Mixing Patterns in a Global Influenza A Virus Network Using Whole Genome Comparisons

Adrienne E. Breland, Mehmet H. Gunes, Karen A. Schlauch and Frederick C. Harris Jr.

Abstract—**Approximating 'real' disease transmission networks through genomic sequence comparisons among pathogenic isolates is increasingly feasible with the current growth in genomic sequence data. Here, we derive a network from over 4,200 globally distributed influenza A virus isolates based on alignment-free sequence comparisons. We then employ network mixing pattern analysis to examine transmission probabilities between isolates from different global regions, host types, subtypes and collection years. While we can not use our results to describe the complete global network of influenza A virus, we present a novel analytical process. In addition, we describe some of the characteristics of this subset of currently available data. Most notable results are the high levels of inter regional links and the important role that avian species seem to play in non human global transmission.**

I. Introduction

FRom a graph theoretic approach, a disease network may be viewed as a vertex and edge graph where individuals or groups are represented by vertices and their pairwise transmissive potential is indicated by edges. Determining 'real' disease networks to characterize the spatial and temporal structure of past epidemiological events can rely on subjective data collection and/or require extensive research [1]. Due to collection methods of the data required for recreating disease transmission networks, resulting graphs may be tree like [1], relatively small, and mis-representative of the networks complexity.

Examining real networks in both the 1981 Hong Kong SARS outbreak [2] and in the persistence of gonorrhea in a localized community [3] led both teams of researchers to similar conclusions regarding transmission network structures. Mainly that, in these cases, disease spread and persistence were attributed to super spreaders. Super spreaders are small numbers of individuals who, by maintaining disproportionately large numbers of contacts, are responsible for a disproportionately large number of transmission events. These indicated that underlying interactions formed scale-free networks [4]. However, much research in disease network modeling has gone in the reverse route of applying existing theoretical network types to derive characteristics about diseases [5] [6]. Other approaches include dynamic network creation through agent based computational models [7], and estimating transmission links based on geographic proximity [8] [9].

In this report, we explore the potential for the use of genomic sequence comparisons to approximate the topology of real and complex disease networks. The use of genomic sequences draws from an ever increasing supply of data and computational power. At the same time, whole genome comparison methods provide a quantitative approach to represent evolutionary relationships. We assume in this report that phylogenetic relatedness can imply approximate transmission pathways through a population. As a network graph, sequenced isolates are viewed as vertices and edge relationships are drawn according to comparative sequence scores. Whole sequence similarity scores are particularly attractive in the case for the analysis of RNA viruses such as influenza A, which mutate rapidly [10] and are relatively small. While small sequence sizes reduce computations, high mutation rates create a potentially traceable micro-evolutionary pathway through sequence comparisons. Isolate data also have the advantage of known collection locations. These may then be converted to geographic coordinates for each node to incorporate the important spatial component of transmission networks.

In our experiments, we build a network from over 4,200 globally distributed influenza A virus isolates. In addition to regional diversity, the dataset includes isolates from multiple host species, subtypes, and a broad range of dates. We infer isolate relationships from alignment free whole genome sequence comparisons. This allows us to examine mixing patterns between host species, geographic regions, subtypes and collection years. Mixing patterns can indicate disease transmission between samples from different classes such as species type or region. In a broader sense, we present an approach to epidemiological exploration through the formation of genomic networks.

In the following, section II presents an overview of current knowledge and approaches to understanding influenza transmission patterns through genomic data. Section III describes the sequence data set we used and the alignment-free method employed to derive sequence distance scores and resulting contact networks. Section IV describes the derivation of mixing patterns among different class types and results. Section V provides a conclusion of findings and goals for future work.

II. Background

The Influenza A virus, upon which this study is based, has been particularly well sampled. The pandemic potential of Influenza coupled with its high distribution across the globe

Adrienne E. Breland, Mehmet H. Gunes, Frederick C. Harris Jr. are with the Department of Computer Science and Engineering, University of Nevada Reno (email: {brelanda, fredh, mgunes}@cse.unr.edu).

Karen A. Schlauch is with the Dpartment of Biochemistry and Molecular Biology at the University of Nevada(email:schlauch@unr.edu).

TABLE I

Dominant classes ($\geq 0.1\%$) found in data set

Descriptor	class type (percent(%))
Region	USA(51.5%), New Zealand (12.8%) China(5.3%), Italy(1.8%)
Host Class	Human(65%), Avian(20.5%), WildAvian(7.8%), Swine(3.1%), Environment(2.7%)
Subtype	H3N2(33.8%), H1N1(32.6%), H5N1(11.1%), H7N1(4.4%), H9N2(2.8%), H3N8(1.9%),H1N2(1.4%), H7N3(1.4%), H6N1(1.3%), H4N6(1.2%)
Year	2009(17.9%), 2007(15.5%), 2005(14.2%), 2004(8.7%), 2003(7.9%), 2000(6.8%),2006(6.6%), 2002(6.5%), 2008(5.9%), 2001(5.7%), 1999(4.6%)

has resulted in a high rate of sampling, analysis and surveillance [11]. The National Institute of Allergy and Infectious Diseases's Influenza Virus Resource website currently houses over 70,000 influenza viral sequences spanning multiple host species, decades, subtypes and geographic locations [12].

Much attention has been given to estimating the geographic transmission routes of influenza virus through both human and avian populations. The overall global directionality of the human host H3N2 subtype has most recently been described as starting in East-Southeast Asia, then passing to Oceania, and through North America and Europe to South America [13]. Extensive seasonal global migration of influenza A viral strains, as opposed to localized persistence and re-emergence, has also been supported by research in [14]. In [13], the majority of sequence comparisons were performed antigenically rather than genetically. While antigenic comparisons can track how the virus evolves with regards to human host immunity, it focuses only on the functionality of specific region(s) of the entire flu genome. In contrast, whole genome comparisons may allow a finer level of differencing as information representing all point mutations and re-assortments are included in complete genomic sequences.

In addition to global circulation tendencies, much of the focus on influenza A epidemiology has been in tracking the origins of emergent subtypes. Methods are generally based in the creation of phylogenetic trees. While phylogenetic trees are basic networks in the form of tree graphs, this methodology is not designed to encompass the amount of data available and required to characterize complex networks. In [14], even though 900 complete genomes from the Northern and Southern Hemispheres were compared with phylogenies, these did not suggest a specific network of viral movement. In similar studies attempting to characterize the movement of flu viruses in North America [15], China and Southeast Asia [16] [17] [18] [19] [20], India [21], Europe [22] [23] and Africa [24], even smaller sample sizes are used and compared via antigenic or phylogenetic tree comparisons. Consistent in most studies is a call for increased surveillance of influenza and more comprehensive data sets [14] [13]. Similarly, in a recent influenza study conducted in [25], it is stated that there has been "no rigorous measurement of viral diversity across time, across space, and among subtypes" despite data availability.

Here we utilize network theory to examine global influenza transmission. In contrast with phylogenies, complex networks may be composed of multiple relationships between samples, or vertices, and are not constrained by requiring the delineation of each samples progenitor. Instead, we may represent our data and all inter relationships as they are measured and then search for network properties which may emerge.

III. Methods

A. Data Sets and Classes

We obtained genomic sequences of influenza isolates from the Influenza Virus Resource [12]. Data included 4,228 complete influenza A genomes collected between 1999-2009. When acquired from the database, each sequence is annotated with information describing host species, collection location, collection date, subtype, and other types of ancillary information. For four descriptors including region (country of origin), host type, subtype, and collection year, we derived class types and assigned all samples one class type for each descriptor. Distributions of the most dominant types are listed in Table I. We considered all host species as either Human, Domestic Avian, Wild Avian, Swine, Mammal, Environment, or Unknown. The Domestic Avian class included all viral samples from hosts labeled chicken, turkey, duck, or goose. Samples from hosts labeled by wild bird species, such as "mallard", or "Egret" were considered Wild Avian. The Mammal class was broad and included all non-human and non-swine mammals, including species such as horse, civet and tiger. The Human, Swine, Environment and Unknown classes were clearly labeled and required no generalization. The dataset also included 61 different subtypes and 58 countries which required no reclassification.

Influenza genomes exist in eight discontinuous segments unlike the more common contiguous stranded model of genetic material found in larger organisms. Thus, the full set of sequences was divided into eight groups containing each of the eight segments composing influenza genomes. Each of these sets were compared separately, and then results were combined to derive overall network relationships.

B. Genomic Comparisons

Minimizing computation time is one of the largest challenges when comparing many sequences of whole genomes. To delineate a network containing N nodes, $N\left(N-1\right)/2$ comparisons are required. Thus, it is necessary to utilize a computationally efficient method for sequence comparisons to create networks in which $N \approx 4200$. Sequence comparison methods may be classified as alignment based or alignment free. The complexity of alignment based methods remains $O\left(n^2\right)$ where $n = \textit{max sequence length}$. Alignment

TABLE II
ACTUAL AND PREDICTED COMPUTATION TIMES

N	FFP	$BLAST$	$MUMmer$
$2^{(act)}$	0.008s	0.051s	0.178s
$10^{(act)}$	0.86s	51.2s	2m 55.2s
$100^{(act)}$	3.49s	3m 27s	11m 47s
$1000^{(pred)}$	1m 33.9s	5h 47m 53.4s	19h 49m 2.8s
$4000^{(pred)}$	1h 33m 52.6s	91h 50m 23s	317h 19m 1.8s

free methods rely on the comparison of global sequence statistics and can run on the order of $O(n)$. Generally, linear time methods are based on sequence-specific *k*-mer distributions which reflect the frequency of *k*-length words in any given sequence. Due to the small alphabet size of all genomic sequences $\{a, c, g, t\}$, there exist only 4^k possible nucleotide words for any integer value assigned to *k*. Comparing the relative frequencies of all possible *k*-length nucleotide words is thus a computationally feasible task for even relatively large values of *k*. Frequencies of all existing *k*-mers differ among even closely related sequences such that distinctions among related prokaryotic groups have been determined based on this approach [26] [27] [28].

We utilized Feature Frequency Profiles (FFP), a *k*-mer based formula described in [26] to derive distance matrices between isolates. The FFP for a given sequence is a vector of length 4^k in which each entry contains the relative frequency of each *k*-mer in that sequence. In summary, to compare two genomic sequences (s_1, s_2), the FFP vector F_k is computed for each sequence. Obtaining the FFP for a given sequence first entails a linear time parsing to obtain counts of each possible *k*-mer for a specified value of *k*, yielding a count vector, C_k. The FFP profile F_k is then obtained by normalizing each vector element in C_k by the total number of *k*-mers found in a sequence such that, $F_k = C_k / \sum_{w=1}^{4^k} c_{w,k}$. A measure of dissimilarity between two sequences can then be computed as the sum of element-wise differences between frequency profiles.

In order to compute element-wise differences, we utilize the Jenson-Shannon (JS) Divergence, drawn from information theory as in [26]. Let P_k and Q_k represent FFP's for s_1 and s_2 respectively and $M_k = (P_k + Q_k)/2$. The JS Divergence is then calculated as, $JS_k(P_k, Q_k) = (1/2)KL(P_k, M_k) + (1/2)KL(Q_k, M_k)$, where the Kullback-Leibler Divergence is $KL(P_k, M_k) = \sum_{w=1}^{4^k} p_{k,w} \log_2 (p_{k,w}/m_{k,w})$. The value given to *k* in this type of comparison need also be determined. In [26], a method for determining the optimal range of *k* values for a given dataset was proposed. This method was derived from optimal word lengths used to distinguish books by different authors and regarding different subjects using the English language. We used the lower limit of this range per segment data set, as smaller word length values allowed slight increases in computational speed. The lower limit for *k* is determined by the value at which increasing *k* does not increase the number of *k*-length words found at least twice in each sequence of the data set. Thus if $N(k)$ is the total number of distinct *k*-mers found to occur at least twice in all members of a set of sequences, the lower limit of *k* may be defined as *k* such that $N(k) \geq N(k+1)$. This indicates the maximum word length at which new functional words are found, or the length of maximum word diversity.

An in house implementation of the FFP method allowed rapid distance calculations so that it was feasible to derive all-against-all comparisons of 4,228 complete sequences. Table II provides real and predicted computation times relative to sequence number for all against all comparisons using FFP and two other commonly used applications for genomic sequence comparison; BLAST [29] and MUMmer [30]. Predicted times were derived from linear interpolation based on real time values. As seen in Table II, the predicted computation time using the FFP method to compare 4,000 nodes is approximately 1.5 hours. In contrast, BLAST and MUMmer were predicted to require over 91 and 317 hours respectively to perform the same number of comparisons.

We compared the FFP method against BLAST to estimate its accuracy for determining most similar sequence, or minimal distance relationships among our data set. These relationships form the basis of contact networks and are described in more detail in the following section. From each of the eight flu segment data sets, ten sequences were randomly selected. Pairwise similarity and distance matrices were computed using BLAST and FFP respectively. For each of the ten sequences, the most similar sequence(s) within the set of ten were determined according to both BLAST and FFP scores. The most similar sequence(s) using BLAST were those which exhibited the highest bit scores. The bit score reflects the overall alignment of high scoring pairs incorporating gap penalties [31]. The most similar sequence(s) using the FFP method were those determined by the minimal computed Jenson-Shannon Divergence scores. For all ten sequences, comparisons were then made between their most similar sequences determined by each method. For example, if s_1 was most similar to s_2 using the FFP comparison, and also most similar to s_2 according to the BLAST comparison, this would account for a match. We repeated this for 200 random samples of size ten. The resulting average percentage of matches was 20% - 30% across all segments. These results indicate a need for more accurate and efficient comparative approaches of subspecies viral isolates. It is the focus of our current research.

C. Contact Networks

In epidemiological studies, a contact network is used to indicate disease transmission between individuals. Here, we derived contact networks based on minimal FFP sequence distance scores. Thus, an edge indicating contact is drawn between each sequence and all others exhibiting a minimum relative score.

Separate contact networks were derived from distance matrices for each of the eight flu segments. These took the form of undirected N x N adjacency graphs G in which indices G_{ij}, are given the value of 1 to indicate minimal

TABLE III
AVERAGE DEGREE PER SEGMENT

segment	$\bar{k}$
1	6.96
2	11.41
3	10.31
4	6.09
5	11.29
6	8.0
7	18.47
8	16.14

differences between sequences i and j , $i \in N, j \in N, i \neq j$. All other indices are marked with a 0.

For each sample $i \in N$, the minimum distance MIN_i between i and all other samples $j \in N$ was determined. Then any sample j for which the computed distance between samples i and j was equal to MIN_i was considered a contact, and a value of 1 was assigned to G_{ij}. This allowed that each sample could be assigned more than one link if the same minimum distance value was found with multiple samples. Multiple links were observed in most cases. The degree(k) of a vertex is the total number edges connecting to it while the average degree ($\bar{k}$) is the average of all vertex degrees in a graph. Each flu segment specific adjacency graph exhibited a high average degree, with values ranging from $6.09 - 18.74$. Table III provides a list of the average degrees of each segment contact network.

IV. MIXING PATTERN RESULTS

Mixing patterns describe the probability of connections between vertices of different types in a network [32]. Mixing patterns in our dataset were examined with regards to region, host class, year and subtype. This was to examine our network for indications of highly probable geographic transmission routes, cross species transmission, subtype mixing and carry over of genotypes between years. Thus for each vertex of type i in a network, we compute the conditional probability that its network neighbor is of type j, i.e. $P(j|i)$.

To examine mixing patterns among vertex types, a mixing matrix $\mathbf{E}$ is built where E_{ij} contains the number of edges connecting vertexes of type i to vertexes of type j. Because our data were divided into eight distinct sets, each entry in the resulting matrix E_{ij} contained the sum of edges found among all segments one through eight. A normalized mixing matrix $\mathbf{e}$ is then derived where

$$\bar{\mathbf{E}} = \frac{\mathbf{E}}{||\mathbf{E}||}$$

$||\mathbf{E}||$ representing the sum of all elements in $\mathbf{E}$. $P(j|i)$ for each vertex type i and all neighbor types j may then be derived by $P(j|i) = \bar{E}_{ij}/\sum_j \bar{E}_{ij}$, as described in [32]. Entries in the matrix $\bar{E}_{ij}$ measure the link strength between all class types of a given descriptor, e.g., country of origin, host class. Mixing matrices and resulting conditional probabilities among all class types were derived from contact network graphs. Links among all eight segment matrices were counted individually. This allowed that each link could represent similarity between samples in a single segment, independent of all others.

An assortativity coefficient was also calculated for each mixing pattern analysis. Assortativity refers to the selectivity of vertices in forming links within the same class. Using assortativity measures, the prevalence of inter host species transmission, "host-jumping" may be examined for example. In [33], a method for deriving an assortativity coefficient (r) is given as

$$r = \frac{Trace\left(\bar{\mathbf{E}}\right) - \left\|\bar{\mathbf{E}}^2\right\|}{1 - \left\|\bar{\mathbf{E}}^2\right\|}$$

More detail regarding this equation can be found in [33]. Thus a higher selectivity is reflected as r approaches 1 in a given data set with regards to a specific descriptor. A value close to 1 would indicate that vertices of a given type tend to form links with only vertices of the same type.

In the following, we report both the assortativity and the interclass (between class) mixing statistics for regional, host, year and subtype groups in our global influenza dataset. Class type relations are displayed as networks in which vertexes represent classes and edges represent their conditional interlinking probabilities. Edge widths are indicative of the magnitude of these values. Tables listing conditional inter class linking probabilities are presented as well.

A. Regional Mixing Patterns

Regional mixing patterns were derived to indicate global geographic transmission patterns. Regional classes described the country of origin among 58 countries for each sample. Regional mixing patterns showed the lowest assortativity out of all class groupings, with an r value of 0.678. Thus, regional selectivity was notably less than 1.0 and several links across geographic borders were found. The average link strength between countries was 0.069 with a range of values between 0.000026 to 1.0. Figure 1 shows an interclass network displaying all inter-regional links that exceed the average value of 0.069. There were close to one hundred $P(j|i)$ values which exceeded this average so we selected a small subset of fifteen to report in Table IV. To report countries with the highest number of inter-regional links, values were ranked by $degree(i)$.

B. Host Class Mixing Patterns

Host classes mixing patterns were examined to detect the occurrence of cross host type transmission. The assortativity coefficient for host class mixing patterns in our data set was 0.789. This reflected selectivity for within class transmission and also indicated a marked presence of inter host transmission. The most linked classes were (not surprisingly) Domestic and Wild Avian. It is worth noting as well that both the Environment and Unknown classes are most prominently linked to the Avian classes. This suggests the introduction of influenza A in to the environment by avian species. The average link strength between classes was 0.065 with a range of values between 0.0000636 to 0.532. Figure 2 displays a

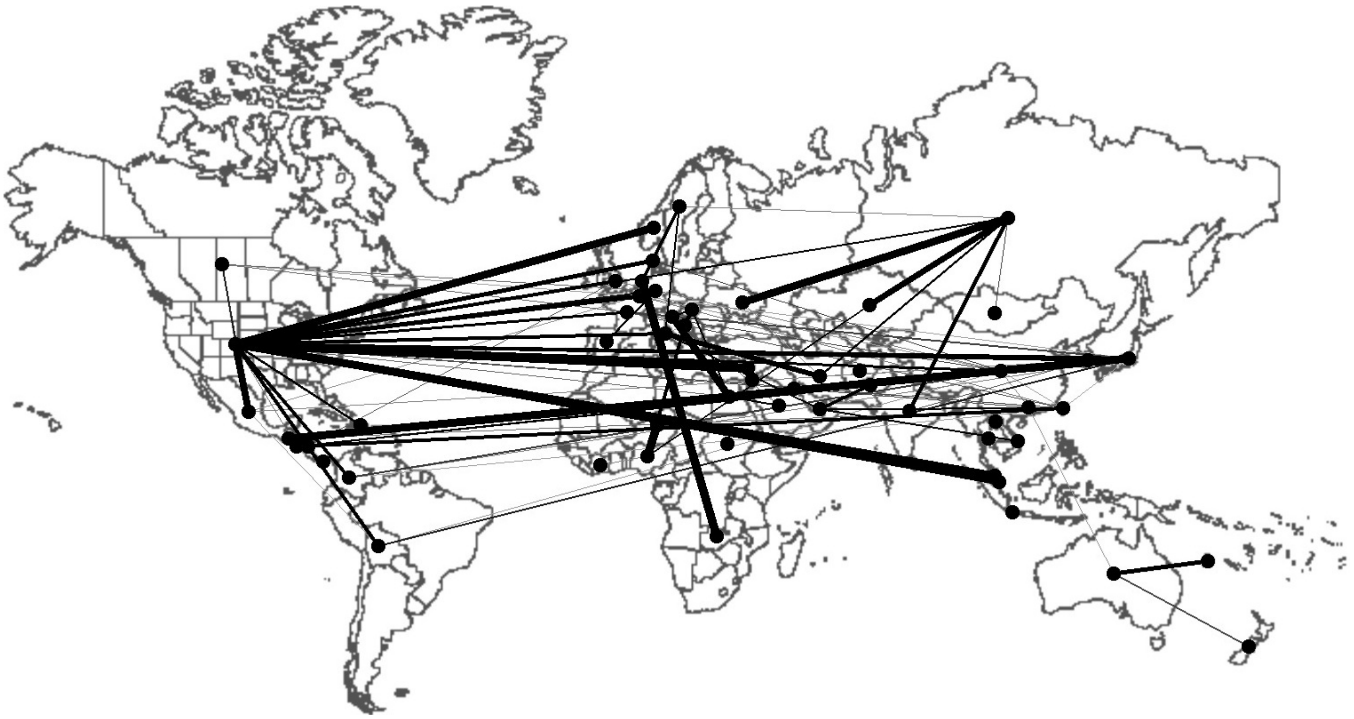

Fig. 1. Inter regional network

TABLE IV
STRONGEST INTER REGIONAL LINK VALUES

$P(neighbor(j)\|class(i))$	$P(j\|i)$	degree(i)
P(USA|China)	0.289	34
P(Japan|China)	0.100	34
P(USA|Russia)	0.206	27
P(USA|Japan)	0.369	26
P(China|Japan)	0.098	26
P(USA|Italy)	0.402	25
P(Russia|Mongolia)	0.146	19
P(Mexico|Canada)	0.178	17
P(USA|Canada)	0.136	17
P(USA|Taiwan)	0.142	17
P(China|Taiwan)	0.073	17
P(Japan|Taiwan)	0.071	17
P(NewZealand|Australia)	0.150	16
P(USA|Colombia)	0.465	16
P(China|Colombia)	0.157	16

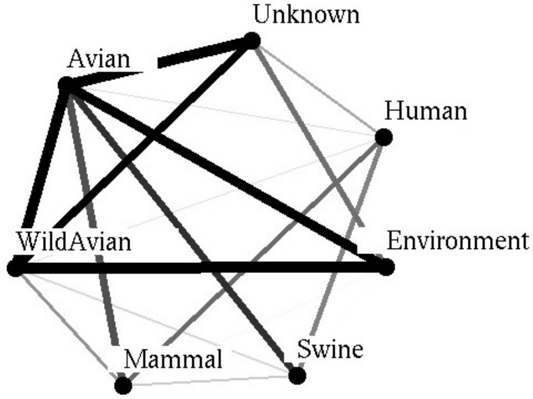

Fig. 2. Inter host network, all links

complete inter host network. Figure 3 displays only links which exceed the average value. Figures 2 and 3 suggest a strong role in non human global transmission attributable to domestic and avian bird species. As most wild avian partake in some form of migration, and domestic avian species are broadly traded, this does not conflict with expected results. Table V lists all higher than average inter class conditional relationships.

C. Subtype Class Mixing Patterns

Of the 61 different subtypes in the dataset, the assortativity coefficient of within subtype links was 0.96. While this indicates a high degree of selectivity, inter subtype links were found as well. The average inter subtype linking probability was 0.057 and ranged from 0.0000118 to 1.0. There were 176 relationships found which exceeded the average value. These inter subtype links are displayed in Figure 4 while Table VI lists only the 15 top ranking relationships based on the highest degree of vertices i and higher than average

TABLE V
STRONGEST INTER HOST TYPE LINK VALUES

$P(neighbor(j)\|class(i))$	$P(j\|i)$	degree(i)
P(Avian|WildAvian)	0.532	7
P(Avian|Unknown)	0.356	5
P(WildAvian|Avian)	0.223	7
P(Avian|Environment)	0.149	5
P(WildAvian|Unknown)	0.119	5
P(WildAvian|Environment)	0.112	5
P(Environment|Avian)	0.0726	7
P(Avian|Swine)	0.0811	6
P(Avian|Mammal)	0.0681	4

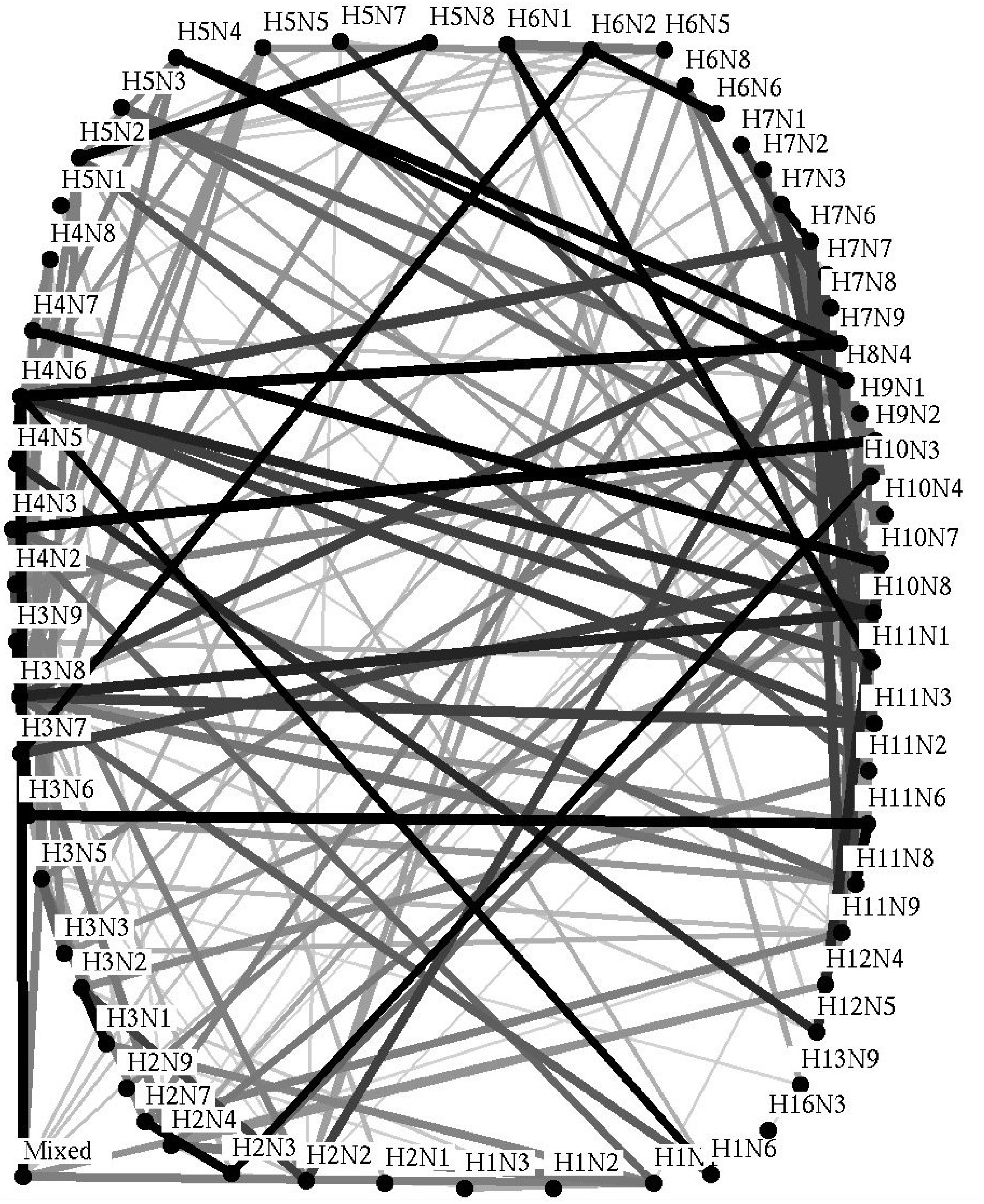

Fig. 4. Inter subtype network

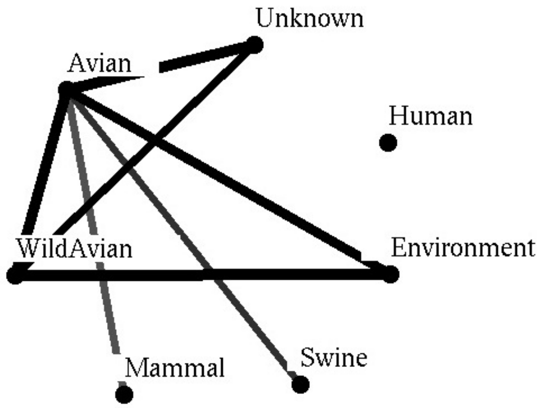

Fig. 3. Inter host network, filtered links

$P(i|j)$ values.

D. Year Class Mixing Patterns

Mixing patterns between years, which ranged from 1999 to 2009, may indicate genotype persistence across successive flu seasons. The assortativity coefficient among year classes was 0.9. The average inter year link probability was 0.028 and ranged from 0.00015 to 0.308. Figure 5 displays links with values above the average and Table VII lists the top ten ranking inter year relationships. Interestingly, eight out of the ten relationships listed in Table VII contain relationships between successive years. The anomalies are links between the pairs (2000,2003) and (2005,2007). These links are visually defined in Figure 5. This may reflect at least partial genotype persistence between years and/or flu seasons which span successive years. However anomalous year pairings merit further investigation.

V. Conclusions and Future Work

Mixing patterns in a network can quantify how individuals, or vertices, of different types exhibit network links. This study describes a methodological approach for deriving mixing patterns within a pathogenic genomic data set. It also describes a novel approach to characterizing influenza disease networks. Examination of networks such as these may provide insight into epidemiological characteristics such as viral movement and transmission though geographic regions and through specific population groups.

TABLE VI
Strongest inter subtype link values

$P(neighbor(j) \mid class(i))$	$P(j \mid i)$	degree(i)
P(H5N2$\mid$H6N2)	0.083	31
P(H6N1$\mid$H6N2)	0.065	31
P(H4N6$\mid$H11N9)	0.064	31
P(H5N2$\mid$H11N9)	0.064	31
P(H5N3$\mid$H5N2)	0.085	29
P(H6N2$\mid$H5N2)	0.067	29
P(H5N8$\mid$H5N2)	0.063	29
P(H3N8$\mid$H4N8)	0.156	24
P(H4N6$\mid$H4N8)	0.102	24
P(H5N7$\mid$H10N7)	0.097	23
P(H3N8$\mid$H3N6)	0.151	20
P(H4N6$\mid$H3N6)	0.067	20
P(H6N1$\mid$H6N5)	0.175	19
P(H5N2$\mid$H6N5)	0.075	19

TABLE VII
Strongest inter year link values

$P(neighbor(j) \mid class(i))$	$P(j \mid i)$	degree(i)
P(2000$\mid$1999)	0.308	11
P(2005$\mid$2004)	0.224	6
P(2007$\mid$2006)	0.194	4
P(2002$\mid$2001)	0.117	9
P(2006$\mid$2005)	0.0912	10
P(2004$\mid$2003)	0.0858	7
P(2008$\mid$2007)	0.0729	5
P(2003$\mid$2002)	0.0619	8
P(2007$\mid$2005)	0.0606	10
P(2001$\mid$2000)	0.059	8
P(2003$\mid$2000)	0.0421	8

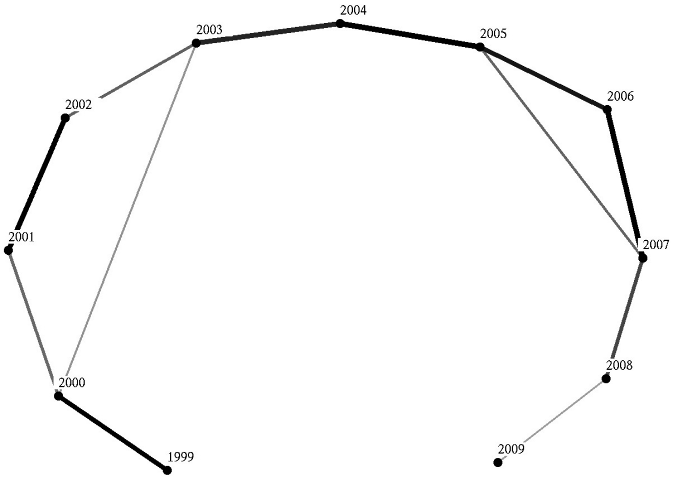

Fig. 5. Inter year network

Our results revealed high levels of selectivity $(r \geq 0.90)$ for year and subtype groups. However, inter class links were found for all classification schemas. This is suggestive of a highly complex global transmission network, with several aspects which can be examined independently and in more depth.

The lowest assortativity measure among all classification schemas was found in regional groupings with several highly weighted links between the USA and countries such as China, Russia, and Japan. However, this may result in part from a large percentage of samples in our data set being from the USA. Host class mixing patterns suggest a strong role in global transmission through avian species, as these host types showed heavily weighted links with all other non human host class types. The mixing patterns among subtypes were numerous although we did not draw any other conclusions. Inter year mixing pattern resulted in the most dominant links being between successive years, suggesting a yearly progression of genotype evolution.

Future work will investigate the underlying network structure of influenza A in more evenly distributed data sets. It will include directed, rather than undirected links to indicate transmission from isolates collected at an earlier date to those collected later. More broadly, we will examine how varied approaches to each component of our study may affect resulting networks. Our current research involves the development of fast comparison methods directed specifically towards whole genome comparisons of subspecies viral sequences. These methods include the selection of specific *k*-mer subsets for better distinction among biological classes and may be extracted from a given dataset at run-time. Subsets may be determined by properties such as GC-rich content or frequency distributions, for example. It is our goal to improve upon the accuracy and robustness of these methods to allow critical insight into the global influenza network.

References

[1] M. Keeling and K. T. D. Eames, "Networks and epidemic models," *J R Soc Interface*, vol. 2, no. 4, pp. 295–307, Sep 2005.

[2] S. Riley, C. Fraser, C. A. Donnelly, A. C. Ghani, L. J. Abu-Raddad, A. J. Hedley, G. M. Leung, L.-M. Ho, T.-H. Lam, T. Q. Thach, P. Chau, K.-P. Chan, S.-V. Lo, P.-Y. Leung, T. Tsang, W. Ho, K.-H. Lee, E. M. C. Lau, N. M. Ferguson, and R. M. Anderson, "Transmission dynamics of the etiological agent of sars in hong kong: impact of public health interventions," *Science*, vol. 300, no. 5627, pp. 1961–1966, Jun 2003.

[3] J. A. Yorke, H. W. Hethcote, and A. Nold, "Dynamics and control of the transmission of gonorrhea," *Sex Transm Dis*, vol. 5, no. 2, pp. 51–56, Apr/Jun 1978.

[4] A.-L. Barabasi and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. [Online]. Available: http://www.sciencemag.org/cgi/content/abstract/286/5439/509

[5] J. M. Read and M. J. Keeling, "Disease evolution on networks: the role of contact structure," *Proc Biol Sci*, vol. 270, no. 1516, pp. 699–708, Apr 2003.

[6] J. Saramäki and K. Kaski, "Modelling development of epidemics with dynamic small-world networks," *J Theor Biol*, vol. 234, no. 3, pp. 413–421, Jun 2005.

[7] J. M. Epstein, "Modelling to contain pandemics," *Nature*, vol. 460, no. 7256, p. 687, Aug 2009.

[8] S. Riley, "Large-scale spatial-transmission models of infectious disease," *Science*, vol. 316, no. 5829, pp. 1298–1301, Jun 2007.

[9] M. Small, D. M. Walker, and C. K. Tse, "Scale-free distribution of avian influenza outbreaks," *Phys Rev Lett*, vol. 99, no. 18, p. 188702, Nov 2007.

[10] E. C. Holmes, "Evolutionary history and phylogeography of human viruses," *Annu Rev Microbiol*, vol. 62, pp. 307–328, 2008.

[11] J. P. Chretien, J. C. Gaydos, J. L. Malone, and D. L. Blazes, "Global network could avert pandemics," *Nature*, vol. 440, no. 7080, pp. 25–26, Mar 2006.

[12] B. Yiming, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the national center for biotechnology information," *Virol J*, vol. 82, no. 2, Jan 2008.

[13] C. A. Russell, T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, J. C. de Jong, A. Kelso, A. I. Klimov, T. Kageyama, N. Komadina, A. S. Lapedes, Y. P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A. D. M. E. Osterhaus, G. F. Rimmelzwaan, M. W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R. A. M. Fouchier, and D. J. Smith, "The global circulation of seasonal influenza a (h3n2) viruses," *Science*, vol. 320, no. 5874, pp. 340–346, Apr 2008.

[14] M. I. Nelson, L. Edelman, D. J. Spiro, A. R. Boyne, J. Bera, R. Halpin, N. Sengamalay, E. Ghedin, M. A. Miller, L. Simonsen, C. Viboud, and E. C. Holmes, "Molecular epidemiology of a/h3n2 and a/h1n1 influenza virus during a single epidemic season in the united states," *PLoS Pathog*, vol. 4, no. 8, p. e1000133, Aug 2008.

[15] R. Chen and E. C. Holmes, "Frequent inter-species transmission and geographic subdivision in avian influenza viruses from wild birds," *Virology*, vol. 383, no. 1, pp. 156–161, Jan 2009.

[16] J. Wang, D. Vijaykrishna, L. Duan, J. Bahl, J. X. Zhang, R. G. Webster, J. S. M. Peiris, H. Chen, G. J. D. Smith, and Y. Guan, "Identification of the progenitors of indonesian and vietnamese avian influenza a (h5n1) viruses from southern china," *J Virol*, vol. 82, no. 7, pp. 3405–3414, Apr 2008.

[17] L. Duan, L. Campitelli, X. H. Fan, Y. H. C. Leung, D. Vijaykrishna, J. X. Zhang, I. Donatelli, M. Delogu, K. S. Li, E. Foni, C. Chiapponi, W. L. Wu, H. Kai, R. G. Webster, K. F. Shortridge, J. S. M. Peiris, G. J. D. Smith, H. Chen, and Y. Guan, "Characterization of low-pathogenic h5 subtype influenza viruses from eurasia: implications for the origin of highly pathogenic h5n1 viruses," *J Virol*, vol. 81, no. 14, pp. 7529–7539, Jul 2007.

[18] D. C. Nguyen, T. M. Uyeki, S. Jadhao, T. Maines, M. Shaw, Y. Matsuoka, C. Smith, T. Rowe, X. Lu, H. Hall, X. Xu, A. Balish, A. Klimov, T. M. Tumpey, D. E. Swayne, L. P. T. Huynh, H. K. Nghiem, H. H. T. Nguyen, L. T. Hoang, N. J. Cox, and J. M. Katz, "Isolation and characterization of avian influenza viruses, including highly pathogenic h5n1, from poultry in live bird markets in hanoi, vietnam, in 2001," *J Virol*, vol. 79, no. 7, pp. 4201–4212, Apr 2005.

[19] R. G. Webster, Y. Guan, M. Peiris, D. Walker, S. Krauss, N. N. Zhou, E. A. Govorkova, T. M. Ellis, K. C. Dyrting, T. Sit, D. R. Perez, and K. F. Shortridge, "Characterization of h5n1 influenza viruses that continue to circulate in geese in southeastern china," *J Virol*, vol. 76, no. 1, pp. 118–126, Jan 2002.

[20] A. N. Cauthen, D. E. Swayne, S. Schultz-Cherry, M. L. Perdue, and D. L. Suarez, "Continued circulation in china of highly pathogenic avian influenza viruses encoding the hemagglutinin gene associated with the 1997 h5n1 outbreak in poultry and humans," *J Virol*, vol. 74, no. 14, pp. 6592–6599, Jul 2000.

[21] K. Ray, V. A. Potdar, S. S. Cherian, S. D. Pawar, S. M. Jadhav, S. R. Waregaonkar, A. A. Joshi, and A. C. Mishra, "Characterization of the complete genome of influenza a (h5n1) virus isolated during the 2006 outbreak in poultry in india," *Virus Genes*, vol. 36, no. 2, pp. 345–353, Apr 2008.

[22] K. Bragstad, P. H. Jørgensen, K. Handberg, A. S. Hammer, S. Kabell, and A. Fomsgaard, "First introduction of highly pathogenic h5n1 avian influenza a viruses in wild and domestic birds in denmark, northern europe," *Virol J*, vol. 4, p. 43, May 2007.

[23] S. L. Salzberg, C. Kingsford, G. Cattoli, D. J. Spiro, D. A. Janies, M. M. Aly, I. H. Brown, E. Couacy-Hymann, G. M. De Mia, D. H. Dung, A. Guercio, T. Joannis, A. S. Maken Ali, A. Osmani, I. Padalino, M. D. Saad, V. Savić, N. A. Sengamalay, S. Yingst, J. Zaborsky, O. Zorman-Rojs, E. Ghedin, and I. Capua, "Genome analysis linking recent european and african influenza (h5n1) viruses," *Emerg Infect Dis*, vol. 13, no. 5, pp. 713–718, May 2007.

[24] F. O. Fasina, S. P. R. Bisschop, T. M. Joannis, L. H. Lombin, and C. Abolnik, "Molecular characterization and epidemiology of the highly pathogenic avian influenza h5n1 in nigeria," *Epidemiol Infect*, vol. 137, no. 4, pp. 456–463, Apr 2009.

[25] A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes, "The genomic and epidemiological dynamics of human influenza a virus," *Nature*, vol. 453, no. 7195, pp. 615–619, May 2008.

[26] G. A. Wu, S.-R. Jun, G. E. Sims, and S.-H. Kim, "Whole-proteome phylogeny of large dsdna virus families by an alignment-free method," *Proceedings of the National Academy of Sciences*, vol. 106, no. 31, pp. 12 826–12 831, August 2009. [Online]. Available: http://dx.doi.org/10.1073/pnas.0905115106

[27] M. D., R. P., W. E. H., and M.-T. H., "Genome phylogeny based on short-range correlations in dna sequences," *Journal of Computational Biology*, vol. 12, no. 5, pp. 545–553, 2005.

[28] A. Breland, S. Nasser, K. Schlauch, M. Nicolescu, and F. C. Harris, "Efficient influenza a virus origin detection," *Journal of Electronics and Computer Science*, vol. 10, pp. 1–11, December 2008.

[29] A. Pertsemlidis and J. W. Fondon, "Having a blast with bioinformatics (and avoiding blastphemy)." *Genome Biol*, vol. 2, no. 10, 2001.

[30] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes." *Genome Biol*, vol. 5, no. 2, 2004.

[31] S. F. Altschul, T. L. Madden, A. A. Schffer, R. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, pp. 3389–3402, 1997.

[32] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[33] M. J. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, pp. 026 126+, February 2003.