/10.1016/j.ijid.2020.07.073

Liu, W., & Yang, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. The Annals of Statistics, 39(4). https://doi.org/10.1214/11-AOS899

Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S., & Stadler, T. (2021). The origin and early spread of SARS-CoV-2 in Europe. Proceedings of the National Academy of Sciences of the United States of America, 118(9), e2012008118. https://doi.org/10.1073/pnas.2012008118

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nguyen, H. Q., & Do, M. N. (2015). Downsampling of Signals on Graphs Via Maximum Spanning Trees. IEEE Transactions on Signal Processing, 63(1), 182–191. https://doi.org/10.1109/TSP.2014.2369013

Nunes, D. R., Braconi, C. T., Ludwig-Begall, L. F., Arns, C. W., & Durães-Carvalho, R. (2022). Deep phylogenetic-based clustering analysis uncovers new and shared mutations in SARS-CoV-2 variants as a result of directional and convergent evolution. PLOS ONE, 17(5), e0268389. https://doi.org/10.1371/journal.pone.0268389

Outbreak identification – Outbreak Toolkit. (n.d.). Retrieved July 28, 2022, from https://outbreaktools.ca/background/outbreak-identification/

Park, L. S., Siraprapasiri, T., Peerapatanapokin, W., Manne, J., Niccolai, L., & Kunanusont, C. (2010). HIV Transmission Rates in Thailand: Evidence of HIV Prevention and Transmission Decline. JAIDS Journal of Acquired Immune Deficiency Syndromes, 54(4), 430–436. https://doi.org/10.1097/QAI.0b013e3181dc5dad

Perelman and Ostfeld, 2011 L. Perelman, A. Ostfeld Topological clustering for water distribution systems analysis Environ. Model. Software, 26 (7) (2011), pp. 969-972

Pons, P., & Latapy, M. (2005). Computing Communities in Large Networks Using Random Walks. In pInar Yolum, T. Güngör, F. Gürgen, & C. Özturan (Eds.), Computer and Information Sciences—ISCIS 2005 (pp. 284–293). Springer. https://doi.org/10.1007/11569596_31

Poon, A. F. Y. (2015). Phylodynamic Inference with Kernel ABC and Its Application to HIV Epidemiology. Molecular Biology and Evolution, 32(9), 2483–2495. https://doi.org/10.1093/molbev/msv123

Populations at Greatest Risk | High-Impact HIV Prevention | Policy and Law | HIV/AIDS | CDC. (2020, May 12). https://www.cdc.gov/hiv/policies/hip/risk.html

Porta, M. S., & International Epidemiological Association (Eds.). (2008). A dictionary of epidemiology (5th ed). Oxford University Press.

Prosperi, M. C. F., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Di Giambenedetto, S., Bruzzone, B., Capetti, A., Vivarelli, A., Rusconi, S., Re, M. C., Gismondo, M. R., Sighinolfi, L., Gray, R. R., Salemi, M., Zazzi, M., & De Luca, A. (2011). A novel methodology for large-scale phylogeny partition. Nature Communications, 2(1), 321. https://doi.org/10.1038/ncomms1325

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology, 5(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Rani, R. R., Ramyachitra, D., & Brindhadevi, A. (2019). Detection of dynamic protein complexes through Markov Clustering based on Elephant Herd Optimization Approach. Scientific Reports, 9(1), 11106. https://doi.org/10.1038/s41598-019-47468-y

Real, L. A., Henderson, J. C., Biek, R., Snaman, J., Jack, T. L., Childs, J. E., Stahl, E., Waller, L., Tinline, R., & Nadin-Davis, S. (2005). Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. Proceedings of the National Academy of Sciences, 102(34), 12107–12111. https://doi.org/10.1073/pnas.0500057102

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S., & Korber, B. (2000). HIV-1 Nomenclature Proposal. *Science, 288*(5463), 55. https://doi.org/10.1126/science.288.5463.55d

Romero-Severson, E. O., Hengartner, N., Meadors, G., & Ke, R. (2020). Change in global transmission rates of COVID-19 through May 6 2020. PLOS ONE, 15(8), e0236776. https://doi.org/10.1371/journal.pone.0236776

Sánchez, D.L., Revuelta, J., De la Prieta, F., Gil-González, A.B., Dang, C. (2016). Twitter User Clustering Based on Their Preferences and the Louvain Algorithm. In: , et al. Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection. PAAMS 2016. Advances in Intelligent Systems and Computing, vol 473. Springer, Cham. https://doi.org/10.1007/978-3-319-40159-1_29

Sharp, P. M., & Hahn, B. H. (2011). Origins of HIV and the AIDS Pandemic. Cold Spring Harbor Perspectives in Medicine:, 1(1), a006841. https://doi.org/10.1101/cshperspect.a006841

Shaw, G. M., & Hunter, E. (2012). HIV Transmission. Cold Spring Harbor Perspectives in Medicine, 2(11), a006965. https://doi.org/10.1101/cshperspect.a006965

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888–905. https://doi.org/10.1109/34.868688

Shih, Y.-K., & Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping Markov clustering. Bioinformatics, 28(18), i473–i479. https://doi.org/10.1093/bioinformatics/bts370

The Stages of HIV Infection | NIH. (n.d.). Retrieved July 27, 2022, from https://hivinfo.nih.gov/understanding-hiv/fact-sheets/stages-hiv-infection

Van Regenmortel MH. Virus species and virus identification: past and current controversies. Infect Genet Evol. 2007 Jan;7(1):133-44. doi: 10.1016/j.meegid.2006.04.002. Epub 2006 May 19. PMID: 16713373.

Villandre, L., Stephens, D. A., Labbe, A., Günthard, H. F., Kouyos, R., Stadler, T., & Study, T. S. H. C. (2016). Assessment of Overlap of Phylogenetic Transmission Clusters and Communities in Simple Sexual Contact Networks: Applications to HIV-1. PLOS ONE, 11(2), e0148459. https://doi.org/10.1371/journal.pone.0148459

Wertheim, J. O., Murrell, B., Mehta, S. R., Forgione, L. A., Kosakovsky Pond, S. L., Smith, D. M., & Torian, L. V. (2018). Growth of HIV-1 Molecular Transmission Clusters in New York City. The Journal of Infectious Diseases, 218(12), 1943–1953. https://doi.org/10.1093/infdis/jiy431

WHO Coronavirus (COVID-19) Dashboard. (n.d.). Retrieved July 27, 2022, from https://covid19.who.int

Worobey, M., Pekar, J., Larsen, B. B., Nelson, M. I., Hill, V., Joy, J. B., Rambaut, A., Suchard, M. A., Wertheim, J. O., & Lemey, P. (2020). The emergence of SARS-CoV-2 in Europe and North America. Science, 370(6516), 564–570. https://doi.org/10.1126/science.abc8169

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

Wu, Z., Wang, X., Fang, W., Liu, L., Tang, S., Zheng, H., & Zheng, Z. (2021). Community detection based on first passage probabilities. Physics Letters A, 390, 127099. https://doi.org/10.1016/j.physleta.2020.127099

Xu, J.-J., Han, M.-J., Jiang, Y.-J., Ding, H.-B., Li, X., Han, X.-X., Lv, F., Chen, Q.-F., Zhang, Z.-N., Cui, H.-L., Geng, W.-Q., Zhang, J., Wang, Q., Kang, J., Li, X.-L., Sun, H., Fu, Y.-J., An, M.-H., Hu, Q.-H., … Shang, H. (2021). Prevention and control of HIV/AIDS in China: Lessons from the past three decades. Chinese Medical Journal, 134(23), 2799–2809. https://doi.org/10.1097/CM9.0000000000001842

Yang, X., Dong, N., Chan, E. W.-C., & Chen, S. (n.d.). Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. Emerging Microbes & Infections, 9(1), 1287–1299. https://doi.org/10.1080/22221751.2020.1773745

Appendices

Appendix A: Accession numbers for Beijing HIV data set

The data we used in Beijing HIIV data is from Ye et al.'s work (Ye et al., 2020). The GenBank accession numbers we used are:

HQ007312-50, JF906562-700, KM011653-849, KY713346-582, AB746342-5, AB773884-5, EU921952-87, FJ036960-71, FJ374975-5126, FJ387028-128, FJ531405-62, FJ752417-20, FM251948-2030, GQ290693-724, GQ845124-6, GU345085-203, GU564221-30, HE590887-1065, HG421451-1735, HQ215552-87, HQ588180-303, JF932468-500, JN848837-955, JQ028198-423, JQ235008-21, JQ302545-755, JQ658474-772, JQ898221-77, JQ901022-97, JX070462-556, JX112796-870, JX392378-84, JX412323-63, JX960597-635, KC183774-83, KC203209-332, KC870027-44, KC888202-745, KC898975-9015, KC924448-4539, KC987968-78, KC988057-166, KC990124-7, KF250366-410, KF267584-704, KF714292-496, KF803577-80, KF835116-250, KF835493-547, KF857358-461, KJ184176-80, KJ193530-636, KJ401414-768, KJ484433-6, KJ570783-851, KJ613998-4226, KJ778895-7, KJ820090-408, KM217833-55, KM258676-875, KM370212-32, KM395730-811, KM974719-20, KP178420-50, KP234972-5200, KP250654-829, KP418582-633, KP698503-8, KP992343-441, KR187186-8450, KT378642-9957, KT625782-884, KT893482-704, KU050197-674, KU161143-5, KU364385-414, KU378038-46, KU871408-88, KU992928-37, KX198562-86, KX305973-6175, KX378999-9000, KX791498-637, MF503154-241, MF684019-335, MG787428-59, MG905777-818.

Curriculum Vitae

Name: Mo Liu

Education:

Dates	Program	Institution and Field
2020-09 – present	M.Sc.	Western University, London ON. Field/Discipline: Pathology Thesis title: Outbreak detection from virus genetic sequence variation by community detection. Thesis advisor: Art Poon, PhD
2016-09 – 2020-05	B.Sc.	Queen's University, Kingston ON. Field/Discipline: Computing and Mathematics (Honors) Thesis title: Construction of environmental risk score using CHILD cohort study. Thesis advisor: Qingling Duan, PhD Award: Dean's Honor List-Distinction

Publications:

Ferreira, Roux-Cil, Emmanuel Wong, Gopi Gugan, Kaitlyn Wade, Molly Liu, Laura Muñoz Baena, Connor Chato, Bonnie Lu, Abayomi S. Olabode, and Art F. Y. Poon. "CoVizu: Rapid Analysis and Visualization of the Global Diversity of SARS-CoV-2 Genomes," July 21, 2021. https://doi.org/10.1101/2021.07.20.453079