Efficient Influenza A Virus Origin Detection

Adrienne Breland[1], Sara Nasser[1],
Karen Schlauch[2], Monica Nicolescu[1], Frederick C. Harris, Jr[1]

[1]Dept. of Computer Science and Engineering, University of Nevada,
Reno NV 89557, USA.
{brelanda, sara, monica, fredh}@cse.unr.edu
[2]Dept. of Biochemistry, University of Nevada, Reno NV 89557, USA.
schlauch@unr.edu

ABSTRACT-This research describes a novel, alignment-free method of genomic sequence comparisons based on absent nucleotide words and expression levels. Testing this method on Influenza A virus isolates, three classifications are presented which successfully identify; 1) the geographic origins of domestic bird H5N1 isolates through China and Southeast Asia during 2006, 2) the country of human H5N1 isolates crossing over from domestic bird hosts and, 3) the historical flu season from which human H3N2 isolates originated. Because comparison methods used do not rely on alignment, they are computationally efficient and well suited for large numbers of sequences in compehensive flu transmission network delineation.

Keywords: classification, alignment free, influenza transmission, word absence

1. INTRODUCTION

Half a million human global deaths per year have been attributed to *influenza* A *virus* [6]. In addition, the highly pathogenic H5N1 subtype has been predicted to cause the next worldwide pandemic while it currently compromises live bird trade economics. The 60% mortality rate found in human H5N1 cases is alarming [11] and its symptoms are similar to those found during the Spanish Flu epidemic [15] which caused acute illness in 25-30% of the world population [32]. Recent human infections with avian subtypes H7N7 and H9N2 have been identified as well in China [18]. There is a pressing need to gain a clear understanding of the geographic routes through which H5N1 and other subtypes of *influenza* A *virus* spread and where they originate. Such efforts require a powerful method for building flu isolate associations which incorporate extensive numbers of genomic sequences.

We propose a novel and efficient sequence similarity detection method. Relationships between flu isolates can be determined through a highly computationally efficient genomic sequence comparison approach. Because of this efficiency, a large number of genomic sequences can be compared and distinguished in terms of lineage and place of origin. This offers a significant improvement over current methods available for influenza isolate comparisons which are not well suited for large amounts of data. It is our overall goal to refine an efficient sequence comparisons technique which will allow the full utilization of existing flu sequence data to help answer many questions regarding the epidemic movement and transmission of this virus.

The rest of this paper is structured as follows. Section 2 presents the background and significance of the work. Section 3 describes a novel sequence comparison measure. Section 4 present three classification schemes based on this measure. Section 5 presents our conclusions and future work.

2. BACKGROUND and SIGNIFICANCE

2.1 CURRENT KNOWLEDGE

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 also through North America and Europe to South America [28]. The tools that researchers have to examine isolate relatedness strongly influence the strength and focus of any non-theoretical geospatial flu transmission mapping. In [28], the majority of sequence comparisons were performed antigenically rather than genetically. Only 130 partial genomic sequences were compared in parallel with antigenic comparisons. While antigenic comparisons elucidate 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, nucleotide sequence comparisons allow a finer level of differencing as information representing all point mutations and re-assortments are included in complete genomic sequences. Standard genomic sequence comparison methods based on rooted phylogenetic trees are not designed to encompass such detail. In [22], 900 complete genomes from the Northern and Southern hemispheres were compared and it is stated that even these did not suggest a specific network of viral movement. In similar studies attempting to characterize the movement of flu viruses in China and Southeast Asia [7, 23, 35, 36], India [27], Europe [3,29] and Africa [9], even smaller sample sizes are generally used and compared via antigenic or genomic sequence differences. Consistent in most studies is a call for increased surveillance of influenza and more comprehensive data sets. Current studies are limited by the number of isolates that can be realistically included and interpreted in a phylogeny and the ability to display results in a geographical manner. Their resolution is both blurred by lack of detail and is often incomplete due to the requirement of focusing on a limited number of isolates.

The number of flu genomes made publicly available have increased exponentially over that last decade (Figure1). While a wealth of viral sequence data exists, software which enables researchers to incorporate this data into comprehensive studies is, to our knowledge, lacking. This represents a common lag in the development of software to match the computational needs involved with analyzing newly available data.

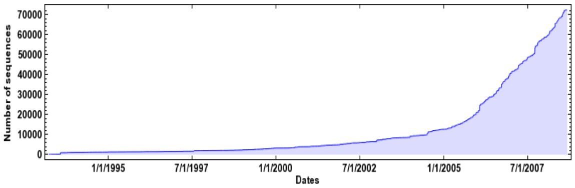


Figure1. Growth of Flu Sequences over past decade (Influenza Resource, website).

This concern has been voiced in publications such as [14] and [37], which call for innovative methods to infer phylogenies and transmission networks respectively. In a recent influenza study conducted by [26], it is stated that there has been "no rigorous measurement of viral diversity across time, across space, and among subtypes" despite data availability. Researchers rely on the use of traditional phylogenetic tree building methods to attempt to elucidate comprehensive geographic transmission networks. Phylogenetic tree building methods and their interpretations are better suited for relatively small numbers of sequences, rather than comprehensive networks spanning multiple continents and species groups. Building a phylogenetic tree requires the derivation of a complete tree structure from isolate relationships so that all isolates can be traced back to a common ancestor. An estimate of the degree of divergence along branches which dictates its final structure is required as well. This presents a computational constraint on the number of sequences which can be included realistically. It is also difficult to asses the accuracy of phylogenetic trees once they have been built as their structures are strongly influenced by both the methods used in preliminary sequence alignments [20] and the tree building algorithms applied afterwards. Even if the phylogenetic tree of large groups of sequences is created, the realistic interpretation of such a tree is a difficult task. For example, in [22] flu sequence dynamics during the US epidemic season 2006-2007 were examined using data from 353 viral isolates. This was described as the largest single season study to date of its kind and could identify no clear spatial patterns of spread.

2.2 RE-ASSORTMENT DETECTION

When quantifying viral sequence similarities, it is optimal to construct comparison methods which are sensitive to the specific genomic characteristics of the virus in question. Influenza genomes remain in eight noncontiguous segments throughout their lifetime. Due to this, they are subject to re-assortment in which the full complement of segments in an isolate may be composed of segment subsets from multiple progenitors. Our approach offers a second improvement over current methods for building influenza isolate relationships which includes the ability to clearly identify re-assortment events. Re-assortment plays an important role in influenza evolution and is quite frequent although not well quantified [26]. The Spanish flu which emerged in 1918 may have resulted from a re-assortment of porcine and human flu segments [12] although this is not a uniformly accepted assertion [32]. Current phylogenetic tree building methods are not conducive to pinpointing exactly when and where re-assortment has occurred as they are limited in isolate numbers and segment distinctions. The sequence comparison method described in this research is sensitive to both re-assortment and point mutations, or antigenic "shift" and "drift". Examining re-assortment requires that each segment in each genome be compared separately rather than treating all segments as part of the same, contiguous sequence. In the recent past, attempts to build influenza phylogenies either concatenated all segment sequences before building trees, or used specific gene segments to represent entire genomes. This obscures the effects of re-assortment. Some more recent approaches have created separate phylogenetic trees for each of the twelve genes encoded in the viral RNA, or for each of the eight genome segments, and then tried to locate inconsistencies which may indicate re-assortment. Through this approach, pinpointing all instances of re-assortment has remained an elusive task, as reported in [17] because it requires trying to find similarities and differences in eight or twelve phlyogenetic trees. Our proposed method for deriving sequence relationships has the potential to treat each segment group separately so that re-assortment is clearly defined in each generational step. Given adequate data, the methods proposed can elucidate the time and place of each re-assortment event in a multi-species, comprehensive data set.

2.3 ALIGNMENT-FREE SEQUENCE COMPARISONS

This paper describes alignment-free sequence comparison methods which optimizes computational time and space. This works to remove the current ceiling on the number of sequences that can be incorporated into viral lineage studies. In the field of bioinformatics, many methods of sequence comparison have been employed which include both alignment-based and alignment-free approaches. Alignment-based methods rely on the preliminary alignment of the sequences in question to enable their comparison. Packages such as ClustalW, Probcons, T-coffee, and DALIGN utilize sequence alignment prior to quantifying sequence differences. Alignment-free methods compare sequences based on global characteristics such as k-nucleotide counts, codon transitional probabilities and complexity measures. A comprehensive review of alignment-free methods is found in [34].

While alignment-based methods are routinely employed in the building of phylogenies, one of the most glaring drawbacks is the computational expense incurred in aligning long and/or large numbers of sequences. To align two sequences using brute force, all possible alignments of the sequence pairs must be tested before the best alignment can be chosen. When multiple sequences are involved, the computational expense grows exponentially as all sequences must be compared against all other sequences in this manner. This presents an NP-hard problem [31], meaning that with large numbers, the problem is essentially unsolvable. This has required that all current alignment-based methods rely on pre-alignment heuristics to reduce the effective number of sequences and sequence lengths to be compared. Often, these heuristics are based on alignment-free statistics. This fact illustrates the additional computation required in alignment-based approaches. The described alignment-free sequence comparison method to allows the processing of several thousand genomic sequences.

3. METHODS

In the following sections, three classifications are presented which successfully identify; 1) the geographic origins of domestic bird H5N1 isolates through China and Southeast Asia during 2006, 2) the country of human H5N1 isolates crossing over from domestic bird hosts and, 3) the historical flu season from which human H3N2 isolates originated. The methods used for genomic sequence comparisons do not rely on alignment and are computationally efficient. Methods and results regarding classifications are presented in the following. All influenza sequences were obtained through the Influenza Virus Resource which gives access to sequences collected by the National Institute of Allergy and Infectious Disease (NIAID) Genome Sequencing Project and those available through GenBank.

3.1 OLIGONUCLEOTIDE SIGNATURES

If genomic sequences were randomly organized, most short nucleotide words would have an equal probability of being found within any given sequence of sufficient length. The study of short word frequencies has shown a biased distribution of words which deviates from random, with some words over- and some under- represented to differing levels [4, 5, 16]. A genomic signature derives frequency patterns by calculating the over- and under- representation of specific base pair sequences when compared to random expectations. Genomic signatures of short word lengths are similar for organisms within kingdom groupings [5] and are sometimes consistent enough to be used in the regrouping of mixed fragments from multiple species genomes [1, 21, 33].

3.2 WORD ABSENCE

Some words have been found to be commonly absent from species groups [13] and have been referred to as nullomers and primes. While the reason as to why certain words are absent and others present in particular genomes is most likely complex, the inheritance of absent words has been examined on a broad evolutionary scale. It has been proposed that word absence is an inherited characteristic through the observation that human and chimp DNA contain 28 common absent words of length 11 and 14 absent words which differ by only one nucleotide [2]. It could also be expected that word absence is inherited by the immediate progeny of microbial samples in a micro-evolutionary sense. Absent words are an integral part of any genomic sequence as much as present words, and by inheriting a nucleotide sequence, or a close derivative of it, a microbial offspring should also inherit many of the words absent from that sequence as well. This may offer the delineation of closely related microbes, including viral pathogens. Researchers in [10] found word absence/presence to show more correlation between genomes within the same species than between genomes of different species. Even so, less correlation was found between same species genomes than was statistically expected, and it was suggested that word absences may offer delineation within species groups as well.

This prediction has been reinforced by our examinations of words absences across related influenza sequences. Cross genome word absence patterns were examined by reclassifying hexa-nucleotide word frequencies into a binary format so that presence was indicated by a one and absence by a zero. Human H3N2 influenza sequences from three distinct epidemics were clearly differentiated in this manner. Words which were distinctly absent or present for one of three epidemics accounted for 82% of all words in signature sets. The full absence/presence table is listed in Table 1 and the dataset description used for this table is provided in Section 4.

3.3 MINIMAL MARKOV MODEL

In genomic word expression analysis, Markov Models are often used as a means of calculating the expected count of each word (E(w)) in a signature set [19, 24, 30]. In Markov chains, the current state of a system is predicted by its previous states. In word signature analysis, this translates to predicting a word frequency based on the observed frequencies of its sub words. Depending on the order of the Markov Model, bias contributed to a word of length *m* from sub words of lengths *1*,…, *m-1* can be

Table 1: Uniquely (0)absent/(1)present word counts.

Inter-epidemic								
Hong Kong 1980		Nicaragua 2007			New South Wales 1999			
s1	s2	s3	s4	s5	s6	s7	s8	# words
0	0	1	1	1	1	1	1	123
1	1	0	0	0	0	0	0	109
1	1	0	0	0	1	1	1	93
0	0	1	1	1	0	0	0	71
1	1	1	1	1	0	0	0	32
0	0	0	0	0	1	1	1	19
1	1	1	1	1	0	0	1	14
0	0	0	0	0	0	0	1	13
0	0	0	0	0	1	1	0	11
0	0	1	1	1	0	0	1	11
0	0	1	1	1	1	1	0	10
1	1	0	0	0	1	1	0	9
0	0	0	1	0	0	0	0	5
1	1	0	0	0	0	0	1	5
1	1	0	0	1	1	1	1	3
0	0	0	0	1	0	0	0	2
0	0	1	1	0	1	1	1	2
1	1	1	0	1	1	1	1	2
1	1	1	1	0	1	1	1	2
0	1	0	0	0	0	0	0	1
0	1	1	1	1	0	0	0	1
1	0	1	1	1	0	0	0	1
1	1	0	0	1	0	0	0	1
1	1	0	1	0	0	0	0	1
1	1	1	0	0	1	1	1	1
1	1	1	1	0	0	0	0	1
1	1	1	1	1	1	0	0	1
						total		544

removed. As an illustration, assume a sequence is dominated by di-nucleotides TA and AG. Unless specifically selected against, TAA and AAG, which are the concatenations of the dominant sub words will naturally show high frequencies as well. Separating the degree of selection for or against exactly these tri-nucleotides from the contributions of their sub words may require the removal of their di-nucleotide sub word frequency bias. With the ultimate goal to match genomic internal word selection mechanisms, the optimal order of the Markov Model to use remains undetermined. Consistent with findings in [24] and [25], minimal order Markov Models allowed the most differentiation between genomic signatures of different prokaryotic species and were thus used to calculate expected values for signature calculations in this research. A minimal order Markov Model does not remove bias from sub words longer than one character. The expected count of a word E(w), in a genomic sequence of length N is expressed as:

$$\mathbf{E(w) = [(A^{a}*C^{c}*G^{g} *T^{t})* N]}$$

A, **C**, **T** and **G** represent specific nucleotide frequencies in the total sequence N and a, c, t, g are the number of each nucleotide in a word w. As described in [24], the ratio of the observed word count over its expected count, O(w)/E(w) can be used to derive the degree of over- or under-representation of each word in a signature set. This ratio is later referred to as an expression level.

3.4 GENOME COMPARISON METHOD

In our method the features, or defining characteristics used in genome comparisons are based on *k*-nucleotide signature absence subsets. More specifically, the expression level of any nucleotide word of length *k* which is absent from at least one genome in question is considered in genome difference calculations. A measure for comparing two sequences based on word expression levels constrained by word absence was derived. To compare two sequences, s1 and s2, let AP be the set of all words present in only one sequence so that for all w ϵ AP, O(ws1)/E(ws1)>0 and O(ws2) = 0, or O(ws2)/E(ws2)>0 and O(ws1) = 0. AA is the set of all words absent from both sequences so that for all w ϵ AA, O(ws1) = O(ws2) = 0. |AP| denotes the total number of words in AP and |AA| denotes the total number of words in AA. The difference between s1 and s2 is calculated as:

$$\frac{\sum_{w}^{AP} | O(w_{s1})/E(w_{s1}) \quad - \quad O(w_{s2})/E(w_{s2}) |}{| AP | \quad + \quad | AA |}$$

Thus the difference between two sequences is the sum of the observed to expected ratios for words which are absent from *exactly* one sequence divided by the total number of words absent from *at least* one sequence. This allows the comparison of only words which exhibit some degree of absence. In contrast to comparing only two sequences, if comparing relative similarities between a group of sequences, removing |AA| from the equation allows a higher degree of distinction between all pairs. This is because words absent from all sequences offer no inter-sequence differentiation and do not contribute to the derivation of relative differences. All ensuing classifications hinged on this differencing measure to compare genomes.

Sections 4.1 and 4.2 describe supervised classifications. In supervised classifications, test data is comprised of “unknown” data from samples which must be classified. Training data denotes “known” data which will be used as class identifiers. The supervised algorithm compares each test data sample to each class identifier, and assigns it to which ever class it is closest to.

In Section 4.3, an unsupervised classification is described. In unsupervised classifications, data are not subdivided into test and training groups and each sample is assigned to the next closest sample in the data set.

4. RESULTS

4.1 AVIAN H5N1 GEOGRAPHIC ORIGIN DETECTION

This classification tests the proposed method in its ability to discriminate between individual avian H5N1 strain lineages and thereby determine their geographic origins. For this application, 94% accuracy is achieved in assigning all genomes to their correct place of origin. A word length of seven allows the highest degree of accuracy. Isolates from domestic bird (chicken, duck, turkey, goose) outbreaks in China, Africa, Thailand, and Vietnam from 2004-2006 are used for this classification. Regions represented in the China dataset include Guanxi province, Hunan province, Guandong province, and HongKong. Countries represented in African dataset include Afganistan, Nigeria and Sudan. Specific regions of genomes collected in Thailand and Vietnam were not indicated on the

Influenza Resource website and are thus referred to by their country name. Training genomes were not chosen randomly, but instead selected in attempts to represent major avian H5N1 outbreaks in each region of interest during specified time periods and included 17 genomes from domestic birds. Test genomes include 61 domestic avian H5N1 genomes from time periods corresponding with training genomes in the same regions. Out of 61 genomes, 57 are accurately assigned to their exact collection region.

All misclassified genomes are assigned to regions directly adjacent to their true collection origins. This coupled with the high degree of accuracy achieved with all other samples suggest that "missclassifications" may indicate border crossing relationships rather than errors in the classification. Our classification associates one genome from Vietnam to China's Guanxi province. The Guanxi province lies north of the Vietnamese border. One genome from Thailand is matched to genomes originating in Vietnam. These two countries are also directly adjacent. Within the Chinese genomes, one genome from the Hunan province is assigned to the Guanxi province while one genome from Shantou is assigned to the Hunan province. Hunan is next to Guanxi, and the Guandong province which contains Shantou is next to the Hunan province. Border crossing relationships are shown in Figure 2. Though samples are limited, this image may indicate that transmission moves towards the coastal regions of Guanxi and Guandong, perhaps following movement towards port regions. Detailed classification results are given in Table 2. Misclassified genomes are indicated by bold lettering and an asterisk.

4.2 AVIAN TO HUMAN H5N1 TRANSMISSION

This classification achieves a cross host species matching of individual strain lineages between avian and human cases. Human cases are assigned to their closest bird counterparts to determine whether the proposed method can point to the location of viral crossover from bird to human. This classification is 100% accurate in matching 18 human cases of H5N1 to chicken cases in their correct countries. A word length of eight is used. Human host H5N1 genomes from Indonesia in 2005, Thailand in 2004, and Vietnam in 2004 are used as test data. The training data set includes all H5N1 strains from domestic bird hosts in Indonesia, Thailand and Vietnam during the same years of 2005, 2004 and 2004 respectively. Accuracy at a higher spatial resolution than country can not be assessed due to a lack of data information. Classification results are given in Table 3.

Table 2: Classification results for 61 domestic bird H5N1 isolates, g = goose, d = duck, t = turkey, c = chicken.

Isolate sequence	Classified As..
China/Guangxi /1898/g	China/Guangxi/150/g
China/Guangxi /224/g	China/Guangxi/150/g
China/Guangxi /32/g	**China/Hunan/856/d ***
China/Guangxi /582/g	China/Guangxi/150/d
China/Guangxi /288/d	China/Guangxi/150/g
China/Guangxi /1830/d	China/Guangxi/150/g
China/Guangxi /2143/d	China/Guangxi/150/g
China/Guangxi /392/d	China/Guangxi/150/g
China/Guangxi /744/d	China/Guangxi/150/d
China/Guangxi /804/d	China/Guangxi/150/d
China/Guangxi /89/d	**Vietnam/10/c ***
Hong Kong /947/d	Hong Kong/282/c
China/Hunan /988/d	China/Hunan/856/d
China/Hunan /324/d	China/Hunan/856/d
China/Hunan /344/d	China/Hunan/856/d
China/Shantou /3295/d	China/Shantou/1233/c
China/Shantou /3265/d	**China/Hunan/856/d ***
China/Shantou /3840/d	China/Shantou/1233/c
China/Shantou /3923/d	China/Shantou/1233/c
Indonesia /175H/c	Indonesia/CDC25/c
Indonesia /PA/c	Indonesia/CDC25/c
Indonesia /Dairi/BPPVI/c	Indonesia/CDC25/c
Indonesi/Deli Serdang/BPPVI/c	Indonesia/CDC25/c
Indonesia/unung Kidal/BPPW/c	Indonesia/CDC25/c
Indonesia /Magetan/BPPW/c	Indonesia/CDC25/c
Indonesia /Parepare/BPPVM/c	Indonesia/CDC25/c
Indonesia /Purworejo/BPPW/c	Indonesia/CDC25/c
Indonesia imalanggang/BPPVI/c	Indonesia/CDC25/c
Indonesia /Tarutung/BPPVI/c	Indonesia/CDC25/c
Indonesia/TebingTinggi/BPPVI/c	Indonesia/CDC25/c
Ivory Coast/4372-3/d	Ivory Coast/4372-2/t
Ivory Coast/4372-4/d	Ivory Coast/4372-2/t
Nigeria/1047-34/d	Nigeria/1047-30/c
Nigeria/1047-54/d	Nigeria/1047-30/c
Nigeria/1047-62/d	Nigeria/1047-30/c
Nigeria/1047-8/d	Nigeria/1047-30/c

Nigeria/SO452/d	Nigeria/SO300/c
Nigeria/SO493/c	Nigeria/SO300/c
Nigeria/SO494/c	Nigeria/SO300/c
Sudan/1784-7/c	Sudan/2115-10/c
Sudan/1784-8/c	Sudan/2115-10/c
Sudan/2115-12/c	Sudan/2115-10/c
Sudan/2115-9/c	Sudan/1784-10/c
Thailand /Ayutthaya/CU23/c	Thailand/39692/c
Thailand/Kanchanburi/CK-160/c	Thailand/Nontaburi/CK-162/c
Thailand/Nakom Patom/CUK2/c	Thailand/39692/c
Vietnam/5/c	Vietnam/10/c
Vietnam/36/c	Vietnam/10/c
Vietnam/38/c	Vietnam/10/c
Vietnam/C57/c	**Thailand/39692/c ***
Vietnam/LD080/c	Vietnam/10/c
Vietnam/TG023/c	Vietnam/10/c
Vietnam/TN025/c	Vietnam/10/c
Vietnam /2/c	Vietnam/10/c
Vietnam/5/c	Vietnam/10/c
Vietnam/8/c	Vietnam/10/c
Vietnam /9/c	Vietnam/10/c

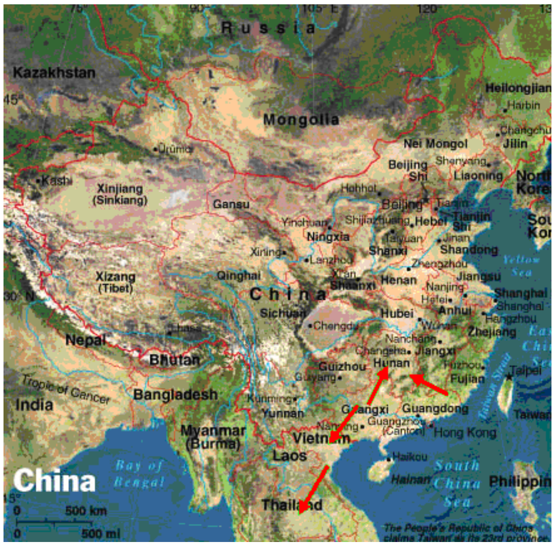


Figure 2: Arrows point from collection region topotential origin.

4.3 EPIDEMIC DISCRIMINATION

The discrimination of inter and intra-epidemic human host H3N2 isolates through unsupervised classifications is also examined. Inter-epidemic sequences are well delineated by difference measures into epidemic specific groups. Intra-epidemic sequences are not consistently well separable in terms of their geographic origins, but show similarities across geographic regions. Inter-epidemic data include eight strains of human host H3N2 representing three distinct epidemics. Two strains were from Hong Kong in 1980, three strains from Managua, Nicaragua in 2007, and three from New South Wales in 1999. Intra-epidemic data include nine human host H3N2 isolates collected in the United States within a three month period during the 2007 flu season. Three are from New York collected between 3/5-3/6. Three isolates were collected in Colorado all on 1/8, and three are from Vermont collected between 2/27-3/1. Identifier strings are listed in Table 4.

Table 3: Classification results, from human host to chicken host, h=human, c = chicken.

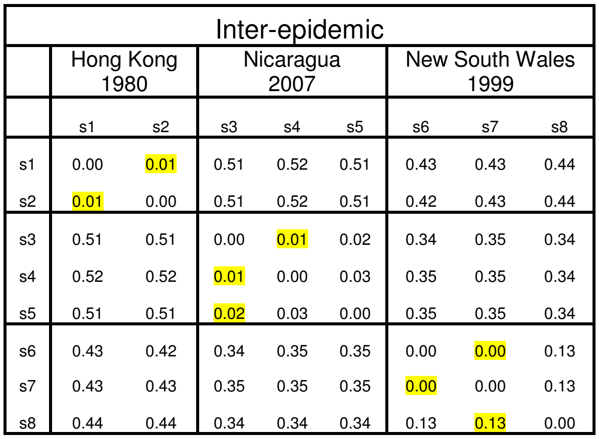
Isolate sequence	Classified As..
Indonesia/5/h	Indo/Magetan/BPPW/c
Indonesia/7/h	Indo/Magetan/BPPW/c
Indonesia/175H/h	Indo/Gunung Kidal/BPPW/c
Indonesia/239H/h	Indo/Parepare/BPPVM/c
Indonesia/245H/h	Indo/Parepare/BPPVM/c
Indonesia/CDC7/h	Indo/Magetan/BPPW/c
Indonesia/CDC184/h	Indo/Magetan/BPPW/c
Indonesia/CDC287/h	Indo/Parepare/BPPVM/c
Indonesia/CDC292T/h	Indo/Parepare/BPPVM/c
Thailand/1(KAN-1)/h	Thailand/9/c
Thailand/2(SP-33)/h	Thailand/9/c
Thailand/5(KK-494)/h	Thailand/9/c
Thailand/16/h	Thailand/9/c
Thailand/SP83/h	Thailand/9/c
Vietnam/1194/h	Vietnam/TN025/c
Vietnam/1203/h	Vietnam/35/c
Vietnam/3062/h	Vietnam/35/c
Vietnam/CL26/h	Vietnam/35/c

Difference metrics for the inter-and intra-epidemic sqequences are shown in Tables 5 and 6. Table 5 shows all sqLences having minimal difference measures with sequences within their respective epidemic groups. These values are highlighted in yellow. For example, s1 is least different from s2, and these two sequences are both members of the Hong Kong 1980 epidemic. This table suggests that samples from distantly related epidemics can be accurately delineated using the proposed measure. In Table 6, samples are not unanimously discriminated based on their geographic location. In this case, only four out of nine samples s4, s5, s7 and s9 show the lowest difference values from same state samples.

Table 4 : Isolate identifier strings.

Inter-epidemic sequences	**Intra-epidemic sequences**
A/Hong Kong/46/1980	A/NewYork/UR06-0510/2007
A/Hong Kong/45/1980	A/NewYork/UR06-0515/2007
A/Managua/14/2007	A/NewYork/UR06-0529/2007
A/Managua/15/2007	A/Vermont/UR06-0469/2007
A/Managua/16/2007	A/Vermont/UR06-0470/2007
A/New S. Wales/20/1999	A/Vermont/UR06-0471/2007
A/New S. Wales/21/1999	A/Colorado/UR06-022/2007
A/New S. Wales/22/1999	A/Colorado/UR06-023/2007
	A/Colorado/UR06-024/2007

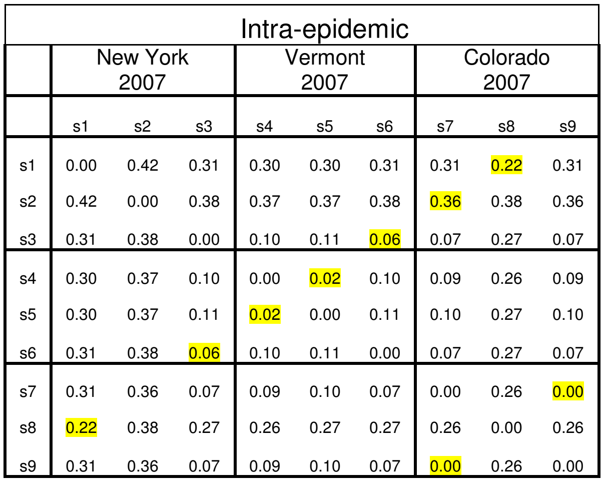
Table 5: Inter-epidemic isolate difference matrix.



	Inter-epidemic							
	Hong Kong 1980		Nicaragua 2007			New South Wales 1999		
	s1	s2	s3	s4	s5	s6	s7	s8
s1	0.00	0.01	0.51	0.52	0.51	0.43	0.43	0.44
s2	0.01	0.00	0.51	0.52	0.51	0.42	0.43	0.44
s3	0.51	0.51	0.00	0.01	0.02	0.34	0.35	0.34
s4	0.52	0.52	0.01	0.00	0.03	0.35	0.35	0.34
s5	0.51	0.51	0.02	0.03	0.00	0.35	0.35	0.34
s6	0.43	0.42	0.34	0.35	0.35	0.00	0.00	0.13
s7	0.43	0.43	0.35	0.35	0.35	0.00	0.00	0.13
s8	0.44	0.44	0.34	0.34	0.34	0.13	0.13	0.00

It is to be expected that intra-epidemic sequences, particularly within a well traveled country such the United States, be highly related. Similarly, it is not surprising that sequences from geographically and temporally distant epidemics show more differences. The clear distinction between sequences from distant epidemics is slightly more surprising and encouraging.

Table 6: Intra-epidemic isolate difference matrix.



	Intra-epidemic								
	New York 2007			Vermont 2007			Colorado 2007		
	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	0.00	0.42	0.31	0.30	0.30	0.31	0.31	0.22	0.31
s2	0.42	0.00	0.38	0.37	0.37	0.38	0.36	0.38	0.36
s3	0.31	0.38	0.00	0.10	0.11	0.06	0.07	0.27	0.07
s4	0.30	0.37	0.10	0.00	0.02	0.10	0.09	0.26	0.09
s5	0.30	0.37	0.11	0.02	0.00	0.11	0.10	0.27	0.10
s6	0.31	0.38	0.06	0.10	0.11	0.00	0.07	0.27	0.07
s7	0.31	0.36	0.07	0.09	0.10	0.07	0.00	0.26	0.00
s8	0.22	0.38	0.27	0.26	0.27	0.27	0.26	0.00	0.26
s9	0.31	0.36	0.07	0.09	0.10	0.07	0.00	0.26	0.00

5. CONCLUSIONS/FUTURE WORK

In sumary, research presented here give examples of k-nucleotide signature subsets enabling the detection of the geographic origins of influenza A viral isolates in supervised and unsupervised classifications. These subsets only include nucleotide words which are absent from at least one genome in question and present in another while multiple genomes may be included in a classification schema. Although difference measures have only been derived for a small number of samples, they are suggestive a highly detailed and quantifiable network among Influenza viral isolates. Large flu transmission networks may become latent through a similar classification scheme which points each isolate to its most similar temporal predecessor. In addition, the described method does not rely on sequence alignment which is computationally expensive. Instead, the differences between closely related genomes are extracted in a relatively inexpensive manner.

Future work will involve the optimization of this comparison method with regards to accuracy and efficiency. We would then like to compare large numbers of sequences to estimate flu networks over large geographic regions. This could prove useful for epidemiological studies, particularly with regards to understanding global transmission networks and the origins of new flu strains.

References

[1] Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T and Ikemura T. (2002) A Novel Bioinformatic Strategy for Unveiling Hidden Genome Signatures of Eukaryotes: Self-Organizing Map of Oligonucleotide Frequency. *Genome Informatics* , 12:12-20.

[2] Acquisti C, Poste G, Curtiss D, Kumar S.(2007) Nullomers: really a matter of natural selection? *PLoS ONE*, 2:e1022.

[3] Bragstad K, Jorgensen P, Handberg K, Hammer AS, Kabell S and Fomsgaard A. (2007) First introduction of highly pathogenic H5N1 avian influenza A viruses in wild and domestic birds in Denmark, Northern Europe. *Virology Journal*, 4:43.

[4] Burge C, Campbell A and Karlin S.(1991) Over- and Under-Representation of Short Oligonucleotides in DNA Sequences. *Proceedings of the National Academy of Sciences of the United States of America,* 89:1358-1362.

[5] Campbell, A., Mrazek J and Karlin S. (1999) Genome signature comparisons among prokaryote, plasmid and mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 96:9184-9.

[6] Cantey JR. (2008) Pandemic Influenza:who will care?. *J S C Med Assoc*., 104:149-51.

[7] Cauthen, AN., Swayne DE, Schultz-Cherry S, Perdue MI and Suarez DL.(2000). Continued Circulation in China of Highly Pathogenic Avian Influenza Viruses Encoding the Hemagglutinin Gene Associated with the 1997 H5N1 Outbreak in Poultry and Humans. *Journal of Virology*, 74:6592-6599.

[8] Duan, L, Campitelli L, Fan XH, et. al. (2007) Characterization of Low-Pathogenic H5 Subtype Influenza Viruses from Eurasia: Implications for the Origin of Highly Pathogenic H5N1 Viruses. *Journal of Virology*, 81:7529-7539.

[9] Fasina FO, Bisschop SP, Joannis TM, Lombin LH, Abolnik C.(2008) Molecular characterization and epidemiology of the highly pathogenic avian influenza H5N1 in Nigeria. *Epidemiol Infect*., Jul 17:1-8.

[10] Fofanov Y et al. (2004) How independent are the appearances of n-*mers* in different genomes? *Bioinformatics* , 20:2421-2428.

[11] Gambotto A, Barratt-Boyes SM, de Jong MD, Neumann G, Kawaoka Y.(2008) Human infection with highly pathogenic H5N1 influenza virus. *Lancet,* 371:1464-75.

[12] Gibbs MJ and Gibbs AJ.(2006) Molecular virology: was the 1918 pandemic caused by a bird flu? *Nature,*Apr 27;440(7088).

[13] Hampikian G and Anderson T. (2007) Absent Sequences: Nullomers and Primes, Pacific Symposium on Biocomputing, 12: 355-366.

[14] Holmes EC. (2008) Evolutionary History and Phlyogeography of Human Viruses. *The Annual Review of Microbiology,* 62:307-28.

[15] Hsieh YC, Wu TZ, Liu DP, Shao PL, Chang LY, Lu CY, Lee CY, Huang FY, Huang LM.(2006) Influenza pandemics: past, present and future. *J Formos Med Assoc*., 105:1-6.

[16] Karlin S and Burge C. (1995) Dinucleotide relative abundance extremes:a genomic signature. *Trends in Genetics,* 11:283-90.

[17] Lam TT-Y, Hon C-C, Pybus OG, Kosakovsky Pons SL, Wong RT-Y, et al. (2008) Evolutionary and Transmision Dynamics of Reassortant Influenza Virus in Indonesia. *PLoS Pathogen* , 4(8):e1000130.

[18] Lee CW and Saif YM(2008) Avian influenza virus. *Comp. Immunol. Microbiol. Infect. Dis*., May 3. [Epub ahead of print].

[19] Leung MY, Marsh GM, and Speed TP. (1996) Under- and overrepresentation of short DNA words in Herpesvirus Genomes. *Journal of Computational Biology*, 3: 345-360.

[20] Lunter G, Miklós I,Drummond A, Jensen J, Hein J.(2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6:83 doi:10.1186/ 1471-2105-6-83

[21] McHardy A, Martin H, Tsirigos A, Hugenholtz P and Rigoutsos I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods,* 4: 63-72.

[22] Nelson MI, Edelman L, Spiro DJ, Boyne AR, Bera J, et. al. (2008) Molecular Epidemiology of A/H3N2 and A/H1N1 Influenza Virus during a Single Epidemic Season in the United States. *PLoS Pathogen,* 4:e1000133. doi:10.1371/journal.ppat.1000133

[23] Nguyen DC, Uyeki TM, Jadhao S et al. (2004) Isolation and Characterization of Avian Influenza Viruses, Including Highly Pathogenic

H5N1, from Poultry in Live Bird Markets in Hanoi, Vietnam, in 2001. *Jounal of Virology*, 79: 4201-4212.

[24] Pride D, Meinersmann R, Wassenaar T and Blaser M. (2003) Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. *Genome Research* , 13:145-58.

[25] Rainer M, Kroger M, Rice P and Joachim-Fritz H. (1992) Statistical Evaluation and biological interpretation of non-random abundances in the E.coli K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair. *Nucleic Acids Research.*, 20:1657- 1662.

[26] Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. (2008) The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453:615-9.

[27] Ray K, Potdar VA, Cherian SS, Pawar SD, Jadhav SM, Waregaonkar SR, Joshi AA, Mishra AC. (2008) Characterization of the complete genome of influenza A (H5N1) virus isolated during the 2006 outbreak in poultry in India. *Virus Genes*, Jan 24 [Epub ahead of print].

[28] Russell Colin. (2008) The Global Circulation of Seasonal Influenza A (H3N2) Viruses. *Science,* 320: 340-346.

[29] Salzberg, Steven L., Carl Kingsford, et al. (2007) Genome Analysis Linking Recent European and African Influenza(H5N1) Viruses. *Emerging Infectious Diseases*, 13:713-718.

[30] Schbath, S, Prum, B and deTurckheim E. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *Journal of Computational Biology*, 2:417-437.

[31] Sze SH, Lu Y, Yang Q. (2006) A polynomial time solvable formulation of multiple sequence alignment. *Journal of Computational Biology,* 13:309-19.

[32] Taubenberger JK. (2006) The origin and virulence of the 1918 "Spanish" influenza virus. *Proceedings of the American Philosophical Society,* 150:86-112.

[33] Teeling H, Meyerdierks A, Bauer M, Amann R and Glockner F. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology,* 6:938-947.

[34] Vinga S, Almeida J. (2003). Alignment-free sequence comparison—a review. *Bioinformatics*, 4:513-523.

[35] Wang J, Vijaykrishna D, Duan L, Bahl J, Zhang JX, Webster RG, Peiris JS, Chen H, Smith GJ, Guan Y. (2008). Identification of the progenitors of Indonesia and Vietnam avian influenza A (H5N1) viruses from southern China. *Journal of Virology*, Jan 23 [Epub ahead of print].

[36] Webster R G., Guan Y, Peiris M et. al. (2002). Characterization of H5N1 Influenza Viruses That Continue To Circulate in Geese in Southeastern China. *Journal of Virology,* 76:118-126.

[37] Welch D, Nicholls GK, Rodrigo A, Solomon W. (2005) Integrating genealogy and epidemiology: the ancestral infection and selection graph as a model for reconstructing host virus histories. *Theoretical Population Biology,* 68:65-75.

Ms. Adrienne Breland is currently a Ph.D. Candidate in the Department of Computer Science and Engineering at the the University of Nevada, Reno, USA. She received her MS in Computer Science in 2008, and her MS in Environmental Resource Sciences in 2003, and her BS in Ecology in 1998. She is a member of the ISCA. Her research interests are in Bioinformatics, Computational Algorithms and solving complex biological problems with computation.

Dr. Sara Nasser is currently a Post Doctoral Reasearcher at TGen in Arizona, USA. She received her MS in Computer Science in 2005 and her Ph.D. in Computer Science and Engineering in 2008 from the University of Nevada, Reno. Her research interests are in Bioinformatics, Clasification Algorithms, and Computational Algorithms.

Dr Karen Schlauch is currently an Associate Professor and Director of Bioinformatics in the Department of Biochemistry at the the University of Nevada, Reno, USA. She received her MS and Ph.D. in Mathematics from New Mexico State University in 1994 and 1998 respectively. She has been on the research staff an faculty at the National Center for Genome Resources, Virginia Bioinformatics Institute, and Boston University. She is a a member of the AMS, AMA, and the AWM. Her research interests are in Bioinformatics and Computational Algorithms.

Dr Monica Nicolescu is currently an Associate Professor in the Department of Computer Science and Engineering and Director of the UNR Robotics Research Lab at the the University of Nevada, Reno, USA. She received her MS and Ph.D. in Computer Science from the University of Southern California in 1999 and 2003 respectively. She is a member of AAAI and the IEEE. Prof Nicolescu's research interests are in the areas of human-robot interaction, robot control and learning, and multi-robot systems.

Dr Frederick C Harris, Jr. is currently a Professor in the Department of Computer Science and Engineering and the Director of the High Performance Computation and Visualization Lab at the the University of Nevada, Reno, USA. He received his BS and MS in Mathematics and Educational Administration from Bob Jones University in 1986 and 1988 respectively, his MS and Ph.D. in Computer Science from Clemson University in 1991 and 1994 respectively. He is a member of ACM, IEEE, and ISCA. His research interests are in Parallel Computation, Graphics and Virtual Reality, and Bioinformatics.