Cladistics
The International Journal of the Willi Hennig Society

Tracking the Geographic Spread of Avian Influenza (H5N1) with Multiple Phylogenetic Trees

Journal:	*Cladistics*
Manuscript ID:	CLA-09-05-0268.R1
Manuscript Type:	Article
Keywords:	Biogeography < Applications, Phylogeography < Biogeography < Applications, Cladistics, Molecular systematics < Genomics, Microbiology, Taxon sampling < Methodology

Tracking the Geographic Spread of Avian Influenza (H5N1) with Multiple Phylogenetic Trees

Rasmus Hovmöller [1,2], Boyan Alexandrov [2], Jori Hardman [2], and Daniel Janies [2]

1 Mathematical Biosciences Institute, The Ohio State University, Columbus, OH, USA
2 Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA.

Abstract

Avian influenza (H5N1) has been of great social and economic importance since it first infected humans in Hong Kong in 1997. A highly pathogenic strain has spread from China and has killed humans in east Asia, west Africa, south Asia, and the Middle East. Recently, several molecular phylogenetic studies have focused on the relationships of various clades of H5N1 and their spread over time, space, and various hosts. These studies examining the geographic spread of H5N1 have based their conclusions on a single tree. This tree often results from the analysis of the genomic segment coding for hemagglutinin (HA) or neuraminidase (NA) proteins and a limited sample of viral isolates. Here we present the first study using multiple candidate trees to estimate geographic transmission routes of H5N1. In addition, we use all high quality HA and NA sequences available to the public as of June 2008. We estimated geographic transmission routes of H5N1 by optimizing multistate characters with states representing different geographical regions over a pool of presumed minimum length trees. We also developed means to visualize our results in Keyhole Markup Language (KML) for virtual globes. We provide these methods as a web application entitled "Routemap" (http://routemap.osu.edu). The resulting visualizations are akin to airline route maps but they depict the routes of spread of viral lineages. We compare our results with the results of previous studies. We focus on the sensitivity of results to sampling of tree space, character coding schemes, optimization methods, and taxon sampling. In conclusion, we

find that using one tree and a single character optimization method will ignore many of the transmission routes indicated by genetic sequence and geographic data.

Introduction

Influenza viruses and pathogenic H5N1

Influenza viruses cause disease in humans, wild and domesticated birds and other homeothermic animals. The current classification divides influenza into three major groups: A, B, and C. Within influenza A, lineages are classified with a system of antibodies to the surface proteins of the virions, hemagglutinin and neuraminidase. This leads to the HXNY nomenclature where X includes 16 antigenic subtypes and Y represents nine different antigenic subtypes. Influenza B is only known to infect humans and pinnipeds (seals, sea lions and walruses). Influenza C is only known to infect humans and swine. On the other hand, influenza A, including subtype H5N1, has a much wider host range. Influenza A infects aves as wells as many mammalian groups such as artiodactyl (swine and bovids), equids (horses), carnivorans (canids, domestic and wild felids, pinnipeds, mustelids, and viverrids). H5N1 has also been found in blow flies (Diptera: Sarcophagidae) in the vicinity of farms containing infected poultry in Japan (Sawabe et al, 2004).

Influenza viruses have two means of creating genetic diversity. The segmented genome of influenza viruses enables genomic reassortment between strains. When a host is simultaneously infected by more that one strain of influenza virions that carry gene segments from two or more original viruses, termed "reassortants" can be produced. In addition, influenza viruses are negative strand-RNA viruses. These viruses use RNA

polymerases without proofreading functions, resulting in a high mutation rate (Holland et al., 1982). Both of these molecular evolutionary processes allow many lineages of influenza to escape host immune responses (Smith et al., 2004) and antiviral therapeutics (Bright, et al., 2005; Hill et al., 2009).

The origins of H5N1 and its variable pathogenicity

The RNA for hemagglutinin (HA) encodes a polyprotein with two subunits that must be cleaved by host endoproteases to enable the virion to fuse with the host's cell membrane. Highly pathogenic strains of H5N1 are characterized by multiple basic amino acids at the cleavage site between the two subunits of the hemagglutinin protein (Kawaoka, 1987; Subbarao, 1998). Pathogenicity is a multigenic trait that is also mediated by features such as whether the polymerase basic 2 F2 protein is produced or not (Zamarin et al., 2006) The earliest known influenza strain characterized as serotype H5N1 was isolated from birds in Scotland in 1959 (WHO, 2004) and sequenced by De et al., in 1988. Descendents carrying the HA of this strain have been found in wild birds in North America since the 1970s (Matrosovich et al., 1999; Zhou et al., 1999; Obenauer et al., 2006). However, thus far these isolates of H5 from North America appear to be strains with low pathogenicity that infect only birds (USDA, 2007).

A separate highly pathogenic lineage of H5N1 has caused significant international concern due to their spread across Eurasia and Africa. This pathogenic lineage can be traced back to a 1996 isolate from a goose in China's Guangdong province (Xu et al.,

1999). As there was little monitoring of H5N1 in Asia before the human cases of H5N1 in Hong Kong in 1997, the deep origins of the pathogenic lineages of H5N1 remains unclear. Further complicating the deep history of H5N1, there are few isolates of H5N1 collected in the period 1960 to 1995 that have been fully sequenced[1] and put into the public domain.

When an outbreak of H5N1 in 1997 in live bird markets in Hong Kong infected 18 people, six of whom died, this was the first indication that H5N1 had not only become very pathogenic in birds but also that it has achieved the ability to infect humans (Xu et al., 1999; Chan, 2002). Since 2003, there have been 438 documented human cases of H5N1 and 60% have died (WHO, 2009). To date, the H5N1 virus has killed people in the Middle East, west Africa, and throughout Asia.

Phylogeographic analysis of H5N1

In most phylogeographic studies of H5N1 (e.g., Ducatez et al., 2007; Guan et al., 2002; Wallace et al., 2007; Wallace and Fitch, 2008; Wang et al., 2008), transmission routes [2] have typically been determined by visual inspection of a single phylogenetic tree. Kilpatrick et al., (2006) used a single tree in conjunction with data from observations on migration routes of wild birds and international poultry trade to estimate opportunities for the virus to spread across various borders. Smith et al., (2006) and Salzberg et al., (2007) optimized geographic regions on a phylogenetic tree. These analyses have the advantage

[1] Most sequences from this period are of low quality and were not included in this study.
[2] With the phrases "transmission route" and "transmission event", we refer to a viral lineage crossing a political border in a character evolution study.

that the nodes of the tree are assigned ancestral states, enabling areas of origin to be estimated. Janies et al., (2007) and Hill et al., (2009) optimized other features such host shifts and key mutations onto a phylogenetic tree for H5N1 and projected the tree into a virtual globe. Using virtual globes also allows for animation over time to reconstruct the geographic spread of viral lineages. All of these studies present a single tree per genetic segment or whole genome for optimization and visualization (Smith et al., 2006; Salzberg et al., 2007; and Janies et al., 2007; 2008; Hill et al., 2009).

In this paper, we compare our results based on multiple trees and large datasets to the work of Wallace et al., (2007) and Wallace and Fitch, (2008). Wallace et al. (2007) analyzed 192 HA sequences and Wallace and Fitch (2008) analyzed 482 HA and 430 NA sequences. These authors used a variety of optimality criteria for tree search. This research group chose to present a single parsimony tree for the 2007 paper and a single maximum likelihood tree for the 2008 paper.

Here we examine whether the analysis of a single tree in phylogeographic studies of avian influenza (H5N1) tends to underestimate possible routes of transmission of the virus across geographical borders. We also update the HA and NA datasets for H5N1 with recently released sequence data and address the implication for regional spread of influenza.

We use combined efficient heuristic search strategies with computing clusters to find large sets a set of presumed most parsimonious trees (MPT). We also explore various coding schemes for political and geographic boundaries as characters.

We developed a novel method for visualization of transmission routes of diseases. The "Supramap" method of Janies et al., (2007) can produce very complex visualizations, especially for large datasets, and only displays a single tree. In order to address these issues we developed a technique, termed "Routemap", that allows the user to 1) display only the routes of interest and 2) summarize transmission between regions implied by one to many candidate trees. The essential difference between Supramap and Routemap is that Supramap projects one tree whereas Routemap distills information from a pool of trees. Moreover Routemap presents the results of character change studies not as a projected tree but rather as an interactive map with lines representing viral transmission routes between localities. We distribute the visualization method as an easy to use web application called "Routemap" (http://routemap.osu.edu). In doing so, we aim that the application will be of general utility in biomedical and natural science applications of phylogeography.

Materials and methods

Current datasets

This study is based on all high quality nucleotide sequences for HA and NA genes from H5N1 available in The National Institutes of Health's GenBank

(http://www.ncbi.nlm.nih.gov) as of June 18, 2008. We add a single sequence made available by via two labs[3] participating in the Global Initiative for Sharing of All Influenza Data (GISAID, http://www.gisaid.org).

The following quality criteria were used to determine sequence quality: Sequences must be >75% complete, must not contain frame breaking insertions or deletions, and must not have a stop codon except at the end. Sequences from viral strains passaged through egg or other cell cultures or laboratory animals several times were excluded. When multiple sequences from the same isolate (as determined by GenBank taxonomic identification number) were available, only the most complete sequence was included. A total of 1646 HA and 1335 NA sequences were included in the final analyses. These sequence alignments are available at http://routemap.osu.edu/publications.

The segmented genome of influenza A viruses presents a difficulty for phylogenetic analysis. Since reassortment between diverse strains is reported to be common, two separate genomic segments are unlikely to share a phylogenetic history (Webster et al., 1992; Guan et al., 2002; Hatchette et al., 2004; Chen et al., 2006; Lindstrom et al., 1998, 2004). Reassortment violates some of the fundamental assumptions of the combined analysis (i.e. total evidence) approach such as the presumption of a shared history among

[3] Lab for virus isolation: Mona M Aly, Abdel S Arafa from National Laboratory for Quality Control on Poultry Production, Dokki (Cairo), Egypt .

Lab for sequencing: Isabella Monne, Alice Fusaro, Giovanni Cattoli, Ilaria Capua from OIE/FAO Reference Laboratory for Avian Influenza and Newcastle Disease, Istituto Zooprofilattico Sper.le delle Venezie, Legnaro (PD), Italy.

various forms of data reflecting the history of the organismal lineages (Kluge, 1989). Therefore, we do not combine data from HA and NA into a single dataset.

Multiple alignment was performed on nucleotide sequences using CLUSTALW (version 2.09 Larkin et al., 2007) using the default settings (gap opening penalty 10, gap extension penalty 0.2, DNA transition weight 0.5). A small number of internal gaps in the alignment were redistributed such that the final alignment did not imply any breaks in the reading frame in ancestral sequences. The 5' and 3' ends of the alignment were trimmed using the start and stop codons as reference. Leading and trailing gaps that remained were replaced with question marks to indicate missing data. Internal gaps were treated as a fifth character state. These datasets are hereafter referred to as "HA1646" and "NA1335" datasets.

Wallace and Fitch datasets

The HA nucleotide sequence dataset analyzed by Wallace and Fitch (2008) was recreated from the list of GenBank accession numbers available as supplemental data to their paper. Following the materials and methods section of Wallace and Fitch (2008), their HA sequence dataset was aligned using CLUSTALW with the default settings without any manual adjustment. Leading and trailing gap markers were replaced with question marks to indicate missing data. Internal gap markers were treated as a fifth character state. This dataset is hereafter referred to as the "HA482WF" dataset.

Phylogenetic analysis

Phylogenetic analysis was performed using TNT (Goloboff et al., 2008) in parallel and sequential modes. All trees were rooted with A/Chicken/Scotland/1959/H5N1 as an outgroup. This study focuses on the highly pathogenic Asian lineage of H5N1 that is typically assumed to have originated with A/Goose/Guangdong/1996. We used the older isolate to check this assumption and polarize geographic character state changes. A preliminary analysis was performed to find the length of presumed most parsimonious trees. The TNT command '`xmult = level 5 replications 100000`' was used for the preliminary runs on a single CPU. The argument '`level 5`' allows for very thorough searches and was used only to find the presumed minimum length for each dataset. The preliminary runs were interrupted after 24 hours if they failed to produce shorter trees than previously found.

To create a representative sampling of the space of MPTs, the presumed minimum tree length was used to guide a new search to find 1000 independent hits to minimum length. The TNT command '`xmult = level 0 hits 1000 giveupscore n`' was used, where '`n`' is the tree length of the most parsimonious trees found by the preliminary run.

To ensure that trees were well sampled, a cumulative series of consensus trees were calculated to check that a stable consensus had been reached. The first consensus tree was based on trees 0-9, the second on the first consensus plus trees 10-19, the third on the

second consensus and trees 20-29 […] until all trees found from sampling 1000 hits to minimum length have been included. Consensus trees were calculated using the TNT command '`nelsen* [startree] [endtree];`' which calculates a strict consensus of all trees in the given range and saves the consensus as the last tree in memory. The TNT command '`tnodes`' was used to count the number of nodes that each consensus tree contained to provide a metric of their stability. A consistent number of nodes in a consensus tree indicate that we have collected a pool of candidate trees that well represents topologies at presumed minimum length.

Character optimization

Transmission events were calculated using a modified version of the method developed by Slatkin and Maddison (1989). To find transmission events, different geographic regions are treated as character states of a single multistate character, and are mapped onto trees obtained from phylogenetic analyses using standard optimization methods (Farris, 1970; Fitch, 1971; Swofford and Maddison, 1987).

To make our results comparable with previous phylogeographic studies of H5N1, we use equal weight between geographical character states. However, if preferred, the multistate character can be coded as a Sankoff character with a transition matrix specifying the relative costs of each type of transition between character states (Sankoff and Rousseau, 1975).

We counted the minimum and maximum number of transmissions between all pairs of geographic character states. When a character state change is observed, this is interpreted as a possible transmission route of the virus. The state changes are directional, so results are expected with different numbers for transformations from state A to state B and state B to state A. Figure 1 illustrates how state changes are counted.

Geographic information for the molecular datasets was extracted from the 'TSeq_orgname' field in GenBank's TinySeq XML records. For each molecular dataset (HA1646, NA1335, and HA482WF), new geographic character matrices were created. Each geographic character matrix contained an entry for a multistate character for each taxon in the nucleotide dataset. The order of taxa was maintained in nucleotide and geographic character datasets as this is important for consistency when using compact tree format and other tree formats that use numerical taxon names in TNT.

Three geographic character data sets of 31 character states were used to find transmission routes. Due to the limitations of the character state space allowed by the tree search software, TNT, we were unable to issue a single set of geographic character states suitable for comparing the HA1646 and NA1335 datasets to some previous studies. However, we were able to compare our results to Wallace and Fitch (2008). The first geographic character set we created, termed "W&F", was used to assign each taxon to a geographic category matching those used by Wallace and Fitch (2008). For taxa collected in localities not matching any of the 28 character states used in W&F, we assigned

character states to these taxa with the most fitting of three new states: "Other China", "Other Asia" and "Other World".

We created a second character set termed "Africa" that focuses on relationships and sources of H5N1 in regions of Africa and Europe. A third character set "World" was used to look at broader patterns among regions of the globe. The character set World treats mainland China as a single entity, treats Europe as four regions (north, east, south and west), and all of Africa as a single region.

Counting transmission events

For each count of transmission events, a geographic character dataset containing a single multistate character and compact tree file containing the MPTs found during the search on the corresponding nucleotide data were loaded into memory. The TNT command `change` was used to calculate the maximum number of character state changes for pairs of geographic character states. The syntax used was '`change ]./[n]/[region X] [region Y]`'. The '`]`' indicates that results should be displayed character by character, and the '`./[n]/`' denotes that character number '`n`' should be optimized in all trees in memory. For example, the command '`change ]./0/HongKong Fujian`' will return the global minimum and maximum number of transmission events from Hong Kong to Fujian, calculated over all MPTs, when only a single multistate character is used. Each direction of change among states (e.g., A to B and B to A) was calculated for all possible pairs of states in the dataset.

From the output of the 'change' command, tables were created containing the global minimum and maximum values for each location pair and direction. When the number of state changes was >0, this result was interpreted as a possible transmission route from one region to another.

Transmission events were calculated for the HA428WF sequence dataset using the W&F character state set. For 1646HA and 1335NA sequence datasets, all three (W&F, Africa, and World) character state sets were used.

We have created a web-application, Routemap, (http://routemap.osu.edu) to produce visualizations of transmission events from the output of the geographic character optimization studies over pools of trees. The required inputs to Routemap are 1) a table of comma-separated values (CSV) containing the taxon label, the geographic place for the isolates, the latitude, and longitude in decimal degrees, and 2) a matrix of aligned nucleotide or amino acid data for the isolates in FASTA format. As an option, the user can input a set of precomputed trees. In the first step, Routemap will produce a TNT file that the user can execute on their machine with the command 'tnt p filename,'. The user then will load the TNT output back into the Routemap via the browser, where the underlying application will complete the geographic calculations. Routemap's final output is a Keyhole Markup Language (KML) file compatible with Google Earth (http://earth.google.com) and other geographic information systems such as ESRI's ArcGIS Explorer (http://www.esri.com/software/arcgis/explorer/index.html. A visualization produced by Routemap is similar to a web-based airline route map. When

using a Routemap in a virtual globe the user selects localities of interest and viral transmission events implied by the analysis of phylogenetic trees are displayed as radiating from origins or arriving at destinations. The user can make one or more origin and destination selections to make as complex or as simple of a visualization as needed. The lines indicating routes are colored to indicate directional, ambiguous or bidirectional travel. Example datasets, KML files, scripts and instructions for users to make their own Routemaps are available at http://routemap.osu.edu.

Results

Alignment and phylogenetic trees

The alignment of the HA1646 dataset consisted of 1646 taxa and 1716 characters. 937 characters in HA1646 were parsimony-informative. The length of the presumed most parsimonious trees was 7284 steps. A search for 1000 independent hits to the MPT length resulted in 2721 unique trees. The number of nodes in the consensus stabilized at 500 trees at 695 nodes.

The alignment of the NA1335 dataset consisted of 1335 taxa and 1410 characters. 744 characters in NA1335 were parsimony-informative. The length of the presumed most parsimonious trees was 4795 steps. A search for 1000 independent hits to the MPT length resulted in 2649 unique trees. The number of nodes in the consensus stabilized at 860 trees at 540 nodes.

The HA482WF dataset consisted of 482 taxa and 1809 characters. 937 characters in HA482WF were parsimony-informative. Under our search strategy, the length of the presumed most parsimonious trees is 3257 steps. A search for 1000 independent hits to the MPT length resulted in 2470 unique trees. The consensus stabilized at 251 nodes after 140 trees were collected.

Reanalysis of Wallace and Fitch (2008)

Results of the reanalysis of Wallace and Fitch's (2008) HA nucleotide dataset (HA482WF) show that using a single tree, and a single optimization path, misestimates the frequency of transmission events. Moreover the use of a single tree can fail to detect possible transmission events. Figure 2 illustrates a comparison of the transmission events reported in Wallace and Fitch's (2008) with our reanalysis of HA428WF and W&F character data for localities. We used the global maximum number of transmissions from the pool of 2470 MPTs in our reanalysis. The global maximum of implied transmission events corresponds to Wallace and Fitch's use of DELTRAN optimization (Swofford and Maddison, 1987; Swofford, 1990). Thus a transmission route is indicated if it is found in any MPT.

There are 756 possible transmission routes in W&F character state data. We compare the results found by Wallace and Fitch's (2008) optimization of these character data over a single tree they derived from HA482WF sequence dataset to our reanalysis using 2470 MPT for HA482WF. 517 of the possible routes are not found by either study, one transmission route result is found only in Wallace and Fitch (2008), 169 new routes were

discovered by our reanalysis, and 69 routes are suggested by both Wallace and Fitch (2008) and our reanalysis. Detailed results are presented in supplemental data at http://routemap.osu.edu/publications.

We also analyzed W&F based on all the MPTs we considered for HA428WF or HA1646 sequence datasets. Of 756 possible routes in W&F, 528 routes were not found when optimizing these character data over MPTs we found for the HA428WF or HA1646 sequence datasets. Four transmission routes are found only in HA482WF, 158 routes are found only in HA1646, and 61 routes occur in both HA482WF and the HA1646 dataset. Thoughtful examination of transmission routes implied by the HA1646 sequence dataset required addition of three character states for viral isolates from regions not included in the HA428WF dataset. 111 additional transmission routes were detected in MPTs implied by HA1646 when we included three localities not considered by Wallace and Fitch, (2008) but considered by other authors (e.g., Ducatez et al., 2006, 2007; Salzberg et al., 2007).

Frequency of discovery of transmission routes

We compared the transmission routes found by Wallace and Fitch (2008) for a single tree to the pool of MPTs we derived from the HA428WF dataset. Wallace and Fitch (2008) did not find 22% of the transmission routes that are found in 95% or more of the MPT that we found for HA428WF. See supplemental data at http://routemap.osu.edu/publications for a summary table.

Visualization

The panels in Figure 3 contain visualizations produced with Routemap of viral traffic in Africa, east Asia, and southeast Asia from the HA1646 dataset and the World geographic character coding scheme. Visualizations of the HA1646 and NA1335 dataset under the W&F, World, and Africa geographic character coding schemes, as well as the HA428WF dataset are available at http://routemap.osu.edu/publications.

Discussion

Phylogenetic analysis and H5N1

Metadata on pathogens such as date, host, and geographic location for viral isolation are often combined with phylogenetic trees based on pathogen sequence data by researchers and public health officials during an epidemic. In order to interpret metadata in a phylogenetic context, character optimization is a necessary tool. Studies that use metadata and character optimization include investigations of: where H5N1 spreads (Wallace et al., 2007, and Wallace and Fitch, 2008; Salzberg et al., 2007), where and when H5N1 has switched hosts (Janies et al., 2007) or experienced mutations that confer resistance to drugs (Hill et al., 2009).

Character optimization in phylogeographic analysis

Here we demonstrate that when reporting transmission events from character state changes, a single tree and a single way of optimizing the character will only tell a fraction of the story. Wallace et al., (2007) and Wallace and Fitch (2008) used MIGRAPHYLA (HoDac et al., 2007) and PAUP* (Swofford, 2008) to calculate transmission events, interpreted as character state changes using a single multistate character with geographical regions as character states. As Wallace and Fitch (2008) point out: "PAUP's assignments may represent only one of several possible most-parsimonious traces". Thus Wallace and Fitch (2008) inferred transmission events from a single possible optimization out of many, of a 28-state character, on a single tree, out of many possible trees. When we examine the length of this 28-state character on the pool of 2470 MPTs from the HA482WF sequence dataset, the lengths are between 107 and 118 steps. This range of lengths is consistent with the 115 steps reported in Figure S2 in Wallace and Fitch (2008). On the other hand, when we examine the average number of transmission events supported by the MPTs we found for the HA482WF dataset, the numbers of transmission events are very different. Each MPT supports on average 131 different transmission routes in contrast to the 70 unique transmission routes reported by Wallace and Fitch (2008).

The difference in results between TNT and PAUP* is because PAUP* only considers a single optimization path of the character, when calculating steps on polarized character state changes. This is the behavior of PAUP* even in cases where there are multiple equally parsimonious optimization paths for the character and each path may support

different sets of polarized state changes. In contrast, in TNT it is possible to calculate the maximum (or minimum) number of polarized state changes (e.g., from state A to state B) between two states of a multistate character. Moreover, in TNT this calculation can be done over multiple trees and multiple optimization paths per tree. Although the sum of all state changes calculated this way with TNT will exceed the minimum length of the character, it is important to find all possible state changes.

We also discovered an artifact in the case of closely related taxa. The number of observed state changes between two states should not necessarily be interpreted as the number of implied transmission events. In cases where the spread of an infectious agent is rapid, it is very likely that isolates from neighboring regions will be genetically identical or fail to have obtained any parsimony informative mutations. In such cases, the number of observed state changes in the MPTs in a clade with no supported branches (*sensu* the command `collapse 3` in TNT) will only reflect the proportion of taxa from each location, when optimizing over a pool of dichotomous, MPTs. In other words, given a character with two geographic states (location A and location B), the maximum number of transmission events between A and B will be the maximum length of the character in the subtree. For example, if there are 10 taxa from location A and five from location B in a subtree, the maximum number of transmission events will be five, when counting changes from A to B or from B to A.

In cases where there are many sequences in a region where there is little or no genetic difference, the artifact we describe above can occur. In other cases, where distinct genetic

lineages have been exchanged between two places, counting the changes in geographic characters is a useful tool to understand viral traffic. Both the HA428WF and HA1646 sequence datasets contain large subclades (e.g., in Africa) of viruses with identical genetic sequences. The recent invasion of H5N1 into Africa without much molecular evolution in HA makes it of interest for future studies to examine these isolates from a multi locus approach as done by Salzberg, et al., (2007) and Janies et al., (2007). These concepts are also applicable to pandemic H1N1, which emerged in early 2009 and has spread rapidly around the world.

Current phylogenetics software (PAUP*, Swofford, 2008; TNT, Goloboff et al., 2008) limit a categorical character to 31 states. These programs are suitable for most analyses of morphological and sequence data in which the data are used to find trees. MESQUITE (Maddison and Maddison, 2007) can summarize character state changes over pools of trees. MESQUITE is thus an alternative for these analyses, but is limited to 56 states and would have to be coupled to a strong tree search program. POY4 (Varón et al., 2008) may provide a solution to the state space limitation as the user can define any alphabet to describe states for features of interest of pathogens and transformation costs among states in a look-up table (e.g., a Sankoff matrix). Heuristic efficiency will be important as an accurate regional map may require hundreds or thousands of geographical character states.

When a widely distributed taxon such as influenza A is studied, geographic character coding requires some abstraction of the globe into regions. We followed the 28-location

state coding scheme of Wallace and Fitch (2008), as well as implementing two alternative coding schemes (i.e. Africa and World) to make our results comparable to other studies. When comparing the HA1646 and NA1335 sequence datasets to the results of Wallace and Fitch (2008), 275 isolates were not compatible with the 28 states defined by Wallace and Fitch (2008). We coded the isolates that fell outside of the character state space with one of three new character states: Other Asia, Other China, and Other World. This coding scheme allowed our analyses to find transmission routes outside the scope of Wallace and Fitch's (2008) original publication, while retaining comparable results.

By including more isolates from all regions sampled by HA1646 and NA1335 sequence datasets, we identified more possible routes of transmission than Wallace and Fitch (2008). For example, when using the same character data as W&F, our results agree with those of an independent study, Wang et al., (2008), in finding a single unambiguous transmission from the Hunan province of China to Indonesia. Wallace and Fitch (2008) suggested a transmission from Guangdong, China to Indonesia. The Guangdong to Indonesia route is not observed under any optimization of character data in W&F across the MPTs we found for the HA1646 sequence dataset.

In all studies of the phylogeography of influenza there are problems associated with the imprecise naming of viruses. For example in the HA1646 sequence dataset, 18 viruses are labeled "China" which represents an area of 9,598,08 km^2. This area is comparable to 10,180,000 km^2 for all of Europe, including Russia west of the Ural Mountains. Other viral isolates are named for cities. In some cases viral isolates are named for regions.

Many workers, including Wallace have suggested that GenBank records for viral isolates should include latitude and longitude records to avoid geographic imprecision (Butler, 2008). Similarly, there is much room for improvement in metadata standards for naming of host taxa (Janies et al., 2007). In the wake of the H1N1 pandemic, GISAID has vastly improved metadata standards. High quality metadata will improve research in viral character evolution and geography studies.

We used the results from HA1646 and NA1335 sequence datasets, and character states from the World coding scheme to find transmission routes discussed in the literature. The major difference between W&F and World is that mainland China is treated as a single region in the World character state set, and countries in Europe and Africa are separate into regions (discussed above). Consolidation of regions within China was necessary to allow all countries (as defined by political borders) in east Asia to fit into the 31 character state limit.

Current state of the H5N1 epidemic

Since 2003, H5N1 has been endemic in Indonesia. All phylogenetic studies including HA sequences isolated from H5N1 in Indonesia have found a monophyletic clade of Indonesian isolates, independent of optimality criterion (neighbor joining: Guan et al., 2004; Smith et al., 2006; Salzberg et al., 2007, Bayesian: Wang et al., 2008; Smith et al., 2006, maximum likelihood: Kilpatrick et al., 2006; Wallace and Fitch, 2008; Hill et al., 2009, parsimony: Janies et al., 2007; Salzberg et al., 2007; Wallace et al., 2007).

It is apparent that the H5N1 strains within Indonesia occurred due to a single transmission event from southern China. Janies et al., (2007) use whole genomes to demonstrate this link. Guan et al., (2004) use HA data to suggest the closest relatives to isolates from Indonesia are viruses from Yunnan province in China. Wallace et al., (2007), and Wallace and Fitch (2008) also us HA data to find that Guangdong province in China is the most likely source. These results stand in contrast to Kilpatrick et al., (2006), who used data from international trade to assert that the source of H5N1 in Indonesia was wild birds imported from Thailand. Our analyses of sequence and geographic data for H5N1 are in disagreement with the assertion of Kilpatrick et al, as we never find a Thailand to Indonesia route. Our analyses of the HA1646 sequence dataset, suggest that the geographic origin of H5N1 for Indonesia was Hunan province in China. Wang et al., (2008) concur. In our analyses of HA1646 sequence data, the Hunan-Indonesia migration route appears in all MPTs independent of optimization method.

Similarly our analyses of the NA1335 sequence dataset suggest Hunan as a possible source for Indonesia but some of the trees for NA1335 also indicate Hong Kong and the Guangxi Zhuang Autonomous Region of China as possible sources for Indonesia. Conflict between genes result could represent a reassortment event in which viruses in Indonesia have HA and NA genes with distinct histories or could simply reflect ambiguity in the data.

Smith et al., (2006), found that all H5N1 from Vietnam, Cambodia and Thailand form a single clade, except for one isolate from Vietnam. We were unable to find this clade structure in any tree found among the most parsimonious trees from the analyses of the HA1646 or NA1335 datasets. Although most isolates from Vietnam, Thailand, Laos, Cambodia and Vietnam are found in a monophyletic clade, isolates from Vietnam and Thailand are always found outside of this clade. This result suggests that the results of Smith et al., (2006) are an artifact of limited taxon sampling.

In the pool of trees resulting from searches of the HA1646 sequence data, we see at least three separate introductions of H5N1 to Japan. There were transmissions events from China to Japan in 2003 and 2008. Under certain optimizations, Vietnam is also a possible source of the 2008 isolates in Japan. In contrast, the source of the 2004 outbreak in Japan appears to be South Korea in all trees and all optimizations based on analysis of HA1646 sequence data and the W&F character data.

These comparisons illustrate the danger in excluding data. Wallace et al., (2007) and Wallace and Fitch (2008) both find the China-Japan connection, but excluded isolates collected in 2002-2003 from South Korea despite the fact these data were available in public databases as of 2005. Due to this exclusion, Wallace et al., (2007) and Wallace and Fitch (2008) did not find South Korea as a source of the Japanese outbreak in 2004. Our phylogenetic analysis indicating transmission of H5N1 from South Korea to Japan in

2004 is in agreement with Kilpatrick et al., (2006) who based their inference on trade and bird migration data.

The H5N1 outbreaks in Africa have been investigated using HA sequences (Ducatez et al., 2006) as well as full genomes (Ducatez et al., 2007, Salzberg et al., 2007). However, these studies were conducted when there was a limited set of putative outgroup(s) for the African lineages in the public domain. These studies conclude that the sources of H5N1 are not southeast Asia, but rather central Asia and Europe. The pools of trees resulting from our analysis of HA1646 agree with Ducatez et al., (2006; 2007) that H5N1 has been introduced to Nigeria on three separate occasions. Using the "Africa" character state set, we find corroboration for the sources to be central Asia or Europe rather than southeast Asia. By examining at all trees and all optimizations we found for HA1646 and the W&F character state set, we can not rule out the source of one transmission to Africa was the Qinghai province of China (Salzberg et al., 2007).

Sources for H5N1 in Europe are unclear [4], as multiple possible sources are indicated for most countries. The basis for the ambiguity is the high similarity of the isolates found in Europe, Africa, and the Middle East in 2005-2006. This is another empirical example of the artifact we described above, where very low genetic variation leads to every region being a possible source for its neighbors.

[4] One important exception is the case of an infected eagle (*Nisaetus nipalensis* Hodgson, 1836) that was smuggled from Thailand to Belgium but did not cause any further spread (Van Borm et al., 2005).

Perils of using a single tree

A single tree is likely to miss several transmission routes found frequently among the MPTs, as illustrated in Fig. 4. All of the publications on phylogeographic analysis of H5N1 we have reviewed here use a single tree for the results they report. This practice is independent of whether they use Slatkin and Maddison (1989) transmission event analysis (Wallace et al., 2007; Wallace and Fitch, 2008) or other optimization methods [POY3 (Wheeler et al., 2005) command `-diagnose` in Janies et al., 2007; MESQUITE command `trace:parsimony optimzation` in Salzberg et al., 2007] or observe transmission events by inspection (e.g., Ducatez et al., 2006, 2007; Wang et al., 2008).

The use of one tree is also independent of which optimality criteria researchers use in tree search. In some cases (e.g., Smith et al., 2006; Wang et al., 2008) only a single neighbor-joining tree is produced. Janies et al., (2007) created a binary tree on which to optimize based on a strict consensus of MPT using the TNT command `randtree*`. In other cases, information from multiple trees is represented as a consensus tree. A consensus tree can be a majority rules consensus from bootstrap resampling under maximum likelihood (Kilpatrick et al., 2006; Smith et al., 2006), parsimony (Guan et al., 2004; Salzberg et al., 2007;), or a distribution of trees from a Bayesian analysis (Smith et al., 2006).

There are several problems with the use of a consensus tree in the context of character optimization. First, a consensus tree does not represent all MPTs. Also the consensus

topology is never one of the most parsimonious trees. Furthermore, trees can be constructed that do not change the consensus tree, but still are not found among the MPTs. Finally, a set of most parsimonious trees can be found, where the strict consensus tree would imply transmission routes that are never found in any of the individual MPTs, (as illustrated in Figure 4). It is thus inappropriate to optimize character state changes on a consensus tree.

Under parsimony, each tree in the pool of MPTs and each optimization of the character states are just as optimal as the next one. The transmission routes found in only a few trees and under certain optimizations are just as supported by the optimality criterion as those trees and optimization found frequently. However, under a model-based optimality criterion, probability calculations for the existence of transmission routes could be calculated. A distribution of trees from a Bayesian analysis, where each tree has a posterior probability could be used to provide confidence values for transitions between the character states in a multistate character for geography optimized on trees of various posterior probability. Alternatively the proportional likelihood can be calculated for ancestral geographic states. One could employing a method such as described to study the evolution morphological characters under likelihood (Lewis, 2001) and is implemented in MESQUITE (Maddison and Maddison, 2003).

Problems with using multiple trees

When examining character state changes over a pool of MPTs, different types of optimization need to be taken into account. If one examines the global minimum of state

changes over a pool of trees, they will find only those state changes that appear in all trees, under all optimizations. When dealing with large numbers of trees, a tree will often be found where a specific state change does not occur. For example, one has a subtree with only unsupported branches and 100 taxa from location A, 100 from location B and a single taxon from location C. When looking at this subtree in the MPTs, one can find a tree and an optimization pathway that does not have a transmission event between location A and B, making the global minimum number of transmission events from A to B equal zero.

In contrast, use of the global minimum number of polarized state changes dramatically changes the number of indicated transmission routes for a set of trees when compared to use of the maximum number of state changes. For example, when considering the maximum number of directional state changes on the HA1646 dataset, we find a total of 267 different transmission routes. This results contrasts with a result of 20 transmission routes found when looking at the global minimum number of directional state changes. In summary, the global minimum is a severe test of putative transmission routes as the route has to appear in every tree in the pool of MPTs to register a global minimum > zero. The global maximum is useful to find the range of possible transmission routes a set of MPTs implies, irrespective of how frequently a route is found.

Conclusions

Phylogenetic tree search and character evolution analyses have been valuable tools in many fields including surveillance of infectious diseases. In this context, parsimony is

sometimes used for evaluation of trees and almost always used for character optimization. The multiple equally optimal trees produced by parsimony reflect of the natural ambiguity of the sequence data. Character optimization under parsimony can find the global minimum and maximum values for state changes in a pool of trees reflecting various degrees of ambiguity underlying transmission routes for organismal lineages. With efficient character optimization and tree search heuristics it is possible to find and analyze very large pools of trees for datasets with thousands of terminal taxa. Moreover, because results are sensitive to tree choice, any tree search method that considers multiple trees has an advantage over methods that consider a single tree, such as neighbor joining.

The methods of character optimization and visualization used and developed here are also applicable to sets of trees produced by maximum likelihood and Bayesian methods. Maximum likelihood methods are commonly used for tree search and sometimes preferred for the operational reason that they support one tree slightly more than near suboptimal candidate trees. While Bayesian methods produce a distribution of trees, conclusions are typically discussed in relationship to the single tree presented (often a majority rules consensus and sometimes a maximum *a posteriori* tree). As we have demonstrated here, a single tree, as well as consensus trees are insufficient for estimating transmission routes of infectious disease. In addition to considering multiple trees, when using Slatkin and Maddison's (1989) method of geographic analysis, it is important to take into account sensitivity to different strategies for character state optimization.

It is our intention that the methods described in this paper should be general and useful for phylogeographic analysis of any type of problem above and beyond infectious diseases. We welcome everyone to our website http://routemap.osu.edu to test our methods and create KML visualization of your own data.

Acknowledgements

We thank the Pablo Goloboff, Steve Farris, Kevin Nixon, and the Willi Hennig Society for making TNT available. Pablo Goloboff, Universidad Nacional de Tucumán and Diego Pol, CONICET, Museo Paleontológico Egidio Feruglio provided insights and discussions on character analysis. We also thank two anonymous reviewers for their constructive criticisms.

We acknowledge the Department of Biomedical Informatics and The Mathematical Biosciences Institute of The Ohio State University (OSU) for space, funding, and administrative support. We thank the Medical Center Information Services team of OSU and the Mathematical Biosciences Institute for hosting computing clusters used in this study. We acknowledge this material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and Office under grant number W911NF-05-1-0271.

Literature cited

Bright, R.A., Medina, M., Xu, X., Perez-Oronoz, G., Wallis, T.R., Davis, X., Povinelli, L., Cox, N., Klimov, A. , 2005. Incidence of adamantane resistance among influenza A (H3N2) viruses isolated worldwide from 1994 to 2005: A cause for concern. 2005. Lancet 366,1175-1181.

Butler, D. 2008. Politically correct names given to flu viruses. Nature News online 23 April 2008. doi:10.1038/452923a .

Chan, P. Outbreak of Avian Influenza A(H5N1) Virus infection in Hong Kong in 1997 (2002). Clin. Infect. Dis. 34 (Suppl 2), S58–64.

Chen, H., Bahri, S., Chen, Y., Cheung, C.L., Duan, L., Fan, X.H., Guan, Y. Guo, C.T., Hassan, S.S., Huang, K., Leung, Y.H.C., Li, K.S., Lu, H.R., Naipospos, T.S.P., Nguyen, T.D., Peiris, J.S.M., Qin, K., Rayner, J.M., Smith, G.J.D., Vijaykrishna, D., Wang, J., Webster, R.G., Wu, W.L., Xia, N.S., Xu, K.M., Yuen, K.Y., Zhang, J.X., Zhang, L.J. 2006. Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. PNAS. 103, 2845-2850.

De, B. K., Brownlee, G. G., Kendal, A. P., Shaw, M. W. 1988. Complete sequence of cDNA clone of the hemagglutinin gene of influenza A/Chicken/Scotland/59 (H5N1)

virus: comparison with contemporary North American and European strains. Nucleic Acids Res. 16, 4181-4182.

Ducatez, M.F., Ammerlaan, W., de Landtsheer, S., Fouchier, R.A.M., Muller, C.P., Niesters, H.G.M., Olinger, C.M., Osterhaus, A.D.M.E., Owoade, A.A. 2006. Avian Flu: Multiple introductions of H5N1 in Nigeria. Nature 442, 37.

Ducatez, M. F., Ammerlaan, W., De Landtsheer, S., Fouchier, R. A. M., Muller, C. P., Olinger, C. M., Osterhaus, A. D. M. E., Ouedraogo, J. B., Owoade, A. A., Sow, A., Tahita, M. C., Tarnagda, Z. 2007. Molecular and antigenic evolution and geographical spread of H5N1 highly pathogenic avian influenza viruses in western Africa. J. Gen. Virol. 88, 2297-2306.

Farris, J.S. Methods for computing Wagner trees. 1970. Syst. Zool. 19, 83-92.

Fitch, W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20, 406-416.

Goloboff, P. A., Farris, J.S., Nixon, K.C. 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774-786

GISAID, 2009. Sample provided by Aly, M. M., Arafa, A. S. National laboratory for quality control on poultry production, Dokki (Cairo), Egypt. Sequence submitted by

Monne, I., Fusaro, A., Capua, I., Cattoli, G. OIE/FAO Reference Laboratory for avian influenza and Newcastle disease, Istituto Zooprofilattico Sper.le delle Venezie, Legnaro (PD), Italy.

Guan, Y., Dyrting, K.C., Ellis, T. M., Kong, K. F., Peiris, Y., Shortridge. K. F., Sit, T., Zhang, L. J. 2002. H5N1 influenza viruses isolated from geese in southeastern China: evidence for genetic reassortment and interspecies transmission to ducks. Virology 292, 16-23.

Guan, Y., Auewarakul, P., Buranathai, C., Chaisingh, A., Chen, H., Duan, L., Estoepangestie, A.T.S., Hanh, N.T.H., Li, K.S., Long, H.T., Nguyen, T.D., Peiris, J.S.M., Poon, L.L.M., Puthavathana, P., Rahardjo, A.P., Shortridge, K.F., Smith, G.J.D., Wang, J., Webby, R.J., Webster, R.G., Xu, K.M., Yuen, K.Y. 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430, 209-213.

Hatchette, T.F., Baker, A., Johnson, C., Pryor, S. P., Walker, D., Webster, R.G. 2004. Influenza A viruses in feral Canadian ducks: extensive reassortment in nature. J. Gen. Virol. 85, 2327-2337.

Hill, A., Guralnick, R. Wilson, M., Habib, F. Janies, D. 2009. Evolution of drug resistance in multiple distinct lineages of H5N1 avian influenza. Infection, Genetics, and Evolution. 9, 169-178.

HoDac, H., Fitch, W.M., Lathrop, R.H., Wallace, R.G. 2007. MigraPhyla: Statistical analysis of migration events through a phylogeny. Version 1.0b manual. http://pd.bio.uci.edu/ee/WallaceR/MigraPhyla.html

Holland, J., Grabau, E., Horodyski, F., Nichol, S., Spindler, K., VandePol, S. 1982. Rapid evolution of RNA genomes. Science 215, 1577-1585.

Huelsenbeck, J.P and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Janies, D., Hill, A.W., Habib, F., Guralnick, R., Waltari, E., Wheeler, W.C. 2007. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). Syst. Biol. 56, 321-329.

Kawaoka Y., Alexander D.J., Nestorowicz A., Webster R.G. 1987. Molecular analyses of the hemagglutinin genes of H5 influenza viruses: origin of a virulent turkey strain. Virology 158, 218-227.

Kilpatrick, A. M., Chmura, A.A., Gibbons, D.W., Fleischer, R.C., Marra, P.P., Daszak, P. 2006. Predicting the global spread of H5N1 avian influenza. PNAS. 103, 19368-19373

Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst. Zool 38, 7–25.

Larkin, M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.

Lewis, P.O. 2001. A likelihood approach to inferring phylogeny from discrete morphological characters. Syst. Biol. 50:913-925.

Lindstrom, S.E., Hiromoto, Y., Nerome, R., Omoe, K., Sugita, S., Yamazaki, Y., Takahashi, T., Nerome, K. 1998. Phylogenetic analysis of the entire genome of influenza A (H3N2) viruses from Japan: evidence for genetic reassortment of the six internal genes. J. Virol. 72, 8021-8031.

Lindstrom, S.E., Cox, N.J., Klimov, A. 2004. Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957-1972: evidence for genetic divergence and multiple reassortment events. Virology 328, 101-119.

Maddison, W. P. and D.R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org

Matrosovich, M., Kawaoka, Y., Webster, R., Zhou, N. 1999. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distiguishable properties. J. Virol. 73, 1146-1155.

Obenauer, J.C., Denson, J., Fan, Y., Finkelstein, D.B., Hoffmann, E., Krauss, S., Ma, J., Mehta, P.K., Mukatira, S., Naeve, C.W., Rakestraw, K.M., Su, X., Wang, J., Webster, R.G., Xu, X., Zhang, Z., Zheng, J. 2006. Large-scale sequence analysis of avian influenza isolates. Science 311, 1576-1580.

Salzberg, S.L., Ali, A.S.M., Aly, M.M., Brown, I.H., Capua, I., Cattoli, G., Couacy-Hymann, E., De Mia, G.M., Dung, D.H., Ghedin, E., Guercio, A., Janies, D.A., Joannis, T., Kingsford, C., Osmani, A., Padalino, I., Saad, M.D., Saviæ, V., Sengamalay, N.A., Spiro, D.J., Yingst, S., Zaborsky, J., Zorman-Rojs, O. 2007. Genome analysis linking recent European and African influenza (H5N1) viruses. Emerg. Infect. Dis. 13, 713-718.

Sankoff, D. and Rousseau, P. 1975. Locating the vertices of a Steiner tree in arbitrary space. Mathematical Programming, p. 240246.

Sawabe, K., Hayashi, T., Hoshino, K., Hotta, A., Isawa, H., Kobayashi, M., Kurahashi, H., Saito, T., Sasaki, T., Tanabayashi, K., Tsuda, Y., Yamada, A. 2006. Detection and isolation of highly pathogenic H5N1 avian influenza A viruses from blow flies collected in the vicinity of an infected poultry farm in Kyoto, Japan, 2004. Am. J. Trop. Med. Hyg. 75, 327-332.

Slatkin, M. and Maddison, W. P. 1989. A cladistic measure of gene flow inferred from the phylogeny of alleles. Genetics 123, 603-613.

Smith, D., Lapedes, A., de Jong, J., Bestebroer, T., Rimmelzwaan, G., Osterhaus, A., Fouchier, R. 2004. Mapping the Antigenic and Genetic Evolution of Influenza Virus. Science 305, 371-376.

Smith, G.J.D., Bui, N.A., Chen, H., Cheung, C.L., Dao, T.V., de Jong, M.D., Farrar, J., Guan, Y., Hassan, S.S., Hien, T.T., Leung, Y.H.C., Li, K.S., Naipospos, T.S.P., Nguyen, T.D., Nguyen, T.V., Nguyen, V.C., Peirvis, J.S.M., Poon, L.L.M., Rayner, J.M., Usman, T.B., Vijaykrishna, D., Webster, R.G., Zhang, J.X., Zhang, L.J. 2006. Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. Virology 350, 258-268.

Subbarao, K., Klimov, A., Katz, J, Regnery, H, Lim, W, Hall, H, Perdue, M, Swayne, D, Bender, C, Huang, J, Hemphill, M, Rowe, T, Shaw, M, Xu, X, Fukuda, K, Cox, N. 1998. Characterization of an avian influenza a (H5N1) virus isolated from a child with a fatal respiratory illness. Science 279, 393-396.

Swofford, D. L., 1990. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0. Manual. Illinois Natural History Survey, Champaign, Ill.

Swofford. D.L. 2008. PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D.L. and Maddison, W.P. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87, 199-229.

USDA, 2007. Release No. 0296.06. Avian influenza. Low pathogenic H5N1 vs. highly pathogenic H5N1. Latest update, 23 July 2007. http://www.usda.gov/2006/08/0296.xml

Van Borm, S., Boschmans, M., Decaestecker, M., Dupont, G., Hanquet, G., Lambrecht, B., Snacken, R., Thomas, I., van den Berg, T. 2005. Highly pathogenic H5N1 influenza virus in smuggled Thai eagles, Belgium. Emerg. Infect. Dis. 11, 702-705.

Wallace, R.G., HoDac, H., Lathrop, R.H., Fitch, W.M. 2007. A statistical phylogeography of influenza A H5N1. PNAS 104, 4473-4478.

Wallace, R.G. and Fitch, W.M. 2008. Influenza A H5N1 immigration is filtered out at some international borders. PLoS ONE 3, e1697. doi:10.1371/journal.pone.0001697.

Wang, J., Bahl, J., Chen, H., Smith, G.J.D., Duan, L.,.Guan, Y. Peiris, J. S. M., Vijaykrishna, D., Webster, R. G., Zhang, J. X. 2008. Identification of the progenitors of Indonesian and Vietnamese avian influenza A (H5N1) viruses from southern China. J. Virol. 82, 3405-3414.

Webster, R.G., Bean, W.J., Chambers, T.M., Gorman, O.T., Kawaoka, Y. 1992. Evolution and ecology of influenza A viruses. Microbiol. Rev. 56, 152-179.

Wheeler,W. C., Delaet, J., Gladstein, D., and Varón, A., 2005. POY (version 3.012). Phylogeny reconstruction via optimization of DNAand other data. http://research.amnh.org/scicomp/projects/poy.php

WHO, 2004. Avian influenza A(H5N1)- update 31: Situation (poultry) in Asia: need for a long-term response, comparison with previous outbreaks. Reported to WHO. 2 March 2004. http://www.who.int/csr/don/2004_03_02/en/

WHO, 2009. Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1) Reported to WHO. 11 August 2009. http://www.who.int/csr/disease/avian_influenza/country/cases_table_2009_08_11/en/index.html

Varón, A., Vinh, L. S., Bomash, I., Wheeler, W. C. 2009. POY 4.1.1. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php

Xu X., Cox, N.J., Guo, Y., Subbarao K. 1999. Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. Virology 261, 15-19.

Zamarin, D., Ortigoza, M.B., Palese, P. 2006. Influenza A Virus PB1-F2 Protein Contributes to Viral Pathogenesis in Mice. Journal of Virology 80: 7976-7983.

Zhou, N. N., Shortridge, K.F., Claas, E.C.J., Krauss, S.L., Webster, R.G. 1999. Rapid evolution of H5N1 influenza viruses in chickens in Hong Kong. J. Virol. 73, 3366-3374.

Figure captions

Figure 1.

The same tree is shown with two different optimizations of a geographic character. The tree on the left shows one transmission from location A to location B and one from location B to location A. The tree on the right has two transmission events from location A to location B, and zero transmission events from location B to location A. The global maximum number of transmissions are thus: location A to location B: 2; location B to location A: 1.

Figure 2.

A comparison of transmission routes found by Wallace and Fitch (2008) and our reanalysis of the WF482HA dataset. Each cell in the grid represents a transmission route between two locations in the 28 state geographical character set used in the original publication. Cells colored black indicate transmission routes that were not found in any analysis. Cells colored yellow indicate that our reanalysis agrees with Wallace and Fitch (2008). Cells colored cyan indicates that the transmission route was not found in our reanalysis. Cells colored red indicate transmission routes found in the reanalysis, but not in Wallace and Fitch (2008).

The rows represent outgoing locations and the columns incoming. For example, the cell in the Vietnam column and the Guangdong row represents a transmission from Guangdong to Vietnam, while the cell in the Guangdong column and the Vietnam row represents a transmission from Vietnam to Guangdong

Figure 3.

Migration routes indicated by the HA1646 dataset.

A. Transmissions into Nigeria, using the Africa character set. The orange lines represent transmission routes are bidirectional or ambiguous. In H5N1 in Africa, this result is due to very low genetic variation between the H5N1 in Nigeria and neighboring regions.

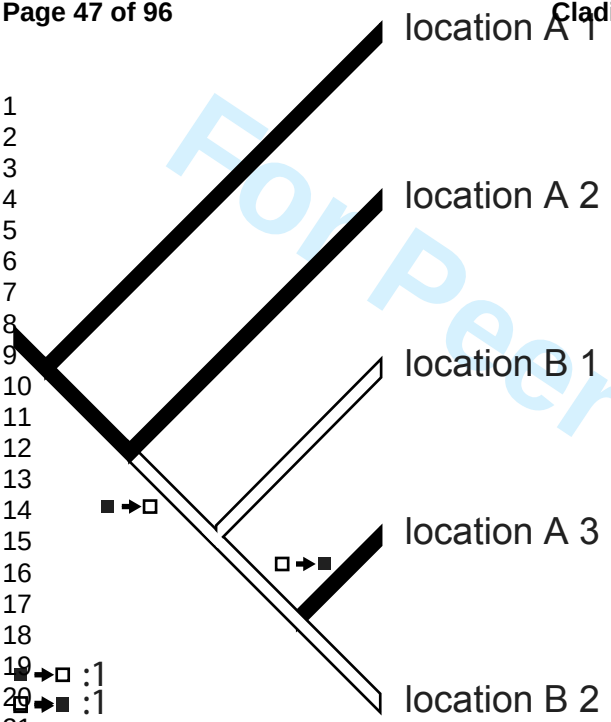
B. Transmissions into Indonesia, using the "W&F" character set. Analyses indicate only one source of H5N1 in Indonesia, Hunan province in China.

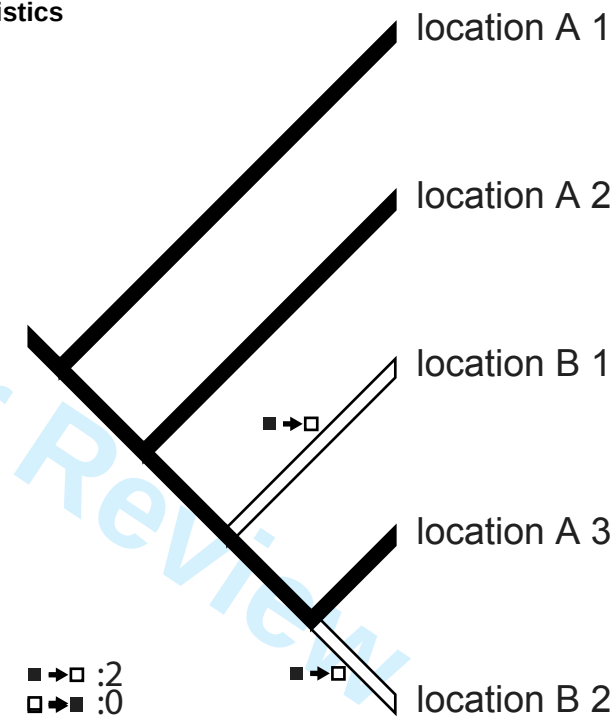
C. Transmission into Japan, using the "World" character set. Both China and South Korea are sources for the H5N1 in Japan. The direction of the transmission route to Vietnam is either bidirectional or ambiguous.

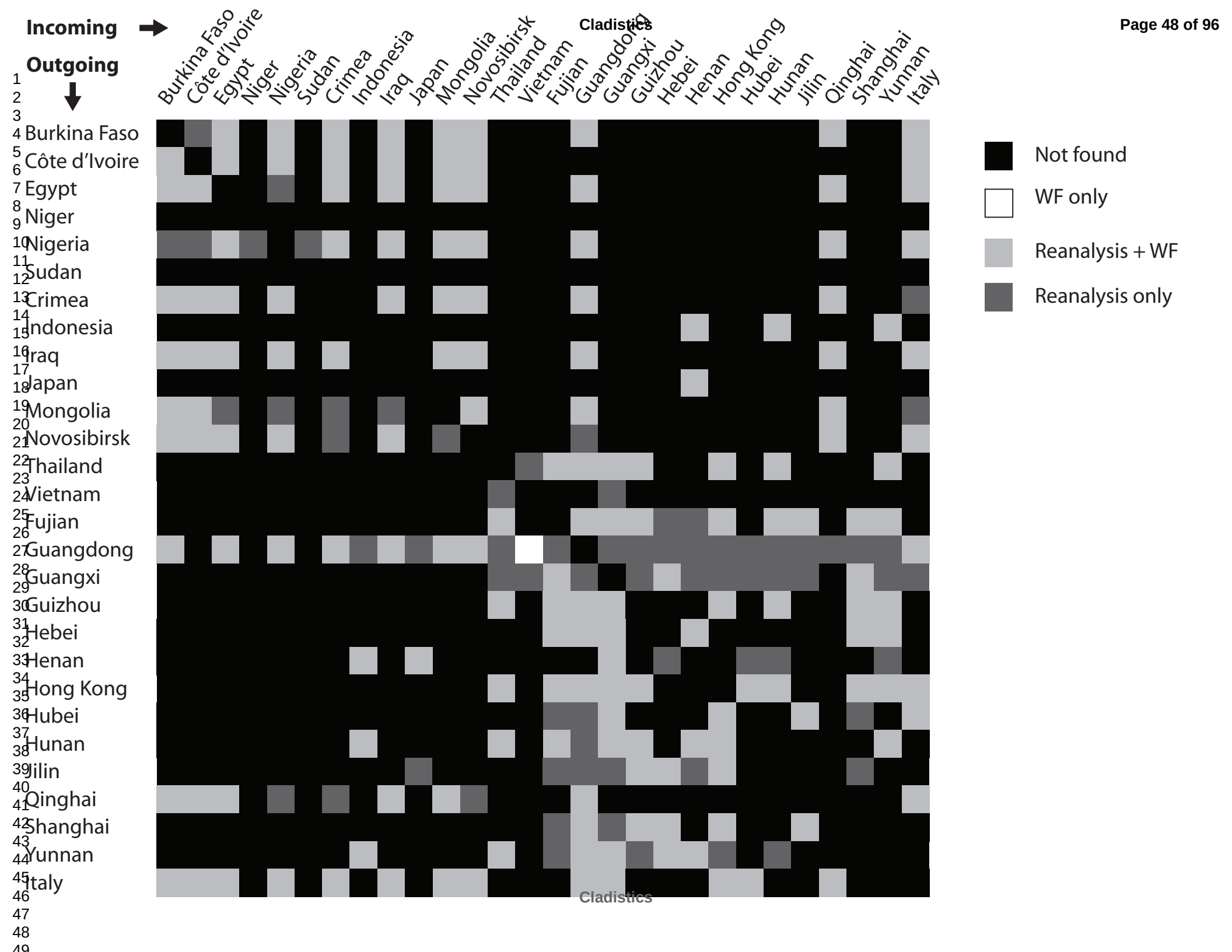
Figure 4:

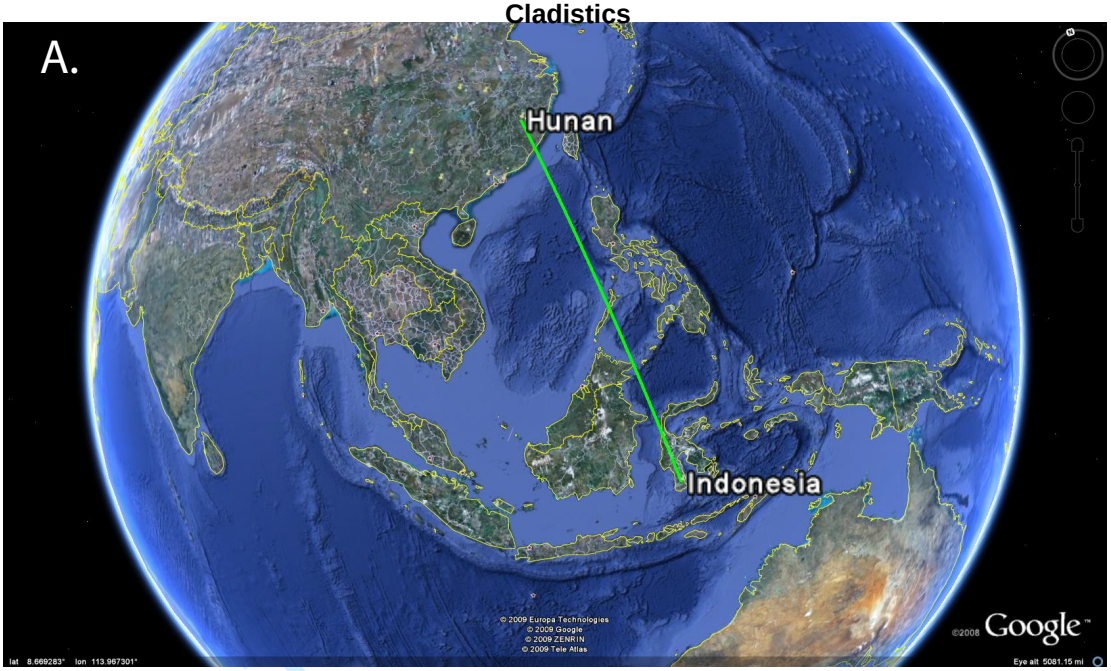
The consensus tree can imply transmission routes never found among the most parsimonious trees.

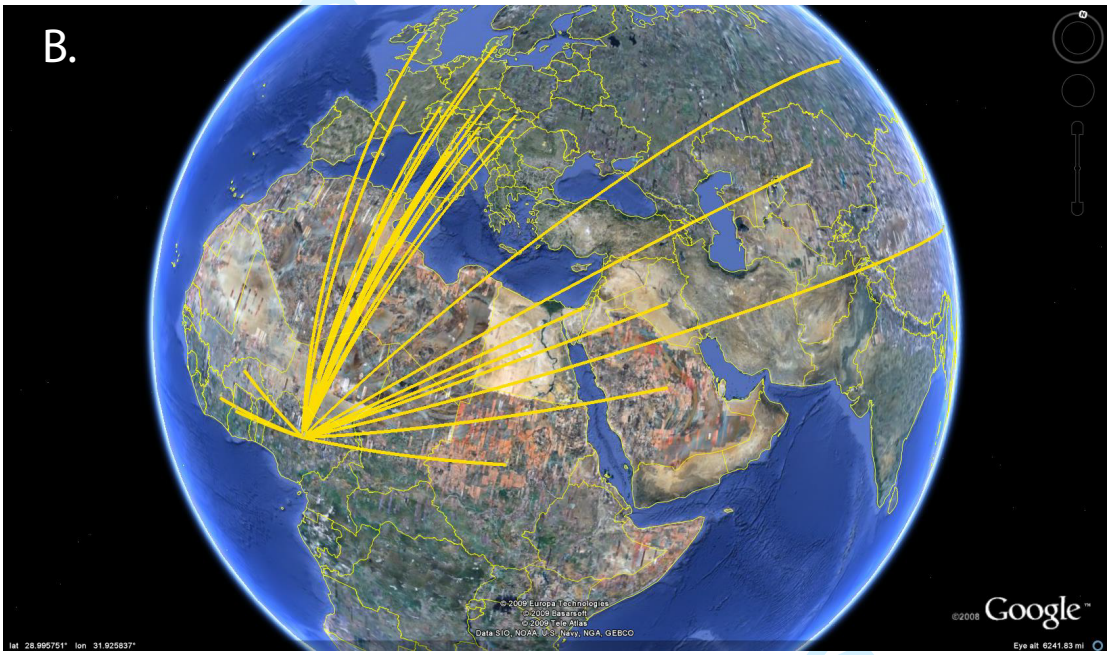
When an unordered multistate character with states 0, A, B and C is optimized on trees a. and b., no optimal optimization path has the directional state change C to A. However, considering optimal optimization paths in the strict consensus tree, one path has the node above the outgroup assigned to state B and thus two changes from B to C, as well as two changes from C to A are required.

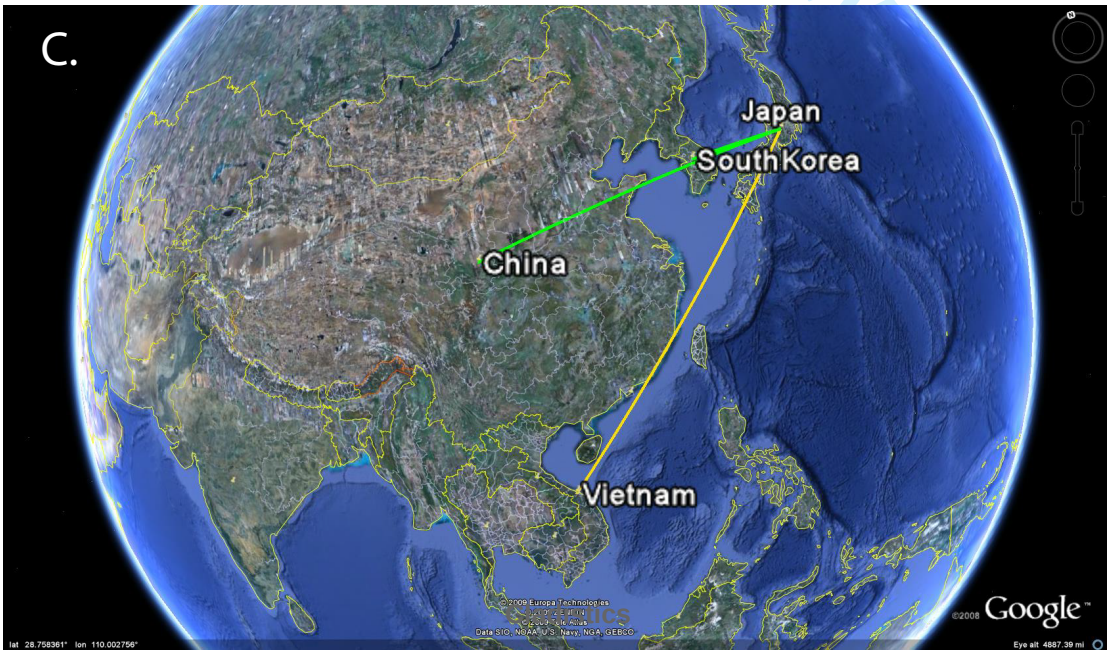


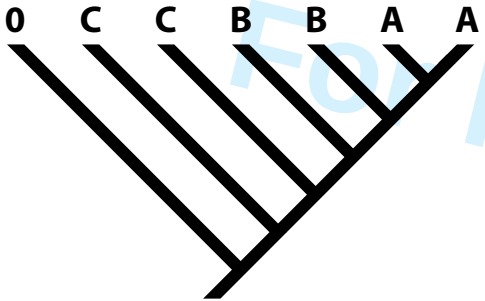



Incoming
Outgoing
Burkina Faso
Côte d'Ivoire
Egypt
Niger
Nigeria
Sudan
Crimea
Indonesia
Iraq
Japan
Mongolia
Novosibirsk
Thailand
Vietnam
Fujian
Guangdong
Guangxi
Guizhou
Hebei
Henan
Hong Kong
Hubei
Hunan
Jilin
Qinghai
Shanghai
Yunnan
Italy
Not found
WF only
Reanalysis + WF
Reanalysis only
Cladistics
Cladistics

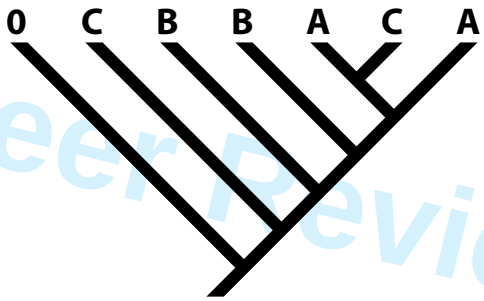

A.
Hunan
Indonesia
© 2009 Europa Technologies
© 2009 Google
© 2009 ZENRIN
© 2009 Tele Atlas
©2008 Google
lat 8.669283° lon 113.967301°
Eye alt 5081.15 mi


B.
© 2009 Europa Technologies
© 2009 Basarsoft
© 2009 Tele Atlas
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
©2008 Google
lat 28.995751° lon 31.925837°
Eye alt 6241.83 mi

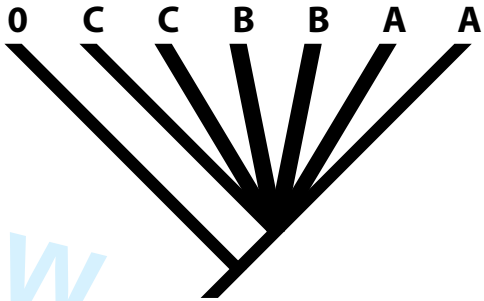

C.
Japan
South Korea
China
Vietnam
© 2009 Europa Technologies
© 2009 Tele Atlas
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
©2008 Google
lat 28.758361° lon 110.002756°
Eye alt 4687.39 mi



a. Max changes C to A: 0



b. Max changes C to A: 0



c. Strict consensus of a. and b.
Max changes C to A: 2

Tracking the Geographic Spread of Avian Influenza (H5N1) with Multiple Phylogenetic Trees

Rasmus Hovmöller [1,2], Boyan Alexandrov [2], Jori Hardman [2], and Daniel Janies [2]

1 Mathematical Biosciences Institute, The Ohio State University, Columbus, OH, USA
2 Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA.

Deleted: A

Deleted: 29
Deleted: 30
Deleted: 31
Deleted: 33
Deleted: 34
Deleted: 44
Deleted: 44
Deleted: 44
Deleted: 45

Abstract

Avian influenza (H5N1) has been of great social and economic importance since it first infected humans in Hong Kong in 1997. A highly pathogenic strain has spread from China and has killed humans in east Asia, west Africa, south Asia, and the Middle East. Recently, several molecular phylogenetic studies have focused on the relationships of various clades of H5N1 and their spread over time, space, and various hosts. These studies examining the geographic spread of H5N1 have based their conclusions on a single tree. This tree often results from the analysis of the genomic segment coding for hemagglutinin (HA) or neuraminidase (NA) proteins and a limited sample of viral isolates. Here we present the first study using multiple candidate trees to estimate geographic transmission routes of H5N1. In addition, we use all high quality HA and NA sequences available to the public as of June 2008. We estimated geographic transmission routes of H5N1 by optimizing multistate characters with states representing different geographical regions over a pool of presumed minimum length trees. We also developed means to visualize our results in Keyhole Markup Language (KML) for virtual globes. We provide these methods as a web application entitled "Routemap" (http://routemap.osu.edu). The resulting visualizations are akin to airline route maps but they depict the routes of spread of viral lineages. We compare our results with the results of previous studies. We focus on the sensitivity of results to sampling of tree space, character coding schemes, optimization methods, and taxon sampling. In conclusion, we

Deleted: out of

Deleted: East

Deleted: West

Deleted: South

Deleted: Several recent

Deleted: ,

Deleted: conclusion

Deleted: dataset in terms

Deleted: sampling of

Deleted: of the virus.

Deleted: -

Deleted: ,

Deleted: develop

Deleted:),

Deleted: , in virtual globes.

find that using one tree and a single character optimization method will ignore many of the transmission routes indicated by genetic sequence and geographic data.

Deleted: the

Introduction

Influenza viruses and pathogenic H5N1

Influenza viruses cause disease in humans, wild and domesticated birds and other homeothermic animals. The current classification divides influenza into three major groups: A, B, and C. Within influenza A, lineages are classified with a system of antibodies to the surface proteins of the virions, hemagglutinin and neuraminidase. This leads to the HXNY nomenclature where X includes 16 antigenic subtypes and Y represents nine different antigenic subtypes. Influenza B is only known to infect humans and pinnipeds (seals, sea lions and walruses). Influenza C is only known to infect humans and swine. On the other hand, influenza A, including subtype H5N1, has a much wider host range. Influenza A infects aves as wells as many mammalian groups such as artiodactyl (swine and bovids), equids (horses), carnivorans (canids, domestic and wild felids, pinnipeds, mustelids, and viverrids). H5N1 has also been found in blow flies (Diptera: Sarcophagidae) in the vicinity of farms containing infected poultry in Japan (Sawabe et al, 2004).

Influenza viruses have two means of creating genetic diversity. The segmented genome of influenza viruses enables genomic reassortment between strains. When a host is simultaneously infected by more that one strain of influenza virions that carry gene segments from two or more original viruses, termed "reassortants" can be produced. In addition, influenza viruses are negative strand-RNA viruses. These viruses use RNA

Deleted: 9
Deleted: ,
Deleted: .
Deleted: artiodactyls (
Deleted:).
Deleted: felids (domestic and wild), pinnipeds, artiodactyls
Deleted: .
Deleted: farms
Deleted: reassortments
Deleted: A
Deleted: can produce reassortant
Deleted: several
Deleted: . Influenza viruses
Deleted: , and on

polymerases without proofreading functions, resulting in a high mutation rate (Holland et al., 1982). Both of these molecular evolutionary processes allow many lineages of influenza to escape host immune responses (Smith et al., 2004) and antiviral therapeutics (Bright, et al., 2005; Hill et al., 2009).

The origins of H5N1 and its variable pathogenicity

The RNA for hemagglutinin (HA) encodes a polyprotein with two subunits that must be cleaved by host endoproteases to enable the virion to fuse with the host's cell membrane. Highly pathogenic strains of H5N1 are characterized by multiple basic amino acids at the cleavage site between the two subunits of the hemagglutinin protein (Kawaoka, 1987; Subbarao, 1998). Pathogenicity is a multigenic trait that is also mediated by features such as whether the polymerase basic 2 F2 protein is produced or not (Zamarin et al., 2006)

The earliest known influenza strain characterized as serotype H5N1 was isolated from birds in Scotland in 1959 (WHO, 2004) and sequenced by De et al., in 1988. Descendents carrying the HA of this strain have been found in wild birds in North America since the 1970s (Matrosovich et al., 1999; Zhou et al., 1999; Obenauer et al., 2006). However, thus far these isolates of H5 from North America appear to be strains with low pathogenicity that infect only birds (USDA, 2007).

A separate highly pathogenic lineage of H5N1 has caused significant international concern due to their spread across Eurasia and Africa. This pathogenic lineage can be traced back to a 1996 isolate from a goose in China's Guangdong province (Xu et al.,

Deleted: ability

Formatted: Font: Times New

Deleted: Origin

Deleted: ¶
¶
The earliest known influenza virus characterized as serotype H5N1 was isolated from birds in Scotland in 1959 (WHO, 2004),

Deleted: first sequenced by De et al. in 1988. H5N1 had been found occasionally in wild birds in North America since the 1970s. However, thus far these isolates of H5N1 from North America appear to be strains with low

Deleted: and only infect birds (USDA, 2007).

Deleted: virions

Deleted: lineages

Deleted: that have

Deleted: an

1999). As there was little monitoring of H5N1 in Asia before the human cases of H5N1 in Hong Kong in 1997, the deep origins of the pathogenic lineages of H5N1 remains unclear. Further complicating the deep history of H5N1, there are few isolates of H5N1 collected in the period 1960 to 1995 that have been fully sequenced[1] and put into the public domain.

When an outbreak of H5N1 in 1997 in live bird markets in Hong Kong infected 18 people, six of whom died, this was the first indication that H5N1 had not only become very pathogenic in birds but also that it has achieved the ability to infect humans (Xu et al., 1999; Chan, 2002). Since 2003, there have been 438 documented human cases of H5N1 and 60% have died (WHO, 2009). To date, the H5N1 virus has killed people in the Middle East, west Africa, and throughout Asia.

Phylogeographic analysis of H5N1

In most phylogeographic studies of H5N1 (e.g., Ducatez et al., 2007; Guan et al., 2002; Wallace et al., 2007; Wallace and Fitch, 2008; Wang et al., 2008), transmission routes[2] have typically been determined by visual inspection of a single phylogenetic tree. Kilpatrick et al., (2006) used a single tree in conjunction with data from observations on migration routes of wild birds and international poultry trade to estimate opportunities for the virus to spread across various borders. Smith et al., (2006) and Salzberg et al., (2007) optimized geographic regions on a phylogenetic tree. These analyses have the advantage

[1] Most sequences from this period are of low quality and were not included in this study.
[2] With the phrases "transmission route" and "transmission event", we refer to a viral lineage crossing a political border in a character evolution study.

Deleted: no
Deleted: a 1997
Deleted: which six
Deleted: cross the species barrier to
Deleted: 394
Deleted: 63% of infected persons
Deleted: .
Deleted: are
Deleted: or a majority rules consensus tree based on bootstrap resampling or Bayesian posterior probabilities (Ronquist and Huelsenbeck, 2001).
Deleted: .
Deleted: .
Deleted: .
Deleted: The isolates in question all fail our
Deleted: requirements,
Deleted: are therefore

that the nodes of the tree are assigned ancestral states, enabling areas of origin to be estimated. Janies et al., (2007) and Hill et al., (2009) optimized other features such host shifts and key mutations onto a phylogenetic tree for H5N1 and projected the tree into a virtual globe. Using virtual globes also allows for animation over time to reconstruct the geographic spread of viral lineages. All of these studies present a single tree per genetic segment or whole genome for optimization and visualization (Smith et al., 2006; Salzberg et al., 2007; and Janies et al., 2007; 2008; Hill et al., 2009).

Deleted: determined.
Deleted: .
Deleted: as
Deleted: and host shifts
Deleted: and animated
Deleted: This analysis combines temporal, genotypic, phenotypic, and
Deleted: studies into one context.

In this paper, we compare our results based on multiple trees and large datasets to the work of Wallace et al., (2007) and Wallace and Fitch, (2008). Wallace et al. (2007) analyzed 192 HA sequences and Wallace and Fitch (2008) analyzed 482 HA and 430 NA sequences. These authors used a variety of optimality criteria for tree search. This research group chose to present a single parsimony tree for the 2007 paper and a single maximum likelihood tree for the 2008 paper.

Deleted: .
Deleted: They choose

Here we examine whether the analysis of a single tree in phylogeographic studies of avian influenza (H5N1) tends to underestimate possible routes of transmission of the virus across geographical borders. We also update the HA and NA datasets for H5N1 with recently released sequence data and address the implication for regional spread of influenza.

We use combined efficient heuristic search strategies with computing clusters to find large sets a set of presumed most parsimonious trees (MPT). We also explore various coding schemes for political and geographic boundaries as characters.

We developed a novel method for visualization of transmission routes of diseases. The "Supramap" method of Janies et al., (2007) can produce very complex visualizations, especially for large datasets, and only displays a single tree. In order to address these issues we developed a technique, termed "Routemap", that allows the user to 1) display only the routes of interest and 2) summarize transmission between regions implied by one to many candidate trees. The essential difference between Supramap and Routemap is that Supramap projects one tree whereas Routemap distills information from a pool of trees. Moreover Routemap presents the results of character change studies not as a projected tree but rather as an interactive map with lines representing viral transmission routes between localities. We distribute the visualization method as an easy to use web application called "Routemap" (http://routemap.osu.edu). In doing so, we aim that the application will be of general utility in biomedical and natural science applications of phylogeography.

Deleted: computationally
Deleted: for
Deleted:)
Deleted: models for
Deleted: the regional
Deleted: methodological implications are expected to

Materials and methods

Current datasets

This study is based on all high quality nucleotide sequences for HA and NA genes from H5N1 available in The National Institutes of Health's GenBank

Deleted: (NIH)

(http://www.ncbi.nlm.nih.gov) as of June 18, 2008. We add a single sequence made available by via two labs[3] participating in the Global Initiative for Sharing of All Influenza Data (GISAID, http://www.gisaid.org).

The following quality criteria were used to determine sequence quality: Sequences must be >75% complete, must not contain frame breaking insertions or deletions, and must not have a stop codon except at the end. Sequences from viral strains passaged through egg or other cell cultures or laboratory animals several times were excluded. When multiple sequences from the same isolate (as determined by GenBank taxonomic identification number) were available, only the most complete sequence was included. A total of 1646 HA and 1335 NA sequences were included in the final analyses. These sequence alignments are available at http://routemap.osu.edu/publications.

The segmented genome of influenza A viruses presents a difficulty for phylogenetic analysis. Since reassortment between diverse strains is reported to be common, two separate genomic segments are unlikely to share a phylogenetic history (Webster et al., 1992; Guan et al., 2002; Hatchette et al., 2004; Chen et al., 2006; Lindstrom et al., 1998, 2004). Reassortment violates some of the fundamental assumptions of the combined analysis (i.e. total evidence) approach such as the presumption of a shared history among

Deleted: The
Deleted: Avian
Deleted: :
Deleted:
Deleted: maintained in
Deleted: , eggs, and cell cultures
Deleted: id
Deleted:
Deleted: genes
Deleted: , or
Deleted: -
Deleted: ,

[3] Lab for virus isolation: Mona M Aly, Abdel S Arafa from National Laboratory for Quality Control on Poultry Production, Dokki (Cairo), Egypt .

Lab for sequencing: Isabella Monne, Alice Fusaro, Giovanni Cattoli, Ilaria Capua from OIE/FAO Reference Laboratory for Avian Influenza and Newcastle Disease, Istituto Zooprofilattico Sper.le delle Venezie, Legnaro (PD), Italy.

various forms of data reflecting the history of the organismal lineages (Kluge, 1989). Therefore, we do not combine data from HA and NA into a single dataset.

Multiple alignment was performed on nucleotide sequences using CLUSTALW (version 2.09 Larkin et al., 2007) using the default settings (gap opening penalty 10, gap extension penalty 0.2, DNA transition weight 0.5). A small number of internal gaps in the alignment were redistributed such that the final alignment did not imply any breaks in the reading frame in ancestral sequences. The 5' and 3' ends of the alignment were trimmed using the start and stop codons as reference. Leading and trailing gaps that remained were replaced with question marks to indicate missing data. Internal gaps were treated as a fifth character state. These datasets are hereafter referred to as "HA1646" and "NA1335" datasets.

Deleted: ClustalW
Deleted: (
Deleted: '
Formatted: Font: Times New
Formatted: Font: Times New
Deleted: 5'.
Deleted: “
Deleted: “

Wallace and Fitch datasets

The HA nucleotide sequence dataset analyzed by Wallace and Fitch (2008) was recreated from the list of GenBank accession numbers available as supplemental data to their paper. Following the materials and methods section of Wallace and Fitch (2008), their HA sequence dataset was aligned using CLUSTALW with the default settings without any manual adjustment. Leading and trailing gap markers were replaced with question marks to indicate missing data. Internal gap markers were treated as a fifth character state. This dataset is hereafter referred to as the "HA482WF" dataset.

Deleted:)
Deleted: ClustalW (Larkin et al. 2007)
Deleted: characters
Deleted: datasets are

Phylogenetic analysis

Phylogenetic analysis was performed using TNT (Goloboff et al., 2008) in parallel and sequential modes. All trees were rooted with A/Chicken/Scotland/1959/H5N1 as an outgroup. This study focuses on the highly pathogenic Asian lineage of H5N1 that is typically assumed to have originated with A/Goose/Guangdong/1996. We used the older isolate to check this assumption and polarize geographic character state changes. A preliminary analysis was performed to find the length of presumed most parsimonious trees. The TNT command '`xmult = level 5 replications 100000`' was used for the preliminary runs on a single CPU. The argument '`level 5`' allows for very thorough searches and was used only to find the presumed minimum length for each dataset. The preliminary runs were interrupted after 24 hours if they failed to produce shorter trees than previously found.

Deleted: .
Deleted: Although this
Deleted: flu
Deleted: can be traced back
Deleted: , we use
Deleted: isolates
Deleted: root the phylogenies
Deleted: the optimizations of
Deleted: states.
Deleted: after 24 hours

To create a representative sampling of the space of MPTs, the presumed minimum tree length was used to guide a new search to find 1000 independent hits to minimum length. The TNT command '`xmult = level 0 hits 1000 giveupscore n`' was used, where '`n`' is the tree length of the most parsimonious trees found by the preliminary run.

To ensure that trees were well sampled, a cumulative series of consensus trees were calculated to check that a stable consensus had been reached. The first consensus tree was based on trees 0-9, the second on the first consensus plus trees 10-19, the third on the

second consensus and trees 20-29 […] until all trees found from sampling 1000 hits to minimum length have been included. Consensus trees were calculated using the TNT command '`nelsen* [startree] [endtree];`' which calculates a strict consensus of all trees in the given range and saves the consensus as the last tree in memory. The TNT command '`tnodes`' was used to count the number of nodes that each consensus tree contained to provide a metric of their stability. A consistent number of nodes in a consensus tree indicate that we have collected a pool of candidate trees that well represents topologies at presumed minimum length.

Character optimization

Transmission events were calculated using a modified version of the method developed by Slatkin and Maddison (1989). To find transmission events, different geographic regions are treated as character states of a single multistate character, and are mapped onto trees obtained from phylogenetic analyses using standard optimization methods (Farris, 1970; Fitch, 1971; Swofford and Maddison, 1987).

Deleted: The method assumes an island biogeographical model, where each location is equally accessible from every other location.

To make our results comparable with previous phylogeographic studies of H5N1, we use equal weight between geographical character states. However, if preferred, the multistate character can be coded as a Sankoff character with a transition matrix specifying the relative costs of each type of transition between character states (Sankoff and Rousseau, 1975).

We counted the minimum and maximum number of transmissions between all pairs of geographic character states. When a character state change is observed, this is interpreted as a possible transmission route of the virus. The state changes are directional, so results are expected with different numbers for transformations from state A to state B and state B to state A. Figure 1 illustrates how state changes are counted.

Geographic information for the molecular datasets was extracted from the 'TSeq_orgname' field in GenBank's TinySeq XML records. For each molecular dataset (HA1646, NA1335, and HA482WF), new geographic character matrices were created. Each geographic character matrix contained an entry for a multistate character for each taxon in the nucleotide dataset. The order of taxa was maintained in nucleotide and geographic character datasets as this is important for consistency when using compact tree format and other tree formats that use numerical taxon names in TNT.

Deleted: the GenBank

Deleted: record.

Deleted: and

Deleted: containing

Deleted: kept the same

Three geographic character data sets of 31 character states were used to find transmission routes. Due to the limitations of the character state space allowed by the tree search software, TNT, we were unable to issue a single set of geographic character states suitable for comparing the HA1646 and NA1335 datasets to some previous studies. However, we were able to compare our results to Wallace and Fitch (2008). The first geographic character set we created, termed "W&F", was used to assign each taxon to a geographic category matching those used by Wallace and Fitch (2008). For taxa collected in localities not matching any of the 28 character states used in W&F, we assigned

Deleted: phylogenetic

Deleted: all

character states to these taxa with the most fitting of three new states: "Other China", "Other Asia" and "Other World". We created a second character set termed "Africa" that focuses on relationships and sources of H5N1 in regions of Africa and Europe. A third character set "World" was used to look at broader patterns among regions of the globe. The character set World treats mainland China as a single entity, treats Europe as four regions (north, east, south and west), and all of Africa as a single region.

Deleted: OtherChina", "OtherAsia

Deleted: OtherWorld". The

Deleted: "

Deleted: "

Deleted: North, East, South

Deleted: West

Counting transmission events

For each count of transmission events, a geographic character dataset containing a single multistate character and compact tree file containing the MPTs found during the search on the corresponding nucleotide data were loaded into memory. The TNT command `change` was used to calculate the maximum number of character state changes for pairs of geographic character states. The syntax used was '`change ]./[n]/[region X] [region Y]`'. The '`]`' indicates that results should be displayed character by character, and the '`./[n]/`' denotes that character number '`n`' should be optimized in all trees in memory. For example, the command '`change ]./0/HongKong Fujian`' will return the global minimum and maximum number of transmission events from Hong Kong to Fujian, calculated over all MPTs, when only a single multistate character is used. Each direction of change among states (e.g., A to B and B to A) was calculated for all possible pairs of states in the dataset.

Deleted: .

From the output of the '`change`' command, tables were created containing the global minimum and maximum values for each location pair and direction. When the number of state changes was >0, this result was interpreted as a possible transmission route from one region to another.

Transmission events were calculated for the HA428WF sequence dataset using the W&F character state set. For 1646HA and 1335NA sequence datasets, all three (W&F, Africa, and World) character state sets were used.

We have created a web-application, Routemap, (http://routemap.osu.edu) to produce visualizations of transmission events from the output of the geographic character optimization studies over pools of trees. The required inputs to Routemap are 1) a table of comma-separated values (CSV) containing the taxon label, the geographic place for the isolates, the latitude, and longitude in decimal degrees, and 2) a matrix of aligned nucleotide or amino acid data for the isolates in FASTA format. As an option, the user can input a set of precomputed trees. In the first step, Routemap will produce a TNT file that the user can execute on their machine with the command '`tnt p filename,`'. The user then will load the TNT output back into the Routemap via the browser, where the underlying application will complete the geographic calculations. Routemap's final output is a Keyhole Markup Language (KML) file compatible with Google Earth (http://earth.google.com) and other geographic information systems such as ESRI's ArcGIS Explorer (http://www.esri.com/software/arcgis/explorer/index.html). A visualization produced by Routemap is similar to a web-based airline route map. When

Deleted: wrote

Deleted: computer program in Java

Deleted: www.java.com

Deleted: input

Deleted: the program is two tables

Deleted:). One table contains

Deleted: potential

Deleted: sources

Deleted: disease agent in rows

Deleted: geographic destination for the disease agent in columns

Deleted: each cell contains the number of character transformation events between two localities discovered by the character optimization process represented by an integer. Another table contains the latitudes and longitudes of all the source and destination localities in the first table expressed

Deleted: Java program's

Deleted: ESRI's

Deleted:)

using a Routemap in a virtual globe the user selects localities of interest and viral transmission events implied by the analysis of phylogenetic trees are displayed as radiating from origins or arriving at destinations. The user can make one or more origin and destination selections to make as complex or as simple of a visualization as needed. The lines indicating routes are colored to indicate directional, ambiguous or bidirectional travel. Example datasets, KML files, scripts and instructions for users to make their own Routemaps are available at http://routemap.osu.edu.

Deleted: NASA's Worldwind (http://worldwind.arc.nasa.gov). The KML displays possible

Deleted: a source or going to a

Deleted: . Colors are used

Deleted: the direction of travel

Deleted: Datasets

Deleted: people.mbi.ohio-state.edu/rhovmoller/

Deleted: /.

Results

Alignment and phylogenetic trees

The alignment of the HA1646 dataset consisted of 1646 taxa and 1716 characters. 937 characters in HA1646 were parsimony-informative. The length of the presumed most parsimonious trees was 7284 steps. A search for 1000 independent hits to the MPT length resulted in 2721 unique trees. The number of nodes in the consensus stabilized at 500 trees at 695 nodes.

The alignment of the NA1335 dataset consisted of 1335 taxa and 1410 characters. 744 characters in NA1335 were parsimony-informative. The length of the presumed most parsimonious trees was 4795 steps. A search for 1000 independent hits to the MPT length resulted in 2649 unique trees. The number of nodes in the consensus stabilized at 860 trees at 540 nodes.

The HA482WF dataset consisted of 482 taxa and 1809 characters. 937 characters in HA482WF were parsimony-informative. Under our search strategy, the length of the presumed most parsimonious trees is 3257 steps. A search for 1000 independent hits to the MPT length resulted in 2470 unique trees. The consensus stabilized at 251 nodes after 140 trees were collected.

Reanalysis of Wallace and Fitch (2008)

Results of the reanalysis of Wallace and Fitch's (2008) HA nucleotide dataset (HA482WF) show that using a single tree, and a single optimization path, misestimates the frequency of transmission events. Moreover the use of a single tree can fail to detect possible transmission events. Figure 2 illustrates a comparison of the transmission events reported in Wallace and Fitch's (2008) with our reanalysis of HA428WF and W&F character data for localities. We used the global maximum number of transmissions from the pool of 2470 MPTs in our reanalysis. The global maximum of implied transmission events corresponds to Wallace and Fitch's use of DELTRAN optimization (Swofford and Maddison, 1987; Swofford, 1990). Thus a transmission route is indicated if it is found in any MPT.

Deleted: and in some cases fails

Deleted: among regions.

Deleted: most closely

Deleted:) optimization.

There are 756 possible transmission routes in W&F character state data. We compare the results found by Wallace and Fitch's (2008) optimization of these character data over a single tree they derived from HA482WF sequence dataset to our reanalysis using 2470 MPT for HA482WF. 517 of the possible routes are not found by either study, one transmission route result is found only in Wallace and Fitch (2008), 169 new routes were

discovered by our reanalysis, and 69 routes are suggested by both Wallace and Fitch (2008) and our reanalysis. Detailed results are presented in supplemental data at http://routemap.osu.edu/publications.

Deleted: people.mbi.ohio-state.edu/rhovmoller/

Deleted: /.

We also analyzed W&F based on all the MPTs we considered for HA428WF or HA1646 sequence datasets. Of 756 possible routes in W&F, 528 routes were not found when optimizing these character data over MPTs we found for the HA428WF or HA1646 sequence datasets. Four transmission routes are found only in HA482WF, 158 routes are found only in HA1646, and 61 routes occur in both HA482WF and the HA1646 dataset. Thoughtful examination of transmission routes implied by the HA1646 sequence dataset required addition of three character states for viral isolates from regions not included in the HA428WF dataset. 111 additional transmission routes were detected in MPTs implied by HA1646 when we included three localities not considered by Wallace and Fitch, (2008) but considered by other authors (e.g., Ducatez et al., 2006, 2007; Salzberg et al., 2007).

Deleted: .

Frequency of discovery of transmission routes

We compared the transmission routes found by Wallace and Fitch (2008) for a single tree to the pool of MPTs we derived from the HA428WF dataset. Wallace and Fitch (2008) did not find 22% of the transmission routes that are found in 95% or more of the MPT that we found for HA428WF. See supplemental data at http://routemap.osu.edu/publications for a summary table.

Visualization

The panels in Figure 3 contain visualizations produced with Routemap of viral traffic in Africa, east Asia, and southeast Asia from the HA1646 dataset and the World geographic character coding scheme. Visualizations of the HA1646 and NA1335 dataset under the W&F, World, and Africa geographic character coding schemes, as well as the HA428WF dataset are available at http://routemap.osu.edu/publications.

Deleted: Figure 3. contains a visual summary of the transmission routes concerning Indonesia, South Korea, Japan and Africa using our system for visualization.¶

Discussion

Phylogenetic analysis and H5N1

Metadata on pathogens such as date, host, and geographic location for viral isolation are often combined with phylogenetic trees based on pathogen sequence data by researchers and public health officials during an epidemic. In order to interpret metadata in a phylogenetic context, character optimization is a necessary tool. Studies that use metadata and character optimization include investigations of: where H5N1 spreads (Wallace et al., 2007, and Wallace and Fitch, 2008; Salzberg et al., 2007), where and when H5N1 has switched hosts (Janies et al., 2007) or experienced mutations that confer resistance to drugs (Hill et al., 2009).

Deleted: during

Deleted: examinations

Deleted: investigations on

Deleted:), and to answer questions as to where and when

Deleted: evolved in H5N1

Character optimization in phylogeographic analysis

Here we demonstrate that when reporting transmission events from character state changes, a single tree and a single way of optimizing the character will only tell a fraction of the story. Wallace et al., (2007) and Wallace and Fitch (2008) used MIGRAPHYLA (HoDac et al., 2007) and PAUP* (Swofford, 2008) to calculate transmission events, interpreted as character state changes using a single multistate character with geographical regions as character states. As Wallace and Fitch (2008) point out: "PAUP's assignments may represent only one of several possible most-parsimonious traces". Thus Wallace and Fitch (2008) inferred transmission events from a single possible optimization out of many, of a 28-state character, on a single tree, out of many possible trees. When we examine the length of this 28-state character on the pool of 2470 MPTs from the HA482WF sequence dataset, the lengths are between 107 and 118 steps. This range of lengths is consistent with the 115 steps reported in Figure S2 in Wallace and Fitch (2008). On the other hand, when we examine the average number of transmission events supported by the MPTs we found for the HA482WF dataset, the numbers of transmission events are very different. Each MPT supports on average 131 different transmission routes in contrast to the 70 unique transmission routes reported by Wallace and Fitch (2008).

The difference in results between TNT and PAUP* is because PAUP* only considers a single optimization path of the character, when calculating steps on polarized character state changes. This is the behavior of PAUP* even in cases where there are multiple equally parsimonious optimization paths for the character and each path may support

Deleted: ¶

Deleted: .

Deleted: MigraPhyla

Deleted: outline

Deleted: ¶
¶

Deleted: This difference in results is because PAUP* only considers a single optimization path of the character, when calculating steps on polarized character state changes. In contrast, in TNT it is possible to calculate the maximum (or minimum) state changes between two specified states of a multi state character, from multiple trees using multiple optimization paths. Although the sum of all state changes calculated this way will exceed the minimum length of the character, it is a more efficient method of finding possible directional state changes. It is evident that when reporting transmission events from character state changes, a single tree and a single way of optimizing the character will only tell a fraction of the story. We also discovered an artifact in the case of closely related taxa.

different sets of polarized state changes. In contrast, in TNT it is possible to calculate the maximum (or minimum) number of polarized state changes (e.g., from state A to state B) between two states of a multistate character. Moreover, in TNT this calculation can be done over multiple trees and multiple optimization paths per tree. Although the sum of all state changes calculated this way with TNT will exceed the minimum length of the character, it is important to find all possible state changes.

We also discovered an artifact in the case of closely related taxa. The number of observed state changes between two states should not necessarily be interpreted as the number of implied transmission events. In cases where the spread of an infectious agent is rapid, it is very likely that isolates from neighboring regions will be genetically identical or fail to have obtained any parsimony informative mutations. In such cases, the number of observed state changes in the MPTs in a clade with no supported branches (*sensu* the command `collapse 3` in TNT) will only reflect the proportion of taxa from each location, when optimizing over a pool of dichotomous, MPTs. In other words, given a character with two geographic states (location A and location B), the maximum number of transmission events between A and B will be the maximum length of the character in the subtree. For example, if there are 10 taxa from location A and five from location B in a subtree, the maximum number of transmission events will be five, when counting changes from A to B or from B to A.

Deleted: there are rapid spreads

Deleted: only carry

Deleted: non-

Deleted: differences.

Formatted: Font: Italic

Deleted: . Given

Deleted: place

Deleted: place

Deleted: either

Deleted: the other way around.

In cases where there are many sequences in a region where there is little or no genetic difference, the artifact we describe above can occur. In other cases, where distinct genetic

lineages have been exchanged between two places, counting the changes in geographic characters is a useful tool to understand viral traffic. Both the HA428WF and HA1646 sequence datasets contain large subclades (e.g., in Africa) of viruses with identical genetic sequences. The recent invasion of H5N1 into Africa without much molecular evolution in HA makes it of interest for future studies to examine these isolates from a multi locus approach as done by Salzberg, et al., (2007) and Janies et al., (2007). These concepts are also applicable to pandemic H1N1, which emerged in early 2009 and has spread rapidly around the world.

Current phylogenetics software (PAUP*, Swofford, 2008; TNT, Goloboff et al., 2008) limit a categorical character to 31 states. These programs are suitable for most analyses of morphological and sequence data in which the data are used to find trees. MESQUITE (Maddison and Maddison, 2007) can summarize character state changes over pools of trees. MESQUITE is thus an alternative for these analyses, but is limited to 56 states and would have to be coupled to a strong tree search program. POY4 (Varón et al., 2008) may provide a solution to the state space limitation as the user can define any alphabet to describe states for features of interest of pathogens and transformation costs among states in a look-up table (e.g., a Sankoff matrix). Heuristic efficiency will be important as an accurate regional map may require hundreds or thousands of geographical character states.

When a widely distributed taxon such as influenza A is studied, geographic character coding requires some abstraction of the globe into regions. We followed the 28-location

Deleted: of
Deleted: can be
Deleted: ¶ ¶
Deleted: .
Deleted: in (
Deleted: is
Deleted: Mesquite
Deleted:
Deleted: and
Deleted: tool that can be explored
Deleted: . POY
Deleted:
Deleted: .
Deleted: ; Sankoff and Rousseau, 1975
Deleted: This is an active area for future research, as trading of coarseness with which the world is divided into a grid to save computational space and time is a challenge for many modelers across the sciences (Lindley, 2009).
Deleted:

state coding scheme of Wallace and Fitch (2008), as well as implementing two alternative coding schemes (i.e. Africa and World) to make our results comparable to other studies. When comparing the HA1646 and NA1335 sequence datasets to the results of Wallace and Fitch (2008), 275 isolates were not compatible with the 28 states defined by Wallace and Fitch (2008). We coded the isolates that fell outside of the character state space with one of three new character states: Other Asia, Other China, and Other World. This coding scheme allowed our analyses to find transmission routes outside the scope of Wallace and Fitch's (2008) original publication, while retaining comparable results.

Deleted: 'Africa'
Deleted: 'World'
Deleted: other
Deleted: other
Deleted: other
Deleted: allows

By including more isolates from all regions sampled by HA1646 and NA1335 sequence datasets, we identified more possible routes of transmission than Wallace and Fitch (2008). For example, when using the same character data as W&F, our results agree with those of an independent study, Wang et al., (2008), in finding a single unambiguous transmission from the Hunan province of China to Indonesia. Wallace and Fitch (2008) suggested a transmission from Guangdong, China to Indonesia. The Guangdong to Indonesia route is not observed under any optimization of character data in W&F across the MPTs we found for the HA1646 sequence dataset.

Deleted: found
Deleted: routes, as implied by the data,
Deleted: .
Deleted: MPT

In all studies of the phylogeography of influenza there are problems associated with the imprecise naming of viruses. For example in the HA1646 sequence dataset, 18 viruses are labeled "China" which represents an area of 9,598,08 km^2. This area is comparable to 10,180,000 km^2 for all of Europe, including Russia west of the Ural Mountains. Other viral isolates are named for cities. In some cases viral isolates are named for regions.

Deleted: “
Deleted: ”
Deleted: is
Deleted: mountains.

Many workers, including Wallace have suggested that GenBank records for viral isolates should include latitude and longitude records to avoid geographic imprecision (Butler, 2008). Similarly, there is much room for improvement in metadata standards for naming of host taxa (Janies et al., 2007). In the wake of the H1N1 pandemic, GISAID has vastly improved metadata standards. High quality metadata will improve research in viral character evolution and geography studies.

We used the results from HA1646 and NA1335 sequence datasets, and character states from the World coding scheme to find transmission routes discussed in the literature. The major difference between W&F and World is that mainland China is treated as a single region in the World character state set, and countries in Europe and Africa are separate into regions (discussed above). Consolidation of regions within China was necessary to allow all countries (as defined by political borders) in east Asia to fit into the 31 character state limit.

Deleted: 'world'

Deleted: states

Current state of the H5N1 epidemic

Since 2003, H5N1 has been endemic in Indonesia. All phylogenetic studies including HA sequences isolated from H5N1 in Indonesia have found a monophyletic clade of Indonesian isolates, independent of optimality criterion (neighbor joining: Guan et al., 2004; Smith et al., 2006; Salzberg et al., 2007, Bayesian: Wang et al., 2008; Smith et al., 2006, maximum likelihood: Kilpatrick et al., 2006; Wallace and Fitch, 2008; Hill et al., 2009, parsimony: Janies et al., 2007; Salzberg et al., 2007; Wallace et al., 2007).

It is apparent that the H5N1 strains within Indonesia occurred due to a single transmission event from southern China. Janies et al., (2007) use whole genomes to demonstrate this link. Guan et al., (2004) use HA data to suggest the closest relatives to isolates from Indonesia are viruses from Yunnan province in China. Wallace et al., (2007), and Wallace and Fitch (2008) also us HA data to find that Guangdong province in China is the most likely source. These results stand in contrast to Kilpatrick et al., (2006), who used data from international trade to assert that the source of H5N1 in Indonesia was wild birds imported from Thailand. Our analyses of sequence and geographic data for H5N1 are in disagreement with the assertion of Kilpatrick et al, as we never find a Thailand to Indonesia route. Our analyses of the HA1646 sequence dataset, suggest that the geographic origin of H5N1 for Indonesia was Hunan province in China. Wang et al., (2008) concur. In our analyses of HA1646 sequence data, the Hunan-Indonesia migration route appears in all MPTs independent of optimization method.

Similarly our analyses of the NA1335 sequence dataset suggest Hunan as a possible source for Indonesia but some of the trees for NA1335 also indicate Hong Kong and the Guangxi Zhuang Autonomous Region of China as possible sources for Indonesia. Conflict between genes result could represent a reassortment event in which viruses in Indonesia have HA and NA genes with distinct histories or could simply reflect ambiguity in the data.

Deleted: endemic
Deleted: singe
Deleted: While
Deleted: .
Deleted: China's
Deleted: ,
Deleted: .
Deleted:) as well as
Deleted: use
Deleted: indicate
Deleted: to be
Deleted: Pathogen
Deleted: is
Deleted: this
Deleted: as
Deleted: ¶ ¶
Deleted: source
Deleted: into
Deleted: appears to be
Deleted: .
Deleted: In
Deleted: , all MPTs contain
Deleted: . Some
Deleted: ,
Deleted: ,
Deleted: This
Deleted: viruses
Deleted: be
Deleted: a reflection of

Smith et al. (2006), found that all H5N1 from Vietnam, Cambodia and Thailand form a single clade, except for one isolate from Vietnam. We were unable to find this clade structure in any tree found among the most parsimonious trees from the analyses of the HA1646 or NA1335 datasets. Although most isolates from Vietnam, Thailand, Laos, Cambodia and Vietnam are found in a monophyletic clade, isolates from Vietnam and Thailand are always found outside of this clade. This result suggests that the results of Smith et al. (2006) are an artifact of limited taxon sampling.

In the pool of trees resulting from searches of the HA1646 sequence data, we see at least three separate introductions of H5N1 to Japan. There were transmissions events from China to Japan in 2003 and 2008. Under certain optimizations, Vietnam is also a possible source of the 2008 isolates in Japan. In contrast, the source of the 2004 outbreak in Japan appears to be South Korea in all trees and all optimizations based on analysis of HA1646 sequence data and the W&F character data.

These comparisons illustrate the danger in excluding data. Wallace et al. (2007) and Wallace and Fitch (2008) both find the China-Japan connection, but excluded isolates collected in 2002-2003 from South Korea despite the fact these data were available in public databases as of 2005. Due to this exclusion, Wallace et al. (2007) and Wallace and Fitch (2008) did not find South Korea as a source of the Japanese outbreak in 2004. Our phylogenetic analysis indicating transmission of H5N1 from South Korea to Japan in

Deleted: .
Deleted:) and Guan et al. (2004)
Deleted: isolates
Deleted: southeast Asia (
Deleted: , Laos, Malaysia, Myanmar, Singapore,
Deleted: and Vietnam)
Deleted: monophyletic clade. In contrast we are
Deleted: support for
Deleted: resulting
Deleted: searches on HA6146
Deleted: sequence data. It is true that most isolates from southeast Asia are found in a single clade in both the HA and NA consensus trees, but we find in all trees a few
Deleted: Malaysia
Deleted: Vietnam
Deleted: Laos placed external to the main clade comprised of
Deleted: southeast Asia. This leads us to believe
Deleted: findings
Deleted: .
Deleted:) and Guan et al. (2004
Deleted:
Deleted: mainland
Deleted: from
Deleted: 'W&F'
Deleted: .
Deleted: .

2004 is in agreement with Kilpatrick et al., (2006) who based their inference on trade and bird migration data.

The H5N1 outbreaks in Africa have been investigated using HA sequences (Ducatez et al., 2006) as well as full genomes (Ducatez et al., 2007, Salzberg et al., 2007). However, these studies were conducted when there was a limited set of putative outgroup(s) for the African lineages in the public domain. These studies conclude that the sources of H5N1 are not southeast Asia, but rather central Asia and Europe. The pools of trees resulting from our analysis of HA1646 agree with Ducatez et al., (2006; 2007) that H5N1 has been introduced to Nigeria on three separate occasions. Using the "Africa" character state set, we find corroboration for the sources to be central Asia or Europe rather than southeast Asia. By examining at all trees and all optimizations we found for HA1646 and the W&F character state set, we can not rule out the source of one transmission to Africa was the Qinghai province of China (Salzberg et al., 2007).

Sources for H5N1 in Europe are unclear [4], as multiple possible sources are indicated for most countries. The basis for the ambiguity is the high similarity of the isolates found in Europe, Africa, and the Middle East in 2005-2006. This is another empirical example of the artifact we described above, where very low genetic variation leads to every region being a possible source for its neighbors.

[4] One important exception is the case of an infected eagle (*Nisaetus nipalensis* Hodgson, 1836) that was smuggled from Thailand to Belgium but did not cause any further spread (Van Borm et al., 2005).

Deleted: .
Deleted: both
Deleted: non-
Deleted: isolates
Deleted: Both
Deleted: .
Deleted: 'Africa'
Deleted: of not southeast origin but
Deleted: .
Deleted: middle east
Deleted: an

Perils of using a single tree

A single tree is likely to miss several transmission routes found frequently among the MPTs, as illustrated in Fig. 4. All of the publications on phylogeographic analysis of H5N1 we have reviewed here use a single tree for the results they report. This practice is independent of whether they use Slatkin and Maddison (1989) transmission event analysis (Wallace et al., 2007; Wallace and Fitch, 2008) or other optimization methods [POY3 (Wheeler et al., 2005) command `-diagnose` in Janies et al., 2007; MESQUITE command `trace:parsimony optimzation` in Salzberg et al., 2007] or observe transmission events by inspection (e.g., Ducatez et al., 2006, 2007; Wang et al., 2008).

Deleted: All of the publications on phylogeographic analysis of H5N1 we have reviewed here use single tree for the reported results. This practices is independent of whether they use Slatkin and Maddison (1989) transmission event analysis (Wallace et al., 2007; Wallace and Fitch, 2008) or observe transmission events by inspection (e.g. Ducatez et al., 2006, 2007; Wang et al., 2008), and which optimality criteria they use in tree search. In some cases (e.g. Salzberg et al., 2007; Smith et al., 2006; Wang et al., 2008), only a single neighbor-joining tree is produced.

The use of one tree is also independent of which optimality criteria researchers use in tree search. In some cases (e.g., Smith et al., 2006; Wang et al., 2008) only a single neighbor-joining tree is produced. Janies et al., (2007) created a binary tree on which to optimize based on a strict consensus of MPT using the TNT command `randtree*`. In other cases, information from multiple trees is represented as a consensus tree. A consensus tree can be a majority rules consensus from bootstrap resampling under maximum likelihood (Kilpatrick et al., 2006; Smith et al., 2006), parsimony (Guan et al., 2004; Salzberg et al., 2007;), or a distribution of trees from a Bayesian analysis (Smith et al., 2006).

Deleted: In other cases, information from multiple trees are condensed into a consensus tree, this

Deleted:),

There are several problems with the use of a consensus tree in the context of character optimization. First, a consensus tree does not represent all MPTs. Also the consensus

Deleted: trees

Deleted: a

topology is never one of the most parsimonious trees. Furthermore, trees can be constructed that do not change the consensus tree, but still are not found among the MPTs. Finally, a set of most parsimonious trees can be found, where the strict consensus tree would imply transmission routes that are never found in any of the individual MPTs, (as illustrated in Figure 4). It is thus inappropriate to optimize character state changes on a consensus tree.

Deleted: a shape that was not found during the tree search itself.

Under parsimony, each tree in the pool of MPTs and each optimization of the character states are just as optimal as the next one. The transmission routes found in only a few trees and under certain optimizations are just as supported by the optimality criterion as those trees and optimization found frequently. However, under a model-based optimality criterion, probability calculations for the existence of transmission routes could be calculated. A distribution of trees from a Bayesian analysis, where each tree has a posterior probability could be used to provide confidence values for transitions between the character states in a multistate character for geography optimized on trees of various posterior probability. Alternatively the proportional likelihood can be calculated for ancestral geographic states. One could employing a method such as described to study the evolution morphological characters under likelihood (Lewis, 2001) and is implemented in MESQUITE (Maddison and Maddison, 2003).

Deleted: Each

Deleted: is

Deleted: Similar to the fallacy of interpreting percentage in majority rules consensus as support for clades, the rarer

Deleted: ,

Deleted: in

Problems with using multiple trees

When examining character state changes over a pool of MPTs, different types of optimization need to be taken into account. If one examines the global minimum of state

Deleted: When we look at

changes over a pool of trees, they will find only those state changes that appear in all trees, under all optimizations. When dealing with large numbers of trees, a tree will often be found where a specific state change does not occur. For example, one has a subtree with only unsupported branches and 100 taxa from location A, 100 from location B and a single taxon from location C. When looking at this subtree in the MPTs, one can find a tree and an optimization pathway that does not have a transmission event between location A and B, making the global minimum number of transmission events from A to B equal zero.

In contrast, use of the global minimum number of polarized state changes dramatically changes the number of indicated transmission routes for a set of trees when compared to use of the maximum number of state changes. For example, when considering the maximum number of directional state changes on the HA1646 dataset, we find a total of 267 different transmission routes. This results contrasts with a result of 20 transmission routes found when looking at the global minimum number of directional state changes. In summary, the global minimum is a severe test of putative transmission routes as the route has to appear in every tree in the pool of MPTs to register a global minimum > zero. The global maximum is useful to find the range of possible transmission routes a set of MPTs implies, irrespective of how frequently a route is found.

Conclusions

Phylogenetic tree search and character evolution analyses have been valuable tools in many fields including surveillance of infectious diseases. In this context, parsimony is

Deleted: (transmissions)

Deleted: we

Deleted: and short branch lengths, for nearly every type of state change

Deleted: can

Deleted: when looking at the minimum number of state changes.¶ ¶ Say you have

Deleted: you

Deleted: 0

Deleted: When optimizing the 'world' geographic character state set on HA1646 under global minimum number of state changes we find only 20 supported transmission routes (directional changes from one state to another). In contrast, when looking at the same character state set and pool of trees, but looking for any transmission in any tree when maximizing the number of state changes we find 267 routes. Wallace and Fitch (2008) found a total of 70 transmission routes, when using DELTRAN optimization (Swofford and Maddison, 1987). ¶

Deleted: , geographic,

Deleted: . Here we demonstrate that

Deleted: analysis,

sometimes used for evaluation of trees and almost always used for character optimization. The multiple equally optimal trees produced by parsimony reflect of the natural ambiguity of the sequence data. Character optimization under parsimony can find the global minimum and maximum values for state changes in a pool of trees reflecting various degrees of ambiguity underlying transmission routes for organismal lineages. With efficient character optimization and tree search heuristics it is possible to find and analyze very large pools of trees for datasets with thousands of terminal taxa. Moreover, because results are sensitive to tree choice, any tree search method that considers multiple trees has an advantage over methods that consider a single tree, such as neighbor joining.

The methods of character optimization and visualization used and developed here are also applicable to sets of trees produced by maximum likelihood and Bayesian methods. Maximum likelihood methods are commonly used for tree search and sometimes preferred for the operational reason that they support one tree slightly more than near suboptimal candidate trees. While Bayesian methods produce a distribution of trees, conclusions are typically discussed in relationship to the single tree presented (often a majority rules consensus and sometimes a maximum *a posteriori* tree). As we have demonstrated here, a single tree, as well as consensus trees are insufficient for estimating transmission routes of infectious disease. In addition to considering multiple trees, when using Slatkin and Maddison's (1989) method of geographic analysis, it is important to take into account sensitivity to different strategies for character state optimization.

Deleted: , and tree search algorithms that employ heuristics enabling analysis of large datasets are key technologies for infectious disease surveillance and research. Parsimony is computationally fast and with tree search heuristics can find very large pools of candidate trees that form a representative sample of the space of possible topologies and the hypotheses they support.¶
¶

Deleted: with respect to the questions of interest.

Deleted: .

Deleted: we use

It is our intention that the methods described in this paper should be general and useful for phylogeographic analysis of any type of problem above and beyond infectious diseases. We welcome everyone to our website http://routemap.osu.edu to test our methods and create KML visualization of your own data.

Acknowledgements

We thank the Pablo Goloboff, Steve Farris, Kevin Nixon, and the Willi Hennig Society for making TNT available. Pablo Goloboff, Universidad Nacional de Tucumán and Diego Pol, CONICET, Museo Paleontológico Egidio Feruglio provided insights and discussions on character analysis. We also thank two anonymous reviewers for their constructive criticisms.

We acknowledge the Department of Biomedical Informatics and The Mathematical Biosciences Institute of The Ohio State University (OSU) for space, funding, and administrative support. We thank the Medical Center Information Services team of OSU and the Mathematical Biosciences Institute for hosting computing clusters used in this study. We acknowledge this material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and Office under grant number W911NF-05-1-0271.

Deleted: authors

Deleted: Tucuman

Deleted: for

Deleted:

Deleted:). We thank The

Deleted: Medical Center Information Services team of OSU and the Mathematical¶
Biosciences Institute hosts for hosting computing clusters used in this¶
study. We acknowledge this material is based upon work supported by, or in¶
part by, the U.S. Army Research Laboratory and Office under grant number¶
W911NF-05-1-0271. ¶

Deleted: .

Literature cited

Bright, R.A., Medina, M., Xu, X., Perez-Oronoz, G., Wallis, T.R., Davis, X., Povinelli, L., Cox, N., Klimov, A. , 2005. Incidence of adamantane resistance among influenza A (H3N2) viruses isolated worldwide from 1994 to 2005: A cause for concern. 2005. Lancet 366,1175-1181.

Butler, D. 2008. Politically correct names given to flu viruses. Nature News online 23 April 2008. doi:10.1038/452923a .

Chan, P. Outbreak of Avian Influenza A(H5N1) Virus infection in Hong Kong in 1997 (2002). Clin. Infect. Dis. 34 (Suppl 2), S58–64.

Chen, H., Bahri, S., Chen, Y., Cheung, C.L., Duan, L., Fan, X.H., Guan, Y. Guo, C.T., Hassan, S.S., Huang, K., Leung, Y.H.C., Li, K.S., Lu, H.R., Naipospos, T.S.P., Nguyen, T.D., Peiris, J.S.M., Qin, K., Rayner, J.M., Smith, G.J.D., Vijaykrishna, D., Wang, J., Webster, R.G., Wu, W.L., Xia, N.S., Xu, K.M., Yuen, K.Y., Zhang, J.X., Zhang, L.J. 2006. Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control. PNAS. 103, 2845-2850.

De, B. K., Brownlee, G. G., Kendal, A. P., Shaw, M. W. 1988. Complete sequence of cDNA clone of the hemagglutinin gene of influenza A/Chicken/Scotland/59 (H5N1)

virus: comparison with contemporary North American and European strains. Nucleic Acids Res. 16, 4181-4182.

Ducatez, M.F., Ammerlaan, W., de Landtsheer, S., Fouchier, R.A.M., Muller, C.P., Niesters, H.G.M., Olinger, C.M., Osterhaus, A.D.M.E., Owoade, A.A. 2006. Avian Flu: Multiple introductions of H5N1 in Nigeria. Nature 442, 37.

Ducatez, M. F., Ammerlaan, W., De Landtsheer, S., Fouchier, R. A. M., Muller, C. P., Olinger, C. M., Osterhaus, A. D. M. E., Ouedraogo, J. B., Owoade, A. A., Sow, A., Tahita, M. C., Tarnagda, Z. 2007. Molecular and antigenic evolution and geographical spread of H5N1 highly pathogenic avian influenza viruses in western Africa. J. Gen. Virol. 88, 2297-2306.

Farris, J.S. Methods for computing Wagner trees. 1970. Syst. Zool. 19, 83-92.

Fitch, W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20, 406-416.

Goloboff, P. A., Farris, J.S., Nixon, K.C. 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774-786

GISAID, 2009. Sample provided by Aly, M. M., Arafa, A. S. National laboratory for quality control on poultry production, Dokki (Cairo), Egypt. Sequence submitted by

Monne, I., Fusaro, A., Capua, I., Cattoli, G. OIE/FAO Reference Laboratory for avian influenza and Newcastle disease, Istituto Zooprofilattico Sper.le delle Venezie, Legnaro (PD), Italy.

Guan, Y., Dyrting, K.C., Ellis, T. M., Kong, K. F., Peiris, Y., Shortridge. K. F., Sit, T., Zhang, L. J. 2002. H5N1 influenza viruses isolated from geese in southeastern China: evidence for genetic reassortment and interspecies transmission to ducks. Virology 292, 16-23.

Guan, Y., Auewarakul, P., Buranathai, C., Chaisingh, A., Chen, H., Duan, L., Estoepangestie, A.T.S., Hanh, N.T.H., Li, K.S., Long, H.T., Nguyen, T.D., Peiris, J.S.M., Poon, L.L.M., Puthavathana, P., Rahardjo, A.P., Shortridge, K.F., Smith, G.J.D., Wang, J., Webby, R.J., Webster, R.G., Xu, K.M., Yuen, K.Y. 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. Nature 430, 209-213.

Hatchette, T.F., Baker, A., Johnson, C., Pryor, S. P., Walker, D., Webster, R.G. 2004. Influenza A viruses in feral Canadian ducks: extensive reassortment in nature. J. Gen. Virol. 85, 2327-2337.

Hill, A., Guralnick, R. Wilson, M., Habib, F. Janies, D. 2009. Evolution of drug resistance in multiple distinct lineages of H5N1 avian influenza. Infection, Genetics, and Evolution. 9, 169-178.

HoDac, H., Fitch, W.M., Lathrop, R.H., Wallace, R.G. 2007. MigraPhyla: Statistical analysis of migration events through a phylogeny. Version 1.0b manual. http://pd.bio.uci.edu/ee/WallaceR/MigraPhyla.html

Holland, J., Grabau, E., Horodyski, F., Nichol, S., Spindler, K., VandePol, S. 1982. Rapid evolution of RNA genomes. Science 215, 1577-1585.

Huelsenbeck, J.P and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Janies, D., Hill, A.W., Habib, F., Guralnick, R., Waltari, E., Wheeler, W.C. 2007. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). Syst. Biol. 56, 321-329.

Kawaoka Y., Alexander D.J., Nestorowicz A., Webster R.G. 1987. Molecular analyses of the hemagglutinin genes of H5 influenza viruses: origin of a virulent turkey strain. Virology 158, 218-227.

Kilpatrick, A. M., Chmura, A.A., Gibbons, D.W., Fleischer, R.C., Marra, P.P., Daszak, P. 2006. Predicting the global spread of H5N1 avian influenza. PNAS. 103, 19368-19373

Kluge, A.G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst. Zool 38, 7–25.

Larkin, M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.

Lewis, P.O. 2001. A likelihood approach to inferring phylogeny from discrete morphological characters. Syst. Biol. 50:913-925.

Lindstrom, S.E., Hiromoto, Y., Nerome, R., Omoe, K., Sugita, S., Yamazaki, Y., Takahashi, T., Nerome, K. 1998. Phylogenetic analysis of the entire genome of influenza A (H3N2) viruses from Japan: evidence for genetic reassortment of the six internal genes. J. Virol. 72, 8021-8031.

Lindstrom, S.E., Cox, N.J., Klimov, A. 2004. Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957-1972: evidence for genetic divergence and multiple reassortment events. Virology 328, 101-119.

Maddison, W. P. and D.R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org

Matrosovich, M., Kawaoka, Y., Webster, R., Zhou, N. 1999. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distiguishable properties. J. Virol. 73, 1146-1155.

Formatted: Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers

Deleted: Lindley, D. 2009. Calculating the Future. Communications of the ACM 52, 9-11.¶
¶

Deleted: http://mesquiteproject.org

Formatted: Font: Courier New, 10 pt

Obenauer, J.C., Denson, J., Fan, Y., Finkelstein, D.B., Hoffmann, E., Krauss, S., Ma, J., Mehta, P.K., Mukatira, S., Naeve, C.W., Rakestraw, K.M., Su, X., Wang, J., Webster, R.G., Xu, X., Zhang, Z., Zheng, J. 2006. Large-scale sequence analysis of avian influenza isolates. Science 311, 1576-1580.

Salzberg, S.L., Ali, A.S.M., Aly, M.M., Brown, I.H., Capua, I., Cattoli, G., Couacy-Hymann, E., De Mia, G.M., Dung, D.H., Ghedin, E., Guercio, A., Janies, D.A., Joannis, T., Kingsford, C., Osmani, A., Padalino, I., Saad, M.D., Saviæ, V., Sengamalay, N.A., Spiro, D.J., Yingst, S., Zaborsky, J., Zorman-Rojs, O. 2007. Genome analysis linking recent European and African influenza (H5N1) viruses. Emerg. Infect. Dis. 13, 713-718.

Sankoff, D. and Rousseau, P. 1975. Locating the vertices of a Steiner tree in arbitrary space. Mathematical Programming, p. 240246.

Sawabe, K., Hayashi, T., Hoshino, K., Hotta, A., Isawa, H., Kobayashi, M., Kurahashi, H., Saito, T., Sasaki, T., Tanabayashi, K., Tsuda, Y., Yamada, A. 2006. Detection and isolation of highly pathogenic H5N1 avian influenza A viruses from blow flies collected in the vicinity of an infected poultry farm in Kyoto, Japan, 2004. Am. J. Trop. Med. Hyg. 75, 327-332.

Slatkin, M. and Maddison, W. P. 1989. A cladistic measure of gene flow inferred from the phylogeny of alleles. Genetics 123, 603-613.

Smith, D., Lapedes, A., de Jong, J., Bestebroer, T., Rimmelzwaan, G., Osterhaus, A., Fouchier, R. 2004. Mapping the Antigenic and Genetic Evolution of Influenza Virus. Science 305, 371-376.

Smith, G.J.D., Bui, N.A., Chen, H., Cheung, C.L., Dao, T.V., de Jong, M.D., Farrar, J., Guan, Y., Hassan, S.S., Hien, T.T., Leung, Y.H.C., Li, K.S., Naipospos, T.S.P., Nguyen, T.D., Nguyen, T.V., Nguyen, V.C., Peirvis, J.S.M., Poon, L.L.M., Rayner, J.M., Usman, T.B., Vijaykrishna, D., Webster, R.G., Zhang, J.X., Zhang, L.J. 2006. Evolution and adaptation of H5N1 influenza virus in avian and human hosts in Indonesia and Vietnam. Virology 350, 258-268.

Subbarao, K., Klimov, A., Katz, J, Regnery, H, Lim, W, Hall, H, Perdue, M, Swayne, D, Bender, C, Huang, J, Hemphill, M, Rowe, T, Shaw, M, Xu, X, Fukuda, K, Cox, N. 1998. Characterization of an avian influenza a (H5N1) virus isolated from a child with a fatal respiratory illness. Science 279, 393-396.

Swofford, D. L., 1990. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0. Manual. Illinois Natural History Survey, Champaign, Ill.

Swofford. D.L. 2008. PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D.L. and Maddison, W.P. 1987. Reconstructing ancestral character states under Wagner parsimony. Math. Biosci. 87, 199-229.

USDA, 2007. Release No. 0296.06. Avian influenza. Low pathogenic H5N1 vs. highly pathogenic H5N1. Latest update, 23 July 2007. http://www.usda.gov/2006/08/0296.xml

Van Borm, S., Boschmans, M., Decaestecker, M., Dupont, G., Hanquet, G., Lambrecht, B., Snacken, R., Thomas, I., van den Berg, T. 2005. Highly pathogenic H5N1 influenza virus in smuggled Thai eagles, Belgium. Emerg. Infect. Dis. 11, 702-705.

Wallace, R.G., HoDac, H., Lathrop, R.H., Fitch, W.M. 2007. A statistical phylogeography of influenza A H5N1. PNAS 104, 4473-4478.

Wallace, R.G. and Fitch, W.M. 2008. Influenza A H5N1 immigration is filtered out at some international borders. PLoS ONE 3, e1697. doi:10.1371/journal.pone.0001697.

Wang, J., Bahl, J., Chen, H., Smith, G.J.D., Duan, L.,.Guan, Y. Peiris, J. S. M., Vijaykrishna, D., Webster, R. G., Zhang, J. X. 2008. Identification of the progenitors of Indonesian and Vietnamese avian influenza A (H5N1) viruses from southern China. J. Virol. 82, 3405-3414.

Webster, R.G., Bean, W.J., Chambers, T.M., Gorman, O.T., Kawaoka, Y. 1992. Evolution and ecology of influenza A viruses. Microbiol. Rev. 56, 152-179.

Wheeler,W. C., Delaet, J., Gladstein, D., and Varón, A., 2005. POY (version 3.012). Phylogeny reconstruction via optimization of DNAand other data. http://research.amnh.org/scicomp/projects/poy.php

WHO, 2004. Avian influenza A(H5N1)- update 31: Situation (poultry) in Asia: need for a long-term response, comparison with previous outbreaks. Reported to WHO. 2 March 2004. http://www.who.int/csr/don/2004_03_02/en/

WHO, 2009. Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1) Reported to WHO. 11 August 2009. http://www.who.int/csr/disease/avian_influenza/country/cases_table_2009_08_11/en/index.html

Varón, A., Vinh, L. S., Bomash, I., Wheeler, W. C. 2009. POY 4.1.1. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php

Xu X., Cox, N.J., Guo, Y., Subbarao K. 1999. Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. Virology 261, 15-19.

Zamarin, D., Ortigoza, M.B., Palese, P. 2006. Influenza A Virus PB1-F2 Protein Contributes to Viral Pathogenesis in Mice. Journal of Virology 80: 7976-7983.

Deleted: 14 January

Deleted: 01_14

Deleted: ¶
World Health Organization/World Organisation for Animal Health/Food and Agriculture Organization H5N1 Evolution Working Group. Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1). 2008. Emerg. Infect. Dis. Available from http://www.cdc.gov/EID/content/14/7/e1.htm doi: 10.3201/eid1407.071681¶

Deleted: ¶

Deleted: ..... *Figure captions*¶ ¶ **Figure 1:** ¶

Formatted: Swedish (Sweden)

Zhou, N. N., Shortridge, K.F., Claas, E.C.J., Krauss, S.L., Webster, R.G. 1999. Rapid evolution of H5N1 influenza viruses in chickens in Hong Kong. J. Virol. 73, 3366-3374.

Figure captions

Figure 1.

The same tree is shown with two different optimizations of a geographic character. The tree on the left shows one transmission from location A to location B and one from location B to location A. The tree on the right has two transmission events from location A to location B, and zero transmission events from location B to location A. The global maximum number of transmissions are thus: location A to location B: 2; location B to location A: 1.

Deleted: first
Deleted: second
Deleted: transmissions
Deleted: none¶
Deleted: hence
Deleted: ,

Figure 2.

Deleted: :

A comparison of transmission routes found by Wallace and Fitch (2008) and our reanalysis of the WF482HA dataset. Each cell in the grid represents a transmission route between two locations in the 28 state geographical character set used in the original publication. Cells colored black indicate transmission routes that were not found in any analysis. Cells colored yellow indicate that our reanalysis agrees with Wallace and Fitch (2008). Cells colored cyan indicates that the transmission route was not found in our reanalysis. Cells colored red indicate transmission routes found in the reanalysis, but not in Wallace and Fitch (2008).

Deleted: &
Deleted: Black indicates
Deleted: A
Deleted: cell indicated
Deleted: the original,
Deleted: and
Deleted: that it is
Deleted: the original study.

The rows represent outgoing locations and the columns incoming. For example, the cell in the Vietnam column and the Guangdong row represents a transmission from Guangdong to Vietnam, while the cell in the Guangdong column and the Vietnam row represents a transmission from Vietnam to Guangdong

Figure 3.

Migration routes indicated by the HA1646 dataset.

A. Transmissions into Nigeria, using the Africa character set. The orange lines represent transmission routes are bidirectional or ambiguous. In H5N1 in Africa, this result is due to very low genetic variation between the H5N1 in Nigeria and neighboring regions.

B. Transmissions into Indonesia, using the "W&F" character set. Analyses indicate only one source of H5N1 in Indonesia, Hunan province in China.

C. Transmission into Japan, using the "World" character set. Both China and South Korea are sources for the H5N1 in Japan. The direction of the transmission route to Vietnam is either bidirectional or ambiguous.

Figure 4:

Deleted: :
Deleted: 'Africa'
Deleted: represents
Deleted: that
Deleted: either
Deleted: ,
Deleted: where the direction is indeterminable.
Deleted: case, this
Deleted: an artefact of
Deleted: Neighboring
Deleted:
Deleted: 'Wallace & Fitch'
Deleted: the H5N1in
Deleted: 'World'
Deleted:
Deleted: not possible to determine.

The consensus tree can imply transmission routes never found among the most parsimonious trees.

When an unordered multistate character with states 0, A, B and C is optimized on trees a. and b., no optimal optimization path has the directional state change C to A. However, considering optimal optimization paths in the strict consensus tree, one path has the node above the outgroup assigned to state B and thus two changes from B to C, as well as two changes from C to A are required.