Fast Graph Approaches to Measure Influenza Transmission Across Geographically Distributed Host Types

Adrienne Breland
Dept. of Computer Science
University of Nevada, Reno
1664 N Virginia St
Reno, Nevada, 89557
ab17744@yahoo.com

Karen Schlauch
Dept. of Biochemistry
University of Nevada, Reno
1664 N Virginia St
Reno, Nevada, 895557
schlauch@unr.edu

Mehmet Gunes
Dept. of Computer Science
University of Nevada, Reno
1664 N Virginia St
Reno, Nevada, 89557
mgunes@cse.unr.edu

Frederick C. Harris Jr.
Dept. of Computer Science
University of Nevada, Reno
1664 N Virginia St
Reno, Nevada, 89557
fredh@cse.unr.edu

ABSTRACT

Recent advances in next generation sequencing are providing a number of large whole-genome sequence datasets stemming from globally distributed disease occurrences. This offers an unprecedented opportunity for epidemiological studies and the development of computationally efficient, robust tools for such studies. Here we present an analytic approach combining several existing tools that enables a quick, effective, and robust epidemiological analysis of large whole-genome datasets. In this report, our dataset contains over $4,200$ globally sampled *Influenza A virus* isolates from multiple host type, subtypes, and years. These sequences are compared using an alignment-free method that runs in linear time. This enables us to generate a disease transmission network where sequences serve as nodes, and high-degree sequence similarity as edges. Mixing patterns are then used to examine statistical probabilities of edge formation among different host types from different global regions and from different localities within Southeast Asia. Our results reflect notable amounts of inter-host and inter-regional transmission of *Influenza A virus*.

Categories and Subject Descriptors

E.1 [**Data**]: Graphs and networks

General Terms

Measurement

1. INTRODUCTION

As technology advances, we are incrementally presented with new combinations of increased computing power and larger numbers of complete genomic sequences. Computing hardware has so far adhered to Moore's law, which states that the number of transistors on a circuit board doubles approximately every two years. This has resulted in rapid and consistent improvements in processing speed and memory capacity [17]. At the same time, developing methods in sequencing technology (Next Generation Sequencing) promise to continue to provide faster and cheaper methods for sequencing entire genomes [10, 9]. The combination of advances in sequencing and computing means that more sequences and processing capabilities are constantly becoming available to researchers.

Epidemiological studies are poised to benefit substantially from such advances. The application of genomic sequence data to epidemiological discovery is becoming increasingly feasible with the current growth in genomic sequence data. Several collections of pathogenic subspecies genomes have and will continue to become publicly available. 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 has resulted in a high rate of sampling, analysis and surveillance [2]. 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 [19]. Advances in computational methods which combine high throughput processing capabilities and increasing data set sizes are now the largest challenge to analysis of pathogenic genomic information this[4]. Our work addresses this developing area of research by presenting a graph theoretic approach to approximating and examining possible disease transmission using networks derived from whole genomic sequence comparisons.

Current applications of genomic data to disease networks are based on phylogenetic tree inferences. These include phylogenies derived from multiple sequence alignments[11] and more recently, Bayesian phylogeography[8]. Because of computational and/or structural constraints, phylogenetic trees do not encompass the amount of data available and required to characterize complex networks or detailed transmission patterns[4, 15, 11]. In contrast, graph based approaches may bypass many of the computational and structural restrictions associated with phylogenetic trees and may be used to investigate very large and comprehensive data sets.

Here, we build a disease network graph from a large set of subspecies, *Influenza A virus* whole genomic sequences. We utilize the majority of *Influenza A virus* whole genome sets which are publicly available from the Influenza Virus Resource database[19]. In this graph, each sequence is viewed as a node, and edges are drawn between nodes which exhibit high degrees of similarity using a string comparison method. The goal here is not to approximate the comprehensive evolutionary history among a group of sequences, but to more simply develop a graph representation with sequences as nodes and strong degrees of similarity as edges. Because each node is associated with a geographic location, such a representation may provide an approximation of the transmission route of a disease through a series of geographically distributed hosts. The underlying assumption is that a strong degree of sequence similarity indicates the best estimate of transmission between hosts from which each sequence was collected.

We base our sequence similarity measures on a string comparison algorithm where the strings are whole genomic nucleotide sequences. Whole sequence similarity scores based on string comparison algorithms are particularly attractive in the case for the analysis of RNA viruses such as *Influenza A virus*, which mutate rapidly [15, 5] and are relatively small. Small sequence sizes reduce computation requirements. Furthermore, high mutation rates create a potentially traceable micro-evolutionary pathway through sequence comparisons [8].

Once a graph is generated from similarity measures, it opens up a number of possible analytical approaches based on graph theory and statistics. As the graph represents a possible disease transmission network, questions pertinent to disease transmission may be addressed with simple existing methods. For example, transmission across global regions or across host types may be examined with mixing pattern analysis [12] to study global circulation routes and interspecies transmission ('host jumping').

In this report we develop a method to build and examine a network based on sequence similarity of influenza A genomes. We examined mixing patterns across several geographic regions of interest and across four hosts groups including wild avian, domestic avian, human, and swine. These analyses address the current need for a more detailed understanding of global influenza circulation [15, 11] and of circulation in Southeast Asia [16]. Southeast Asia has been suggested as a global seeding network producing each year's seasonal epidemic strains in more temperate global regions. The first documented cases of the 'bird flu', H5N1 also originated in Southeast Asia where it crossed the species border between human and avian hosts [3]. Recent human infections with avian subtypes H7N7 and H9N2 have been identified as well in China [7].

2. METHODS

2.1 Genomic Sequence Data

We obtained genomic sequences of influenza isolates from the Influenza Virus Resource [19]. 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 descriptors of region (country of origin) and host type, we derive class types and assigned samples one class type for each descriptor. All of the represented host types included 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 dataset also included 61 different subtypes and 58 countries.

Influenza genomes exist in eight discontinuous segments unlike the more common contiguous stranded model of genetic material found in larger organisms. Sequence data were separated into eight groups associated to the genomic segments, and each group was examined individually. Individual results were combined to derive overall network relationships.

2.2 Sequence Comparison Method

Sequence comparisons comprise the first major computational step in the methodology we present here. Given a set of genomic sequences of size N, an $N \times N$ matrix containing distance scores is computed between all sequence pairs. In this report, $N = 4,228$, thus our choice of sequence comparison method was dictated by both computational speed and accuracy.

We used a k-mer based comparative algorithm which runs in linear time with respect to sequence lengths. A k-mer is a nucleotide word of length k, where k is an integer and $k > 0$. 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. A genomic sequence need only be parsed once to determine the number of times that each possible k-mer appears. Methods for comparing two sequences based on k-mer occurrences are generally performed on a word by word basis. Thus, the difference between 2-mer profiles of sequences A and B is the sum of differences between occurrences of all possible 2-mers $\{aa, ..., tt\}$ in each profile.

2.2.1 d_k^2 Distance

The sequence comparison method that we utilize here is a derivation of the k-mer based, d_k^2 ('d squared') distance described in [18]. The d_k^2 distance between two sequences is the sum of squared differences between the 4^k pairwise k-mer

counts in each sequence. Count based k-mer comparisons reflect differences in both sequence length and composition among different infA strains.

Computing the d_k^2 distance between two sequences (A and B) is presented in [18] as

$$d_k^2(A,B) = \sum_{i=w}^{4^k} p_w \left(c_w(A) - c_w(B)\right)^2 \quad (1)$$

where k is a fixed integer word length, $c_w(A)$ and $c_w(B)$ indicate counts of k-mer$_w$ in sequences A and B, respectively, and p_w is a weight associated with each k-mer.

2.2.2 Presence/Absence Weighting

In previous work [1], we compared influenza genomic sequences using observed to expected ratios of k-mer frequencies [14], and restricted the computation of distance to those k-mers that exhibited variation of presence and/or absence across sequences. These were k-mers present in at least one sequence and absent in at least one other sequence in the data set. We found that this method, when restricted to the subset of k-mers exhibiting presence/absence variation, classified sequences from the same epidemic as most similar. Lineage specific occurrences k-mers may be a reflection of inherited point mutations which cause measurable differences in k-mer counts and their presence or absence.

In this analysis, we also restrict our comparison to a subset of k-mers. A simple weighting scheme was derived in which weights are used to limit comparisons to k-mers which exhibit presence/absence variation among our entire set of sequences. Any k-mer that is absent in one sequence and present in one sequence is assigned a weight of value 1; all other k-mers are assigned a weight of 0. Then if p_w represents the weight for word w and $c_{w,i}$ the count of word w in sequence $i, i \in N$, then p_w in Equation 1 is given as

$$p_w = \begin{cases} 1 & if\ \sum_i^N c_{w,i} > 0 \text{ and } \prod_i^N c_{w,i} = 0 \\ 0 & else \end{cases} \quad (2)$$

2.2.3 Accuracy

We examined the performance of the d_k^2 comparison measure by comparisons with ClustalW, a standard multiple sequence alignment tool [6]. In small randomly selected sequence sets, we determined which sequence pairs were the most similar using the d_k^2 measure, and using the maximal ClustalW pairwise alignment scores.

Ten iterations of comparisons using randomly selected sets of 20 sequences we conducted per flu segment. Any nearest neighbor sequence pair found using d_k^2 distances which was also a nearest neighbor pair using the ClustalW alignment scores was considered an accurate pair. Conversely, any nearest neighbor pair formed using d_k^2 distances which was not formed using ClustalW was considered inaccurate. Accuracy scores were computed as the ratio of accurate verses inaccurate pairs. These scores were averaged over each iteration for each segment.

We also examined the accuracy scores to select the best integer value for k. Accuracy scores were computed for values of k ranging from 4 to 10. Our goal was to develop a fast *and* accurate sequence comparison method. Average accuracy was computed as the average score across the eight segments, as reported in Table 1.

Table 1: Average accuracy scores

k	s1	s2	s3	s4	s5	s6	s7	s8	ave
4	0.79	0.89	0.78	0.80	0.84	0.84	0.82	0.90	**0.83**
5	0.89	0.97	0.91	0.90	0.97	0.93	0.94	0.96	**0.93**
6	0.97	0.97	0.93	0.90	0.98	0.95	0.94	0.96	**0.95**
7	0.98	0.98	0.95	0.90	0.98	0.93	0.96	0.95	**0.95**
8	0.97	0.98	0.96	0.90	0.98	0.92	0.95	0.95	**0.95**
9	0.98	0.97	0.95	0.91	0.98	0.92	0.95	0.96	**0.95**
10	0.98	0.97	0.94	0.90	0.98	0.91	0.95	0.97	**0.95**

As seen in Table 1, the highest overall accuracy measure achieved with varied word lengths was 95%. This value was found in comparisons where $k = 6$, and was not improved with increasing word lengths. Using this value for k, accuracy scores among individual segments ranged from 90% to 97%. For this reason, we computed distance measures using k-mers of length six. Using this value of k, the computation time per segment was approximately 15 minutes.

2.3 Edge Graphs

From each of the $N \times N$ distance matrices computed for each flu segment, we derived a $N \times N$ edge graph G_{seg}, where $seg = 1, ..., 8$. For each segments sequence in a given sample, $s_{seg,i}, i \in N$, there exists at least one nearest neighbor, $s_{seg,j}$, $j \in N$, $j \neq i$ such that the distance score computed between sequences $s_{seg,i}$ and $s_{seg,j}$ is the minimal, non-diagonal entry in the i^{th} row of the $N \times N$ distance matrix computed for segment seg. This distance score was considered minimal so that any other sequence $s_{seg,k}$, $k \in N$, $k \notin \{i, j\}$ which also showed this distance score from $s_{seg,i}$ was also considered a nearest neighbor. Edges were added between each sequence and its nearest neighbor(s) in edge graphs by setting all nearest neighbor pair entries to 1.

As the notion of similarity is not symmetric, these edges were not bi-directional. Edges here represented transmission events between sequence hosts, and as we did not approximate the direction of transmission, we viewed the edges as undirected. Thus upon adding an edge to position (i, j), an edge was also added at position (j, i).

Table 2 displays the average degree for each segment specific graph. The segments with the lowest degrees were $1, 4, 6$ while the highest degrees were found for segments 7 and 8. A lower degree indicates fewer overall nearest neighbor pairs and a higher level of sequence variability.

A summed $N \times N$ graph G was then created by summing all pairwise edges across all segments:

$$\forall i, j \in N,\ G_{i,j} = \sum_{seg=1}^{8} G_{seg,i,j} \quad (3)$$

This graph G contained the sum of edges between all samples, where edges represented segment specific links. Thus, the maximum number of edges recorded at any position (i, j) was 8. A score of 8 at position (i, j) would indicate that

Table 2: Average degree per segment

segment	*average degree*
1	3.66
2	7.72
3	5.68
4	3.35
5	6.2
6	4.44
7	9.88
8	9.0

sample j was a nearest neighbor to i in all segment specific graphs 1 through 8.

It is of interest to note that a number of sequence pairs exhibited a bi-directional similarity, with the a nearest neighbor of $i = j$ and a nearest neighbor of $j = i$. These sequence pairs demonstrate a much stronger notion of similarity than pairs with only uni-directional minimal distance, and their connection was thus designated by two edges in the (i, j) and (j, i) entry of matrix G. The number of total edges was computed by summing the lower diagonal entries of G.

3. NETWORK CHARACTERISTICS

3.1 Mixing Patterns

Quantifying the mixing patterns of a network is performed by examining edge distributions among different node types of interest. If edges are approximations of transmission among individual hosts, then examining edges across regionally distributed host types may offer insight into more generalized global transmission through larger host populations. Such transmission patterns may be discovered using mixing patterns, which are described in detail in [12]. Mixing patterns describe the frequency of edges between nodes of different types in a network.

Here we examined mixing patterns between the host classes wild avian, domestic avian, human, and swine and across globally distributed regions and regions specific to Southeast Asia. Global regions included North America, South America, Africa, Europe, Asia, Oceania, and the Middle East. Regions within Southeast Asia included China, Indonesia, Vietnam, Thailand, Laos and Malaysia/Singapore. These analyses were accomplished by deriving two-way mixing matrices based on regional and host type classes. Thus, each sample collected from a region of interest and a host type of interest was considered as part of a group described by region and host type. Note that not all 4,228 samples of our dataset were members of these groups. All samples were used in generating over all network characteristics such as the total number or edges and the total number of nodes, but mixing patterns described in the following were only examined between select groups.

For studying mixing patterns among T node types, a $T \times T$ mixing matrix E is used to measure edge numbers between different types. This provides a reduction in dimensionality as there are generally less node types than nodes in a network. This mixing matrix E is then normalized by dividing each entry by the total number of undirected edges in the summed graph G(eq.3). Thus, each matrix entry contains the percentage of the total number of edges in the overall network which are found between nodes of each type being examined.

For each node type, the conditional probability that a node of one type is connected to a node of a different or similar type may then be computed [13]. Assume types a, b, then given a node of type a, the probability that it is directly connected to a node of type b is written as $P(b|a)$. To compute $P(b|a)\ \forall b \in T$ from E, each entry in the a^{th} row is divided by the total percentage of edges connecting to nodes of type a in the overall network (eg.3).

Because data were not evenly distributed among groups, we required a method for examining bias among edge types which took this clustered distribution of sample types in to account. We compared the observed occurrence of inter-type edges to what would be expected in a network of random edge formation. We considered the ratio $p_{ab}/p_a p_b$, where p_{ab} represents the frequency of edges between types a and b, p_a and p_b represent individual frequencies of each node type in the entire network. If p_{ab} is notably greater than $p_a p_b$, a bias may be indicated. Because edges were undirected in this analysis, there was no assumed order of node pairs and $p_{ab} = p_{ba}$.

3.1.1 Global/Host Mixing Patterns

Table 3 lists conditional edge formation between different sample types based on their global region and host type. Table 3 indicates a notable amount of mixing across host types. For example, in the North American region, the P(Wild Avian|Domestic Avian)=34%. Similarly, in the North American samples, P(Domestic Avian|Domestic Avian)=50%. This indicates that given any sequence sampled from a domestic avian species in North America, the probability that that sequence forms and edge with another domestic avian sequence is 50%. The probability that the sequence forms an edge with a wild avian sequence is 34%.

Transmission across global regions as well as host types is represented in Table 3 as well. P(Asia, Domestic Avian| South America, Domestic Avian) is equal to 1.00. This means that all samples collected from domestic avian species in South America formed and edge with domestic avian samples from Asia.

Table 4 shows a list of inter-group mixing patterns with observed/expected edge frequencies greater than 1. Ratios greater than 1 are of interest because they indicate a probability of mixing much greater than expected by chance alone. For example, mixing between human South American strains and human European strains occurs at a rate more than three times expected in a randomly generated graph. Similarly, the wild and domestic avian strains within Europe are five times more likely to be mixing than by chance alone. Most of the edge relationships in this table are between avian and avian or human and human node types which differ in their regions of origin. The exception to this is the higher than expected edge frequency between North American Wild Avian and North American, Human.

3.1.2 Southeast Asian/Host Mixing Patterns

Table 3: **Global Regional/Host Mixing**
$P(b|a)$ values for 2-way mixing among host types and global regions. Only relations with values greater than 0.001 are shown. Host type abbreviations, WA = Wild Avian, H = Human, DA = Domestics Avian, S = Swine. Region abbreviations, NA = North America, SA = South America, AF = Africa, EU = Europe, O = Oceania, ME = Middle East

			b																					
			NA				SA		AF		EU				AS				O			ME		
			WA	H	DA	S	H	DA	WA	DA	WA	H	DA	S	WA	H	DA	S	WA	H	DA	WA	H	DA
a	NA	WA	0.40		0.532	0.002																		
		H		0.894			0.045					0.008				0.048				0.004				
		DA	0.34	0.007	0.5	0.014																		
		S		0.0513	0.065	0.787						0.004						0.012		0.065				
	SA	H	0.646			0.217					0.016				0.12									
		DA															1.00							
	AF	WA	0.25		0.187	0.064					0.187		0.187		0.063		0.063							
		DA								0.947	0.02		0.007		0.002		0.015							0.01
	EU	WA	0.002	0.005				0.005	0.021	0.349	0.002	0.502		0.015		0.053				0.002			0.029	
		H		0.585		0.002	0.08					0.079	0.004			0.24				0.009				
		DA			0.002				0.002	0.002	0.174	0.008	0.739		0.011		0.044		0.001		0.003			0.002
		S												0.961				0.039						
	AS	WA								0.005		0.009		0.333	0.022	0.6	0.013						0.002	
		H		0.444			0.077					0.031			0.002	0.435	0.005			0.007				
		DA						0.002		0.001	0.004		0.009		0.145	0.016	0.773	0.02	0.001		0.001			0.002
		S		0.007	0.003	0.014								0.006	0.017	0.002	0.108	0.727		0.102	0.001			
	O	WA										0.047				0.209				0.721				
		H		0.019		0.002										0.003		0.003		0.972				
		DA	0.04		0.013						0.0132		0.066		0.013		0.132	0.026	0.408		0.289			
	ME	WA																					1.000	
		S		0.825			0.079									0.048				0.016			0.032	
		H								0.006	0.017		0.003		0.003		0.019					0.015		0.939

Table 4: **Global/Host Bias**. Edge bias between types, only values > 1 are listed, (numbers) indicate number of samples of each type.

type a	type b	$p_{ab}/p_a p_b$
Oceania, Domestic Avian (3)	Oceania, Wild Avian (5)	87.50
Middle East, Wild Avian (1)	Middle East, Domestic Avian (54)	11.80
Europe, Human (29)	Asia, Human (232)	5.59
Europea, Wild Avian (30)	Europe, Domestic Avian (82)	5.04
South America, Human (125)	Europe, Human (29)	3.47
South America, Human (125)	Asia, Human (232)	3.19
North America, Human (1609)	South America, Human (125)	2.47
Africa, Wild Avian (2)	Europe, Wild Avian (30)	2.12
North America, Human (1609)	Europe, Human (29)	1.97
South America, Domestic Avian (1)	Asia, Domestic Avian (383)	1.88
Asia, Wild Avian (73)	Asia, Domestic Avian (383)	1.74
North America, Wild Avian (218)	North America, Human (312)	1.60
North America, Human (1609)	Asia, Human (232)	1.43

Table 5: **Southeast Asian Regional/Host Mixing**
$P(b|a)$ values for 2-way mixing among host types and Southeast Asian region. Only relations with values greater than 0.001 are shown. Host type abbreviations, WA = Wild Avian, H = Human, DA = Domestics Avian, S = Swine. Region abbreviations, CH = China, IN = Indonesia, VI = Vietnam, TH = Thailand, LA = Laos, ML/S = Malaysia and Singapore.

			b																		
			CH				IN				VI				TH				LA	ML/S	
			WA	H	DA	S	WA	H	DA	S	WA	H	DA	S	WA	H	DA	S	DA	H	DA
a	CH	WA	0.08	0.015	0.72	0.084							0.015						0.007		
		H		0.209	0.002											0.002				0.003	
		DA	0.077	0.011	0.732	0.054							0.007						0.001		0.002
		S	0.02		0.121	0.619		0.016										0.002			
	IN	WA						0.111	0.889												
		H				0.014	0.002	0.971	0.013												
		DA			0.002		0.026	0.026	0.947												
		S									0.241		0.759								
	VI	WA		0.011	0.011					0.079		0.023	0.854		0.154	0.077	0.385		0.023		
		H			0.077						0.154		0.154		0.154						
		DA	0.003	0.003	0.013					0.015	0.053	0.001	0.894		0.001		0.006		0.004		
		S												0.054	0.243	0.081	0.622				
	TH	WA										0.004	0.004	0.017	0.358	0.036	0.581				
		H		0.076								0.004	0.004	0.011	0.072	0.166	0.181			0.011	
		DA			0.001							0.006	0.01	0.028	0.374	0.059	0.522				
		S				0.011												0.947			
	LA	DA	0.007	0.007	0.01						0.007		0.017						0.921		
	ML/S	H		0.021												0.002				0.089	
		DA			0.444																

Table 6: **Southeast Asian Edge Bias**
Edge bias between types, only values > 1 are listed, (numbers) indicate number of samples of each type.

type a	type b	$p_{ab}/p_a p_b$
Indonesia, Swine (2)	Vietnam, Wild Avian (4)	37.10
Indonesia, Wild Avian (1)	Indonesia, Domestic Avian (23)	29.50
Thailand, Wild Avian (18)	Thailand, Domestic Avian (31)	23.20
Vietnam, Wild Avian (4)	Vietnam, Human (1)	21.20
Vietnam, Swine (2)	Thailand, Domestic Avian (31)	15.70
Vietnam, Wild Avian (4)	Vietnam, Domestic Avian (67)	12.00
Vietnam, Swine (2)	Thailand, Wild Avian (18)	10.60
Indonesia, Swine (2)	Vietnam, Domestic Avian (67)	6.95
Vietnam, Human (1)	Thailand, Domestic Avian (31)	6.83
Thailand, Human (12)	Thailand, Domestic Avian (31)	5.47
Vietnam, Swine (2)	Thailand, Human (12)	5.29
Vietnam, Human (1)	Thailand, Wild Avian (18)	4.71
China, Wild Avian (13)	China, Domestic Avian (143)	4.51
Thailand, Wild Avian (18)	Thailand, Human (12)	3.73
Vietnam, Human (1)	Thailand, Human (12)	3.53
Indonesia, Wild Avian (1)	Indonesia, Human (44)	1.93
China, Wild Avian (13)	China, Swine (41)	1.83
China, Human (71)	Malaysia/Singapore, Human (11)	1.63
Vietnam, Wild Avian (4)	Laos, Domestic Avian (16)	1.32
Vietnam, Human (1)	Vietnam, Domestic Avian (67)	1.26
China, Domestic Avian (143)	Malaysia/Singapore, Domestic Avian (1)	1.18

Table 5 shows conditional edge type probabilities among different host types only within the region of Southeast Asia. These results are similar to those for globally distributed regions in that notable levels of edge formation among different types and localities are found. For example, P(Indonesia, Domestic Avian| Indonesia, Wild Avian) = 90% and P(Thailand, Domestic Avian| Vietnam, Swine) = 62%.

Notable observed to expected frequency ratios of edges among different types are listed in Table 6. Unlike the majority of edges being found among similar host types in global regions (Table 3), this table lists a higher number of inter-host type edges. The highest scoring edge type in this table is between Indonesia Swine and Vietnamese Wild Avian samples. This may reflect a relatively high amount of host jumping in this region.

4. MAXIMAL CLIQUES

We also examined the connectivity structure of the network graph to study inter- and intra-group transmission by looking at the maximal cliques formed by edges. A maximal clique is a maximal subset of vertices such that every pair of vertices is connected by an edge. In the flu transmission graph, a maximal clique represented a set of sequences that were all of minimal distance to one another. Thus a maximal clique of sequences may represent a very closely related set.

Although maximal cliques were relatively small size in our network, ranging from 2 to 58 vertices, a number of mixing patterns were evident. Regional mixing patterns as described by the cliques were limited to human strains. More specifically, a number of maximal cliques exhibited strong inter-group mixing of human strains collected in North America, South America, and Asia. These were the only inter-regional transmission patterns reflected by the maximal cliques. Additional maximal cliques presented very strong intra-regional connections within each of these three regions. Higher than expected levels of transmission between humans across these regions were also reflected in Table 3.

5. DISTANCE THRESHOLDS

One of our goals in this research is to develop a more biologically meaningful notion of transmission by incorporating distance score thresholds. The minimal k-mer count of a sequence across all others may be quite large, and in such cases, it may be more meaningful to assign no neighbors, and thus no transmission, from that sequence. For example, the mean pairwise distance of segment 1 data was $2,260$, and its mean intra-group (Global Region/Host) pairwise distance was $1,290$. Nearest-neighbor distances ranged from 0 to $1,848$, with mean of 59.

To establish a more stringent notion of similarity (transmission), a threshold can easily be instated in the nearest-neighbor definition: the nearest neighbor of node n_i is n_j only if the distance between n_i and n_j is minimal and less than distance threshold d. We are examining the effects of different values of d with larger complex datasets and randomly generated data.

6. CONCLUSION

The tools and methods described here provide a computationally efficient and robust method to derive a graph from similarity scores among whole-genome sequences. Sequence similarity is computed in linear time, which is optimal in cases of very large datasets. A graph is easily generated from these comparisons, and then examined with simple graph theoretic approaches. Many graph based approaches are amenable to analysis of very large data sets.

In our example, graph based analysis enabled us to examine regional influenza transmission patterns among different host types. We considered transmission among host types including wild avian, human, domestic avian, and swine with respect to global regions and with respect to localities in Southeast Asia. We detected a considerable amount of transmission across host types and regions in both cases. Host jumping in Southeast Asia seemed more prevalent when compared to global results. However, this may be an artifact of sampling bias and/or the fact that we did not examine host specific transmission within any other localized regions.

The graph representation allowed us to determine conditional probabilities of edges forming among different node types, as well as look at stronger notions of connectivity using the maximal clique approach. Quantifying these types of connections may form the basis of predictive transmission models. In conjunction with adequate sampling, these methods may prove a useful tool in tracking and managing current and future disease outbreaks of many types.

7. REFERENCES

[1] A. Breland, S. Nasser, K. Schlauch, M. Nicolescu, and F. Harris Jr. Efficient Influenza A Virus Origin Detection. *Journal of Electronics & Computer Science*, 10(2), 2008.

[2] J. P. Chretien, J. C. Gaydos, J. L. Malone, and D. L. Blazes. Global network could avert pandemics. *Nature*, 440(7080):25–26, Mar 2006.

[3] J. De Jong, E. Claas, A. Osterhaus, R. Webster, and W. Lim. A pandemic warning? *Nature*, 389(6651):554, 1997.

[4] E. Holmes. RNA virus genomics: a world of possibilities. 2009.

[5] G. Jenkins, A. Rambaut, O. Pybus, and E. Holmes. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution*, 54(2):156–165, 2002.

[6] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947, 2007.

[7] C. Lee and Y. Saif. Avian influenza virus. *Comparative immunology, microbiology and infectious diseases*, 32(4):301–310, 2009.

[8] P. Lemey, A. Rambaut, A. Drummond, and M. Suchard. Bayesian phylogeography finds its roots. 2009.

[9] M. Margulies, M. Egholm, W. Altman, S. Attiya, J. Bader, L. Bemben, J. Berka, M. Braverman, Y. Chen, Z. Chen, et al. Genome sequencing in

microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[10] M. Metzker. Sequencing technologiesŮthe next generation. *Nature Reviews Genetics*, 11(1):31–46, 2009.

[11] 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*, 4(8):e1000133, Aug 2008.

[12] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[13] M. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126+, February 2003.

[14] D. Pride, R. Meinersmann, T. Wassenaar, and M. Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome research*, 13(2):145, 2003.

[15] A. Rambaut, O. Pybus, M. Nelson, C. Viboud, J. Taubenberger, and E. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619, 2008.

[16] 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*, 320(5874):340–346, Apr 2008.

[17] W. Stallings. *Computer organization and architecture: designing for performance*. Prentice Hall, 2009.

[18] D. C. Torney, C. Burks, D. Davison, and K. M. Sirotkin. *Computers and DNA, SFI Studies in the Sciences of Complexity*. Addison-Wesley Publishing Co., 1990.

[19] 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*, 82(2), Jan 2008.