Journal of Bioinformatics and Computational Biology
© Imperial College Press

UBIQUITOUS REASSORTMENTS IN INFLUENZA A VIRUSES

XIU-FENG WAN[*ξ]

Systems Biology Laboratory, Department of Microbiology, Miami University; Department of Computer Science and Systems Analysis, Miami University, Oxford, Ohio 45056 USA
wanhenry@yahoo.com

MUFIT OZDEN

Department of Computer Science and Systems Analysis, Miami University, Oxford, Ohio 45056 USA
ozdenm@muohio.edu

GUOHUI LIN

Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada
ghlin@cs.ualberta.ca

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Influenza A virus is a negative stranded RNA virus, composed of eight segmented RNA molecules, including polymerases (PB2, PB1, PA), haemaglutin (HA), nucleoprotein (NP), neuraminidase (NA), matrix protein (MP), and nonstructure gene (NS). The influenza A viruses are notorious for rapid mutations, frequent genomic reassortments, and possible recombinations. Among these evolutionary events, genetic reassortments refer to exchanges of internal fragments (PB2, PB1, PA, NP, MP, NS) between co-infected viruses, and they have been responsible for generating pandemic and epidemic strains. Thus, identification of reassortments will be critical for pandemic and epidemic prevention and control. This paper presents a reassortment identification method based on distance measurement using complete composition vector (CCV) and segment clustering using a minimum spanning tree algorithm. By applying this method, we identified 34 potential reassortment clusters among 2,641 PB2 segments of influenza A viruses. Among the 83 serotypes tested, at least 56 serotypes (67.46%) exchanged their fragments with another serotype of avian influenza viruses. These identified reassortments involve 1957 H2N1 and 1968 H3N2 influenza pandemic strains as well as H5N1 avian influenza virus isolates, which have generated a threat for a future pandemic. More frequent reassortments were found to occur in wild birds, especially migration birds. This MST clustering program is written in Java and will be available upon request.

Keywords: Reassortment; Avian influenza virus; Influenza A virus; CCV; Minimum spanning tree.

[*] Present address: Molecular Virology and Vaccine Branch, Influenza Division, The Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, Georgia 30333. E-mail: fvq7@cdc.gov.

[ξ] Correspondence should be addressed to Xiu-Feng Wan by e-mail wanhenry@yahoo.com or fvq7@cdc.gov.

1. Introduction

Influenza virus is a negative-stranded RNA virus that belongs to the *Orthomyxoviridae* family [1]. There are three serotypes, A, B, and C, of which B and C are reported to infect mammals only. Influenza A and B viruses have 8 genomic segments (segment 1-8) with varying lengths from about 890 to 2,341 nucleotides which encode at least 11 proteins: PB2 by segment 1, PB1 and PB1-F by 2, PA by 3, haemagglutinin (HA) by 4, nucleoprotein (NP) by 5, neuraminidase (NA) by 6, membrane protein M1 and M2 by 7, and nonstructural protein NS1 and NS2 by 8. Genetic reassortment refers to the exchange of one or more discrete internal RNA segments (PB2, PB1, PA, NP, M, or NS) into multipartite viruses. The influenza virus may cause a pandemic disaster that will impact multiple continents. Three influenza pandemics occurred in 1918, 1957, and 1968 [2, 3]. More than 40 million people were killed in the 1918 influenza pandemic, which was caused by the H1N1 influenza A virus. Since our first isolation of two H5N1 AIVs from Guangdong farmed geese in 1996 [4], these viruses have caused outbreaks in both domestic and wild birds in Asia, Africa and Europe. Moreover, at least 327 confirmed cases and 199 deaths have been reported (www.who.int). The recent family cases in Indonesia alert us a potential threat for human to human transmission [5]. Thus, the World Health Organization (WHO) has issued a warning for a potential influenza pandemic in the near future.

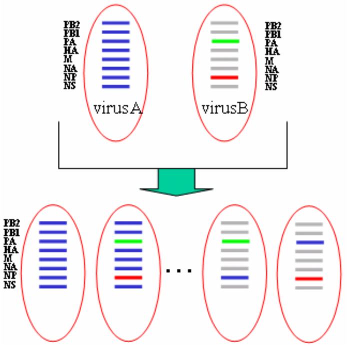


Fig. 1. Genetic reassortment model.

Genetic reassortment is referred to as antigenic shift. Genetic reassortment occurs frequently for all types of influenza viruses [6-10]. Reassortments between two co-infecting influenza A viruses possibly generate 2^8 genotypes in their offsprings (Fig. 1). Genetic reassortment, mutation, and possible recombination lead to a rapid emergence of novel flu genotypes [11-14]. Genetic reassortment facilitates generations of these pandemic influenza strains. It is generally accepted that both the 1957 and 1968 influenza pandemic strains were reassortants from the circulating H1N1 strains before 1957 [15]. Many reassortments have been reported [11, 14] for the recent H5N1 AIVs as well. Identification of reassortment events among AIVs will help us understand the underlying evolutionary pathways thus facilitate a preparation of strategy for influenza prevention and control. Furthermore, studies on the reassortment will help us understand the “teaming” harmony between segments from different AIVs, and thus predict an emergence of a pandemic or epidemic strain by considering our knowledge of the segment pool.

The traditionally computational reassortment identification generally involves three steps: (1) the sequences are selected to perform a multiple sequence alignment using software such as Clustal W [16], T-coffee [17], and Muscle [18]; (2) the phylogenetic trees are constructed using Neighborhood Joining (NJ), Maximum Parsimony (MP), Maximum Likelihood (ML), or Bayesian approach using software such as PAUP* [19], PHYLIP [20], or MrBayes [21]; (3) the conflicting tree topologies between gene segments will be identified as a potential reassortment. For instance, if an internal gene in H5N1 AIVs is located with a clade of H9N2 AIVs, a potential reassortment will be proposed.

However, reassortment identification is a non-trivial task. Multiple sequence alignments and phylogenetic tree construction are normally very time-consuming, and their performance is negatively correlated with the number of sequences. With the bench mark testing, the accuracy of multiple sequence alignments is even less than 70% [22]. Although many influenza gene segments are relatively conserved, it is very common to obtain an incorrect sequence alignment. Based on these available methods, it is impossible to identify the reassortments by considering all the sequences in the databases. Thus, many times, people have to arbitrarily take a subset of data from the database for their analyses thus potentially miss many reassortment events.

Previously we designed a computational method for reassortment identification based on module identification [23]. However, it is still limited by the data size. In this paper, we present another new reassortment identification based on the evolutionary distance measurement using complete composition vector (CCV) and a minimum spanning tree (MST) clustering algorithm. We applied this method in identifying the reassortments of PB2 gene in Influenza A viruses. In this paper, the discussion will be focused on the reassortments between different serotypes.

2. Materials and Methods

2.1 Reassortment identification using minimum spanning tree

2.1.1 Pairwise evolutionary distance calculation

In this study, we applied CCV to calculate pairwise distance between gene segments of the viruses. This method has been described in our previous works [23-25]. The performance of CCV is not limited by the number of sequences, which adversely correlates with the accuracy of most other available evolutionary distance measurement approaches. Briefly, we use S to denote one of the eight gene segments, $S\in$ {HA, NA, NP, PB2, PB1, PA, NS, M}. By using window size k ($k\geq2$), we scan S to generate $L - k + 1$ strings with the length k; each string can be represented as $\alpha_1\alpha_2\ldots\alpha_k$, $\alpha\in$ {A, U, G, C} for an influenza RNA segment. We denote the number of occurrences in S as $f(\alpha_1\alpha_2\ldots\alpha_k)$. Thus, the frequency of string $\alpha_1\alpha_2\ldots\alpha_k$ is $p(\alpha_1\alpha_2\ldots\alpha_k) = f(\alpha_1\alpha_2\ldots\alpha_k)/(L - k + 1)$. Similarly, we can calculate the frequencies of string $\alpha_1\alpha_2\ldots\alpha_{k-1}$ and $\alpha_2\alpha_3\ldots\alpha_k$. Thus, we can calculate the expected frequency for string $\alpha_1\alpha_2\ldots\alpha_k$ through a Markov model:

$$p^e(\alpha_1\alpha_2...\alpha_k) = \begin{cases} \dfrac{p(\alpha_1\alpha_2...\alpha_{k-1}) \times p(\alpha_2\alpha_3...\alpha_{k-2})}{p(\alpha_2\alpha_3...\alpha_{k-2})}, \text{if } p(\alpha_2\alpha_3...\alpha_{k-1}) \neq 0 \\ 0 \end{cases} \tag{1}$$

When k = 1, we set p^e (α_1) = p (α_1). We can then calculate the difference $d(\alpha_1\alpha_2\ldots\alpha_k)$ between the real frequency and the expected frequency as the evolutionary information carried by string $\alpha_1\alpha_2\ldots\alpha_k$:

$$d(\alpha_1\alpha_2...\alpha_k) = \begin{cases} \dfrac{p(\alpha_1\alpha_2...\alpha_k) - p^e(\alpha_1\alpha_2...\alpha_k)}{p^e(\alpha_1\alpha_2...\alpha_k)}, \text{if } p^e(\alpha_1\alpha_2...\alpha_k) \neq 0 \\ 0 \end{cases} \tag{2}$$

We denote the composition vector containing all the $d(\alpha_1\alpha_2\ldots\alpha_k)$ values as $V^k(S)$, in which k is the length of the strings. For the influenza RNA sequence, the number of entries in V^k will be 4^k. To maintain as much information as possible, we include all strings from length m to n, that is $m \leq k \leq n$. Thus, the CCV is defined as $V(S) = \{V^m(S), V^{m+1}(S), \ldots V^k(S), \ldots, V^{n-1}(S), V^n(S)\}$. We determine the optimal values for m and n by checking the amount of information contained. In this study, we set m = 1 and n = 40. The distance between two segments is computed using the *Euclidean* distance between their CCVs [24]:

$$D(S,S') = \sqrt{\sum_{i-1}^{N}(d_i - d_i^{'})^2}\ , \tag{3}$$

where d and d' are the evolutionary information for string $\alpha_1\alpha_2\ldots\alpha_k$ in segments S and S' in two AIVs, and $N = 4^n$ for nucleotides, n is the maximum length of the string. For instance, S and S' can be HA segments in two AIVs, respectively.

2.1.2 Minimum spanning tree clustering algorithm

There are a number of clustering algorithms appropriate for a working cluster definition, for example hierarchical schemes, *k*-means method, or the methods that try to cover all the points with the minimum total distance using the analytical algorithms, the traveling salesman algorithms or minimal spanning tree (MST) algorithms [27, 28] or some meta-heuristics, such as genetic algorithms. Among these clustering algorithms, MST algorithms have a unique advantage over most of the other clustering algorithms in which we do not need to define the number of clusters before clustering process. The MST algorithms have been applied extensively in high dimensional datasets, such as gene expression data analysis [29]. In addition to being a very efficient method for clustering, MST algorithms can handle any shapes of clusters even with the background noise created by the outlier points. Since we have little knowledge about the gene segment pools in the influenza viruses, MST will be potentially good fit for our application.

We define that if a virus belongs to a cluster, it is closest to some members of the cluster and the cluster distances (the edge weights in the constructed MST) increase gradually toward the boundaries of the clusters. It is rather subjective where the boundary of a cluster ends. Here, we assume that the boundary of a cluster is reached when an estimated distance percentile (threshold value) of the data points is exceeded. There may be some outlier viruses that do not belong to any identifiable clusters. The threshold value t is estimated on the basis of the mean and standard deviation of the minimum virus distances as follows:

$$t = \mu + m\sigma$$

where μ is the mean; σ is the standard deviation; and m is a multiplier.

As in Xu *et al.* [30], we choose Prim's MST algorithm since it provides the optimal solution to the problem with a linear representation of the MST. Given a connected, undirected graph $G = <V, E>$, the MST problem is to find a tree $T=<V, E^{/}>$ with no cycle such that $E^{/}$ is a subset of E and the total distance of T is minimal. An MST is not necessarily unique. A tree over $|V|$ vertices contains $|V|$-1 edges. Except for the root of the tree, every vertex in the tree constructed by Prim's algorithm will have exactly one parent vertex and may have many children. The leaf nodes have no children.

Prim's algorithm starts with an arbitrary vertex forming the initial tree, $<p_1, \_>$, and expands it until it covers all the vertices in the graph. At each step, a vertex not yet in the tree but closest to (some vertex that is in) the tree is determined and added to the tree. The insertion order of the vertices is recorded in $L(V)$, called "Linear representation of clusters". The Fibonacci heap implementation of Prim's algorithm runs in $O(|V|^2)$ time.

Even though we know some subtrees of an MST form the optimal clusters of the points, there is no unique method that can identify what these subtrees are. When the number of cluster is specified at the outset, a corresponding number of the longest edges in the MST can be cut to form the desired number of clusters. In our case, we do not know the true number of clusters. Therefore, we use the insertion order $L(V)$ built during the Prim's construction of an MST as in [30]. That is, once the MST is constructed, we estimate the average and the standard deviation of virus distances in the MST to define the first-level threshold value that will be used to determine the memberships of the first-level clusters. Starting from the data points at the beginning of the list $L(V)$, we compare the virus distances from their parents to the threshold value to determine the left boundary point of a cluster. When the distance of a point is detected to be less than the threshold, we begin a new cluster and continue comparing and inserting the viruses into the cluster until a virus distance becomes larger than the threshold value again. The same process is repeated to the end of the list $L(V)$ identifying all the clusters. We found it useful to identify the second-level embedded clusters that may reveal a closer evolutionary relationship among the viruses within the large first-level clusters by repeating the same process with the mean and deviations estimated now from the cluster distances.

2.1.3. Potential reassortment assignment

The reassortment denotes the exchange of genomic segments between co-infected influenza viruses. It will be very challenging to identify the reassortments when two closely related viruses co-infected the same cells altogether since we do not know whether the mutations in the segments resulted from concurrent mutations in the viruses, inherited from same ancestral viruses, or exchanged between different viruses (so called reassortments). Generally, the segments from different serotypes of influenza viruses form distinct genotypes (clades) during the evolution. In this paper, we simplified our analyses and focused on the reassortments between different serotypes. Thus, if we find the two viruses from different serotypes are located in the same cluster, we can believe the reassortments occurred between these two viruses. These reassortments may occur in the same generation of the viruses found in the surveillance or previous generations. Based on this feature, we are able to assign potential reassortment events between influenza A viruses.

2.2. Phylogenetic tree construction

The phylogenetic trees were constructed by Maximum Likelihood using PAUP* 4.0 [19], Maximum Parsimony using PAUP* 4.0 [19], and Neighbor-Joining method using Phylip 3.65 [20] as described previously [12, 23, 25]. The bootstrap values were based on the maximum parsimony bootstrap values with 1000 replicates. The nucleotide substitution model is selected by Modeltest 3.7 [31].

2.3. Cluster visualization

To visualize the clusters generated by Prim's MST algorithm, we use the shortest pairwise distance between two clusters to represent the difference between two clusters. The resulting distance matrix will be used to construct a tree using Neighbor-Joining method using Phylip 3.65 [20]. This type of tree will show the overall relationship between clusters.

2.4. Datasets

The influenza data sets were downloaded from Influenza Virus Resource database at GenBank (http://www.ncbi.nlm.nih.gov/ genomes/FLU/FLU.html), which were updated in November of 2006. In this study, we analyzed the 2,641 PB2 genes. The PB2 gene lengths vary from 2,100 to 2,341 nucleotides.

3. Results

3.1. Phylogenetic tree confirms the MST clustering results

We demonstrate our clustering approach by using a relatively small data set with the HA genes of 594 H5 AIVs. Through clustering, the MST had the minimum distance of 0 and the maximum distance of 44.99. The mean, standard deviation, and threshold value estimated from the MST distances turned out to be 15.03, 9.6, and 34.39 with m = 2, respectively. The procedure yielded 15 nonsingular clusters. The MST distance frequency distribution is shown in Fig. 1A. The *L(V)* list distances are shown in Fig. 1B. In Fig. 1B, the y-axis represents the distance of the successive viruses from their immediate parents in the list *L(V)* and the x-axis records the index number of those viruses in *L(V)*. The largest cluster (cluster 14 in Fig. 3) among the 15 first-level clusters had 478 viruses for which the mean and the standard deviation of the distances were 13 and 7.3.

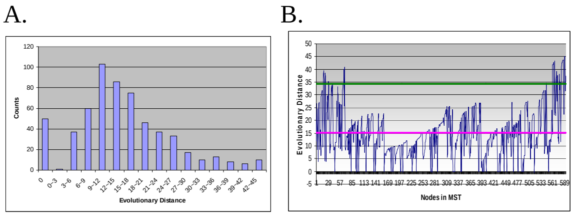


Fig. 2. Data distribution and MST for 594 HA segments from H5N1 AIVs. (A). Distribution of pair-wise distance between HA segments in 594 H5N1 AIVs; (B) MST constructed using Prim's algorithm, red line is the distance average and green line is the threshold.

To validate the concepts of the clustering results with traditional genotyping concepts, we constructed the phylogenetic tree using NJ method of Phylip based on the distance matrix from CCV. As shown in Fig. 3, our clustering results showed the recent emerging lineages of H5N1 AIVs are distinct from other earlier or circulating lineages. The Guangxi H5N1 AIVs have evolved to form a niche and be a distinct cluster from other H5N1 AIVs. The inner clustering will further separate this large cluster into more than 10 small clusters, which reflect those previously reported genotypes (data not shown). These results demonstrated our MST clustering algorithm is able to efficiently connect the HA segments with closer evolutionary distance. Thus, with a proper threshold, this MST graph will be able to separate different clusters.

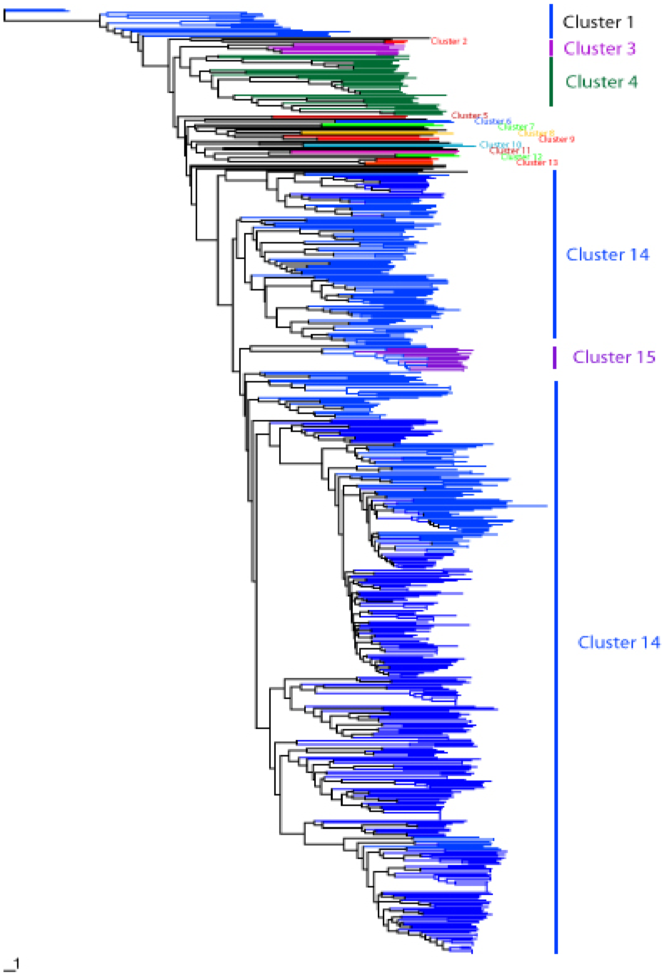


Fig. 3. The phylogenetic tree of 594 H5 AIVs. The resulting 15 clusters are marked in different colors. The phylogenetic tree were constructed based on CCV distance and Neighbor-Joining method using Phylip 3.65 based on distance from CCV [20].

3.2. PB2 reassortment identification in influenza A viruses

We applied MST clustering method to identify potential reassortments in PB2 segments of influenza A viruses. The average of distance is 14.76, and the standard deviation is 12.66. Here we utilized the threshold of 40.08 (m = 2), which is similar to the niche threshold (38.9) of PB2 segments in H5N1 AIVs, where we applied Bayesian approaches and found H5N1 avian influenza viruses form niches through natural selection [25]. Thus, we believe this threshold would be a proper threshold for us to define the cluster boundary of influenza A viruses using Prim's MST algorithm. Through the clustering, we identified 67 clusters with size from 2 to 1,426, 34 of which have potential reassortments. Among the 83 serotypes tested, at least 56 serotypes (67.46%) exchanged their fragments with another serotype of avian influenza viruses. This is the first systematic analyses of PB2 gene segments in influenza A viruses. Fig. 4 shows the relationships between these clusters, which are discussed as follows.

3.2.1. Frequent reassortments of PB2 in human influenza

As shown in Fig. 4, cluster 3 includes 1,426 PB2 gene segments involving 1957 H2N2 pandemic strains, 1968 H3N2 pandemic strains, as well as many historical and recent H3N2 seasonal outbreaks, including those in North America. The PB2 segments in this cluster form 14 inner clusters, each of which reflects the "jumping clusters" for each outbreak. For instance, inner clusters 3-13 reflect the 2003 H3N2 seasonal outbreak in New York. However, different clusters of PB2 segments may co-exist in the same outbreaks. From 2005, the New York isolates contain PB2 segments in different clusters. Generally, the viruses will remain more likely to be in a single cluster if the influenza cases remain active in that period since these "powerful" viruses are more likely to dominate the isolates. The segments in this cluster were possibly originated from earlier H1N1 influenza A virus since many H1N1 influenza viruses from 1940s and 1950s were also located in this cluster. It was reported that PB2 segments of both 1957 segments and 1968 pandemic strains were from 1918 H1N1 pandemic strain [32]. In our results, the PB2 of 1918 influenza H1N1 virus was defined as an outlier, probably because this segment has been evolved rapidly to form different niches through over 20 years from 1918 to 1940s. The PB2 segments of cluster 3 include also the H1N2 influenza A viruses identified in 2003 New York H1N2 epidemics.

The recently emerged H5N1 AIVs have caused at least 327 cases and 199 deaths in human. In the clustering results, these viruses are located in eight clusters (clusters 29, 32, 37, 38, 39, 43, 50, and 52). These demonstrated the dynamic evolution of these viruses during their adaptation. From current available sequences, we can see the frequent reassortments between H5N1 AIVs and the available segment pool, such as H9N2 and H6N1 AIVs (in clusters 52, 38, 32). These results were consistent to previous reports [11, 12, 14, 23, 33-35]. Cluster 52 includes the H5N1 AIVs caused human cases in 1997 Hong Kong outbreak. Cluster 38 has not caused any human infection yet, and the reassortants were transmitted among domestic and wild birds. Cluster 32 is the major circulating PB2 segments in H5N1 AIVs, and this cluster has caused the majority of human cases since 2003, including the family cases in Indonesia. This cluster contains at least 13 serotypes, including H5N1, H5N3, H6N1, H3N6, H3N2, H10N9, H11N2, H4N2, H3N2, H3N8, H7N1, H9N2, and H4N4. Some of these viruses have been circulating in the wild ducks in Southeast Asia since 1970s. Cluster 32 has 294 viruses, and they can be further clustered into three inner clusters. Inner cluster 32-1 is the major cluster, including 267 viruses and all of the reassortants. Inner cluster 32-2 contains eight PB2 segments of H6N1 AIVs circulating in Taiwan from 1997 to 2000. Inner cluster 32-3 includes ten PB2 segments of H5N1 AIVs circulating in the environments and waterfowls in China from 1996 and 2000, which including the strain of A/Goose/Guangdong/1/1996 (H5N1).

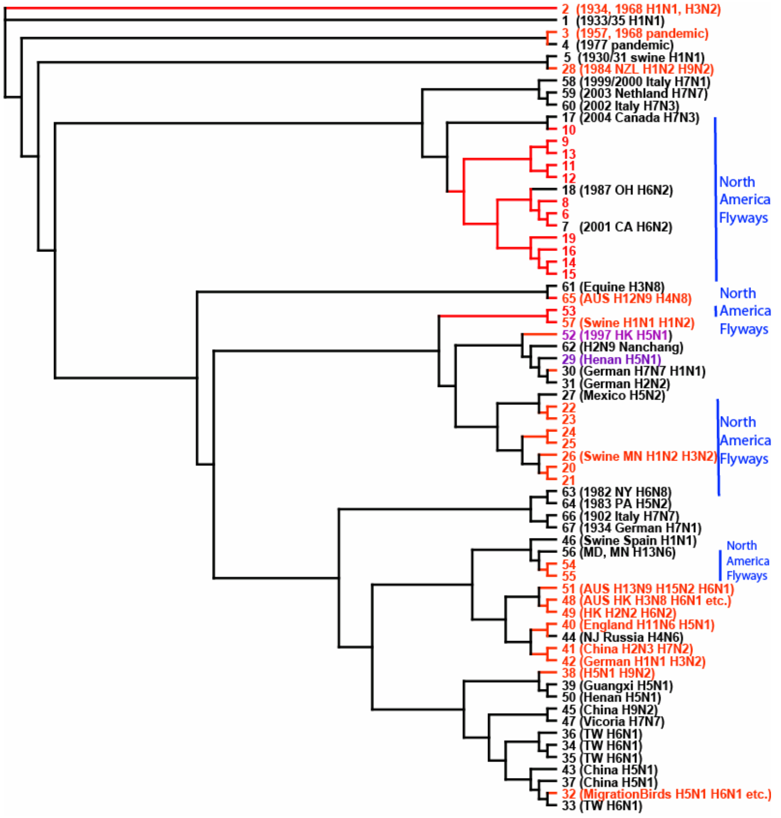


Fig. 4. Visualization of the clusters. The clusters in red is with multiple HA and NA serotypes.

3.2.2. Ubiquitous PB2 reassortments among the flyways of migration birds

To date, there are 16 subtypes of HA and 9 subtypes of NA in influenza A viruses [35]. Thus, there is a possibility of 144 serotypes of influenza A viruses. In the migration birds of North America, there are at least 65 serotypes. In our MST results, these viruses were grouped into 20 clusters. Our results showed 18 of these clusters have multiple serotypes, thus potential reassortments. Among these clusters, the birds at Alberta are found to be involved in most of these clusters.

Alberta is located in the western Canada and connects Montana in its south end. Alberta is one major breeding area for migration birds in North America. It is also a transient area for many migration birds. The avian influenza viruses in our clusters follow mainly two principle flyways in North America, including Atlantic flyways and Pacific flyways (Fig. 5). It is noted that we only consider the reassortments among different serotypes in this paper. Thus, the real reassortment cases will be even more frequent than those shown in our datasets.

Cluster 6 and 64 are two largest clusters among these two clusters. The AIVs in Cluster 64 belong to 33 serotypes. These viruses are found in Alberta, Minnesota, Delware, New York, New Jersey, and Tennessee, which are located along two major Atlantic flyways although Tennessee is located at the boundary of Mississippi pathway and Atlantic pathway (Fig. 5). Cluster 6 has 73 viruses belonging to 36 serotypes. These viruses are located in Alberta, Minnesota, Massachussetts, and Delware, which are also located along with the Atlantic flyway (Fig. 5).

The frequent reassortments in migration birds have facilitated emergence of new influenza strains. These emerged strains may not only cause outbreaks in domestic poultries and also donate gene segments to human influenza viruses. It was reported that all of documented pandemic strains were originated from avian species [32].

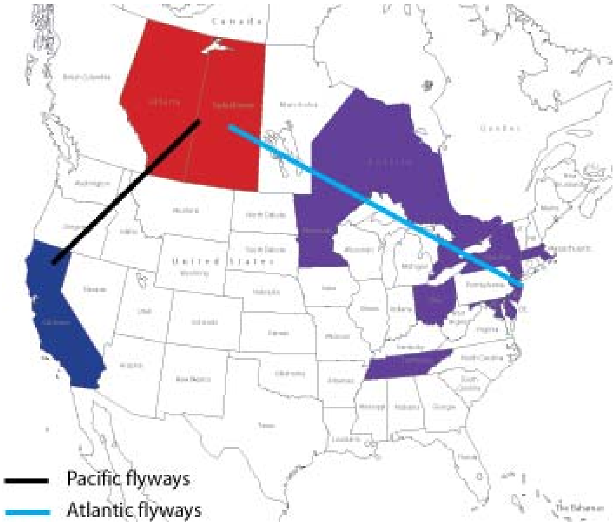


Fig. 5. The locations and simplified Pacific and Atlantic flyways, where potential reassortments were identified to be linked with migration birds at Alberta.

3.2.3. Surrounding of "ancient" PB2 segment in nature

The MST clustering results showed that the PB2 segments from recent isolates are clustered with those from AIVs isolated more than 20 and even 40 years earlier. For instance, the recent H5N1 AIVs isolates are clustered with earlier isolates. The isolates in the swine of Ontario, A/swine/Ontario/48235/04(H1N2), were located within the swine isolates from Tennessee in 1977, A/swine/Tennessee/24/1977(H1N1). A validation through phylogenetic tree was shown in Fig. 6. The strain of A/chicken/Hebei/1/2002(H7N2) is clustered with A/duck/Germany/1215/1973(H2N3). Strikingly, recent H5N1 AIVs (Cluster 32) have picked up ancient PB2 segments of AIVs in waterfowls circulating in Southeastern Asia, for instances, A/duck/HongKong/7/1975(H3N2), A/duck/HongKong/562/1979(H10N9), A/duck/HongKong/24/1976(H4N2), A/goose/HK/10/1976(H3N2), and A/duck/HK/784/1979(H9N2). Another study found similar results in PB2 gene, NP of Ck/Hebei/718/01 has a potential reassortment with Africanstarling/England-Q/983/1979(H7N1) (Wan, X.F., unpublished).

These phenomena are present for those non-reassortment strains as well. The PB2 from A/ruddyturnstone/NewJersey/47/1985(H4N6) is very close to A/duck/Czechoslovakia/1956(H4N6) but far away from other segments.

It is still unknown how these reassortments work, especially how these "ancient" segments remain with low evolutional rates in such a long time, which is much in contrast to high evolutional rates in the circulating AIVs. However, our results show that most of these "ancient" segments were originally present in waterfowls. On the other hand, it is unknown whether these types of segments are continuously circulating in hosts or dormant in environments for a while.

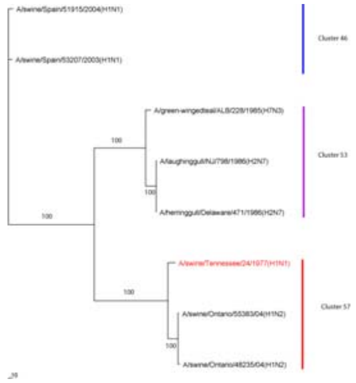

A/swine/Spain/51915/2004(H1N1)
A/swine/Spain/53207/2003(H1N1)
Cluster 46
A/green-wingedteal/ALB/228/1985(H7N3)
100
A/laughinggull/NJ/798/1986(H2N7)
100
A/herringgull/Delaware/471/1986(H2N7)
Cluster 53
100
A/swine/Tennessee/24/1977(H1N1)
100
A/swine/Ontario/55383/04(H1N2)
100
A/swine/Ontario/48235/04(H1N2)
Cluster 57
.10

Fig. 6. Circulation of "ancient" segments in recent avian influenza isolates.

4. Discussion

This paper presents a reassortment identification based on CCV and MST clustering algorithm. This study discovers reassortments in such a large data set automatically for the first time. We applied this method in reassortment identification in the PB2 gene segments of influenza A viruses. Our results showed the reassortment occurred very frequently in both influenza viruses in human and animal hosts, especially waterfowls. Since our annotations only focused on the reassortment events among viruses with different serotypes in each cluster, the actual reassortment cases are even much more frequent than our annotations.

The major challenge of the MST clustering method is to assign the cluster boundaries. Due to the limitation of current surveillance sequence data, it is difficult to obtain a reliable clustering result based only on the density of data in the network. Nevertheless, it might be feasible for certain influenza viruses such as H5N1, H9N2, and H3N2 influenza A viruses since they have much more data than other viruses. By means of certain estimated thresholds, for instance, $t = \mu + m\sigma$ or niche threshold [25], it is possible to overcome this type of limitations. Especially, the niche threshold value reflects the threshold for virus to form "jumping clusters" [25]. In this study, MST clusters are formed based on the *t* value, which is similar to the niche boundary we identified [25].

Our application of the MST based clustering on PB2 gene segments showed the effectiveness of such a method in influenza reassortment analysis. Future studies will focus on deeper analyses of the reassortments in the viruses with the same serotypes by linking the information from the clusters of other gene segments.

Migration birds have been shown to be an important factor in influenza viral epidemiology. Recently, they have spread H5N1 AIVs from Southeast Asia to all over Asia, Europe, and Africa, although it has been shown the poultry movements played a critical role in viral spreading in Southeast Asia [37, 38]. The close interactions among the birds with the same species, different species, local species, and migration species in breading areas will enhance the selection of a powerful influenza entity since the large influenza viral pool in these birds. The bird migration activities overcome the geographic barrier between viral pools at different locations. On the other hand, these selections may present an epidemic/pandemic strain for human outbreaks. It is worth noting that some of current human cases since 2004 were closely related to the genotypes presently circulating in migration birds [25]. One major concern is that the AIVs may reassort with mammal influenza viruses, which may acquire "powerful" resources from the human influenza viral pool. Thus, reassortments may facilitate the emergence of an H5N1 AIV strain with the ability of human to human transmission. Frequent human cases of H5N1 AIVs have increased this type of possibility. Thus, prevention and control of seasonal influenza will be important for preventing an emergence of a pandemic influenza strain. In addition, more surveillance will be required to understand the influenza segment pool in nature.

5. Acknowledgments

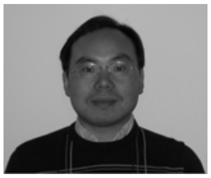
The authors thank computing support from the Ohio Supercomputer Center and the research computing support group of Miami University. This work was supported by Miami University CFR grant.

6. References

1. S.J. Flint, L.W. Enquist, V.R. Racaniello & A.M. Skalka. , *Principles of virology: molecular biology, pathogenesis, and control of animal viruses*. 2nd ed. 2004: ASM press.
2. C.R. Parrish and Y. Kawaoka, *The origins of new pandemic viruses: the acquisition of new host ranges by canine parvovirus and influenza A viruses.* Annu Rev Microbiol, 2005. **59**: p. 553-86.
3. P. Palese, *Influenza: old and new threats.* Nat Med, 2004. **10**(12 Suppl): p. S82-7.
4. X.-F. Wan, *Isolation and characterization of avian influenza viruses in China.* , in *College of Veterinary Medicine*. 1998, South China Agricultural University: Guangzhou.
5. E.R. Sedyaningsih, S. Isfandari, V. Setiawaty, L. Rifati, S. Harun, W. Purba, S. Imari, S. Giriputra, P.J. Blair, S.D. Putnam, T.M. Uyeki, and T. Soendoro, *Epidemiology of cases of H5N1 virus infection in Indonesia, July 2005-June 2006.* J Infect Dis, 2007. **196**(4): p. 522-7.
6. J.A. McCullers, T. Saito, and A.R. Iverson, *Multiple genotypes of influenza B virus circulated between 1979 and 2003.* J Virol, 2004. **78**(23): p. 12817-28.
7. C. Li, K. Yu, G. Tian, D. Yu, L. Liu, B. Jing, J. Ping, and H. Chen, *Evolution of H9N2 influenza viruses from domestic poultry in Mainland China.* Virology, 2005. **340**(1): p. 70-83.
8. I.H. Brown, P.A. Harris, J.W. McCauley, and D.J. Alexander, *Multiple genetic reassortment of avian and human influenza A viruses in European pigs, resulting in the emergence of an H1N2 virus of novel genotype.* J Gen Virol, 1998. **79 (Pt 12)**: p. 2947-55.

9. Y. Matsuzaki, K. Mizuta, K. Sugawara, E. Tsuchiya, Y. Muraki, S. Hongo, H. Suzuki, and H. Nishimura, *Frequent reassortment among influenza C viruses.* J Virol, 2003. **77**(2): p. 871-81.
10. L. Widjaja, S.L. Krauss, R.J. Webby, T. Xie, and R.G. Webster, *Matrix gene of influenza a viruses isolated from wild aquatic birds: ecology and emergence of influenza a viruses.* J Virol, 2004. **78**(16): p. 8771-9.
11. Y. Guan, L.L. Poon, C.Y. Cheung, T.M. Ellis, W. Lim, A.S. Lipatov, K.H. Chan, K.M. Sturm-Ramirez, C.L. Cheung, Y.H. Leung, K.Y. Yuen, R.G. Webster, and J.S. Peiris, *H5N1 influenza: a protean pandemic threat.* Proc Natl Acad Sci U S A, 2004. **101**(21): p. 8156-61.
12. X.F. Wan, T. Ren, K.J. Luo, M. Liao, G.H. Zhang, J.D. Chen, W.S. Cao, Y. Li, N.Y. Jin, D. Xu, and C.A. Xin, *Genetic characterization of H5N1 avian influenza viruses isolated in southern China during the 2003-04 avian influenza outbreaks.* Arch Virol, 2005. **150**(6): p. 1257-66.
13. R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, and Y. Kawaoka, *Evolution and ecology of influenza A viruses.* Microbiol Rev, 1992. **56**(1): p. 152-79.
14. Y. Guan, J.S. Peiris, A.S. Lipatov, T.M. Ellis, K.C. Dyrting, S. Krauss, L.J. Zhang, R.G. Webster, and K.F. Shortridge, *Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR.* Proc Natl Acad Sci U S A, 2002. **99**(13): p. 8950-5.
15. R.G. Webster, K.F. Shortridge, and Y. Kawaoka, *Influenza: interspecies transmission and emergence of new pandemics.* FEMS Immunol Med Microbiol, 1997. **18**(4): p. 275-9.
16. J.D. Thompson, D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
17. C. Notredame, D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.
18. R.C. Edgar, *MSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
19. D.L. Swofford, *PAUP*: Phylogenic analysis using Parsimony*. 1998: Sinauer,. Sunderland, Massachusetts.
20. J. Felsenstein, *PHYLIP - Phylogeny Inference Package (Version 3.2).* Cladistics, 1989. **5**: p. 164-166.
21. F. Ronquist and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-4.
22. T. Lassmann and E.L. Sonnhammer, *Quality assessment of multiple alignment programs.* FEBS Lett, 2002. **529**(1): p. 126-30.
23. X.-F. Wan, X. Wu, G. Lin, S.B. Holton, R.A. Desmone, C.R. Shyu, Y. Guan, and M. Emch, *Computational Identification of Reassortments in Avian Influenza Viruses.* Avian Diseases, 2007. **51(s1)**: p. 434-439.
24. X. Wu, X.-F. Wan, G. Wu, D. Xu, and G. Lin, *Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method.* International Journal of Bioinformatics Research and Application, 2006. **2**(3): p. 219-248.
25. X.F. Wan, G. Chen, F. Luo, M. Emch, and R. Donis, *A quantitative genotyping method reflecting H5N1 avian influenza niches.* Bioinformatics, 2007. **In press**.
26. X.F. Wan, S.M. Bridges, and J.A. Boyle, *Revealing gene transcription and translation initiation patterns in archaea, using an interactive clustering model.* Extremophiles, 2004. **8**(4): p. 291-9.
27. S. Varma and R. Simon, *Iterative class discovery and feature selection using Minimal Spanning Trees.* BMC Bioinformatics, 2004. **5**: p. 126.

28. S. Climer and W. Zhang. *A Traveling Salesman's Approach to Clustering Gene Expression Data*. in *Washington University, WUCSE-2005-5*. 2005.
29. Y. Xu, V. Olman, and D. Xu, *Minimum spanning trees for gene expression data clustering.* Genome Inform Ser Workshop Genome Inform, 2001. **12**: p. 24-33.
30. Y. Xu, V. Olman, and D. Xu, *Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees.* Bioinformatics, 2002. **18**(4): p. 536-45.
31. D. Posada and K.A. Crandall, *MODELTEST: testing the model of DNA substitution.* Bioinformatics, 1998. **14**(9): p. 817-8.
32. R.B. Belshe, *The origins of pandemic influenza--lessons from the 1918 virus.* N Engl J Med, 2005. **353**(21): p. 2209-11.
33. Y. Guan, M. Peiris, K.F. Kong, K.C. Dyrting, T.M. Ellis, T. Sit, L.J. Zhang, and K.F. Shortridge, *H5N1 influenza viruses isolated from geese in Southeastern China: evidence for genetic reassortment and interspecies transmission to ducks.* Virology, 2002. **292**(1): p. 16-23.
34. Y. Guan, K.F. Shortridge, S. Krauss, P.S. Chin, K.C. Dyrting, T.M. Ellis, R.G. Webster, and M. Peiris, *H9N2 influenza viruses possessing H5N1-like internal genomes continue to circulate in poultry in southeastern China.* J Virol, 2000. **74**(20): p. 9372-80.
35. Y. Guan, J.S. Peiris, L.L. Poon, K.C. Dyrting, T.M. Ellis, L. Sims, R.G. Webster, and K.F. Shortridge, *Reassortants of H5N1 influenza viruses recently isolated from aquatic poultry in Hong Kong SAR.* Avian Dis, 2003. **47**(3 Suppl): p. 911-3.
36. R.A. Fouchier, V. Munster, A. Wallensten, T.M. Bestebroer, S. Herfst, D. Smith, G.F. Rimmelzwaan, B. Olsen, and A.D. Osterhaus, *Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls.* J Virol, 2005. **79**(5): p. 2814-22.
37. G.J. Smith, T.S. Naipospos, T.D. Nguyen, M.D. de Jong, D. Vijaykrishna, T.B. Usman, S.S. Hassan, T.V. Nguyen, T.V. Dao, N.A. Bui, Y.H. Leung, C.L. Cheung, J.M. Rayner, J.X. Zhang, L.J. Zhang, L.L. Poon, K.S. Li, V.C. Nguyen, T.T. Hien, J. Farrar, R.G. Webster, H. Chen, J.S. Peiris, and Y. Guan, *Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam.* Virology, 2006. **350**(2): p. 258-68.
38. A.M. Kilpatrick, A.A. Chmura, D.W. Gibbons, R.C. Fleischer, P.P. Marra, and P. Daszak, *From the Cover: Predicting the global spread of H5N1 avian influenza.* Proc Natl Acad Sci U S A, 2006. **103**(51): p. 19368-73.



Xiu-Feng (Henry) Wan received his Ph.D. in Veterinary Medicine and M.Sc. in Computer Science from Mississippi State University both at 2002. He received his M.S. in Avian Medicine from South China Agricultural University and Veterinary Degree from Jiangxi Agricultural University. He finished his bioinformatics and computational biology postdoc trainings in Oak Ridge National Laboratory (2002-2003) and University of Missouri, Columbia (2003-2005). He was Assistant Professor in Department Microbiology at Miami University from 2005-2007. Recently, he joined Molecular Virology and Vaccine Branch, Influenza Division, the Centers for Disease Control and Prevention, Atlanta, Georgia. Dr. Wan is a member of American Society of Microbiology and International Society for Computational Biology.

Mufit Ozden received his Ph.D in Engineering Systems from UCLA. His research interest is in optimization techniques and simulation models that he has applied to a variety of problems ranging from computer systems to bioinformatics. For the last 10 years, he has been engaged in developing software for optimal design and operation of large-scale complex systems with random behavior. He has developed a set of novel simulation model components based on a machine-learning technique that can be used as the adaptive optimizers in large-scale models. He is currently developing programs for 3D-visualization of clusters.

Guohui Lin is an Associate Professor of Computing Science at the University of Alberta. He received his PhD in Theoretical Computer Science from the Chinese Academy of Sciences in 1998. His research interests include Bioinformatics and Computational Biology, and the recent work focuses on whole genome phylogenetic analysis for viruses, cancer bioinformatics, and bovine genomics. He is a member of ACM and a member of IEEE Computer Society.