SCALABLE COMPUTING: PRACTICE AND EXPERIENCE
Volume 8, Number 2, pp. 143–146. http://www.scpe.org
ISSN 1895-1767

LARGE-SCALE PHYLOGENETIC ANALYSIS FOR THE STUDY OF ZOONOSIS AND ASSESSMENT OF INFLUENZA SURVEILLANCE

DANIEL A. JANIES*, PABLO A. GOLOBOFF†, AND DIEGO POL‡

Abstract. We performed a phylogenetic analysis of 2359 hemagglutinin sequences of influenza A. We find multiple host shifts of all polarities among swine, humans, and birds. We also describe novel methods to assess the quality of surveillance and apply these methods to the public sequence record.

Key words. influenza, avian influenza, zoonoses, host-parasite relationship, phylogeny, hemagglutinin, epidemics, genome, surveillance, high performance computing

1. Introduction. Influenza viruses are one of the major threats to public health, as shown by the recent outbreak of H5N1 [1]. Anitbody surveillance of the hemagglutinin (HA) and neuraminidase (NA) is the lens through which the World Health Organization monitors the spread of influenza A. High throughput genomic sequencing and phylogenetic analyses will soon eclipse HA and NA antibody screening [2]. These analyses are important to make public health decisions such as the identification of animal reservoirs, vaccine design, and pandemics warnings [3]. However, most analyses have focused on disjoint subsets of influenza isolates. Here we use a comprehensive sample of 2359 HA sequences ranging from 1902 to present, isolated from human, avian, several mammal hosts, and representing 16 antigenic subtypes (figure 1.1).

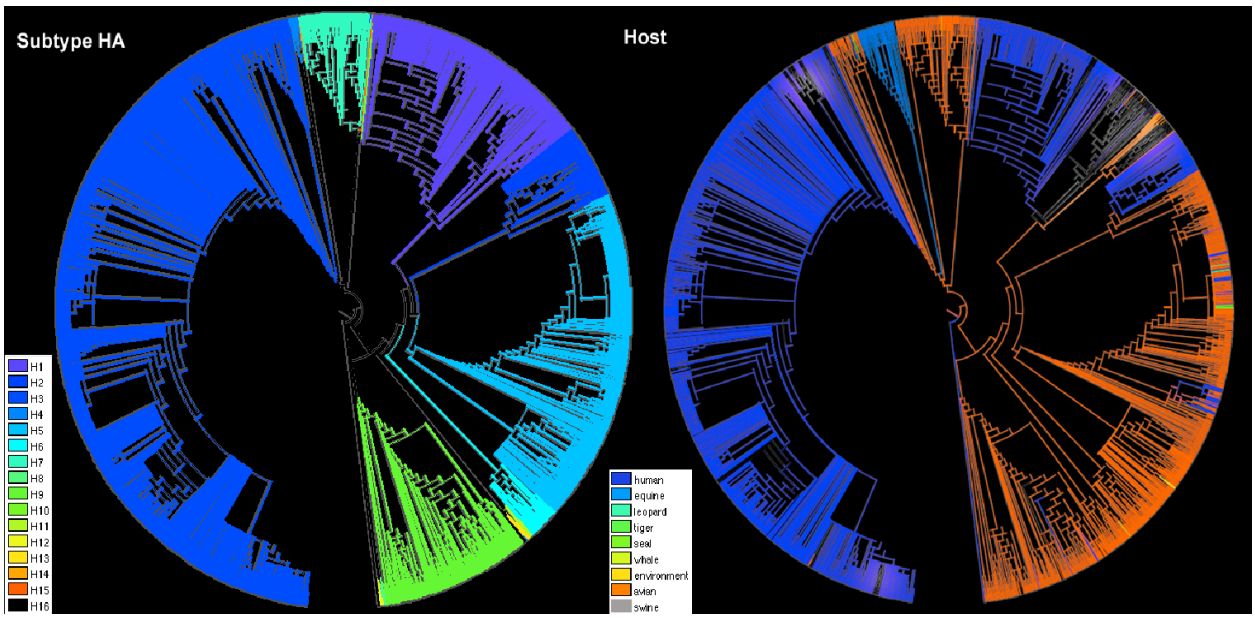

FIG. 1.1. *Two color-coded character optimizations on the same strict consensus tree for hemagglutinin (HA) sequences representing 2358 isolates of influenza A, with an influenza B outgroup at the root. The colors tracing the branches of the tree on the left depict the optimization of the character "HA antigenic subtype". The colors tracing the branches of the tree on the right represents optimization of the character "host". Color legends are provided at the lower left of each tree.*

Phylogenetic analysis of large datasets is difficult because the cost of computation scales combinatorially with the number of isolates. Large-scale phylogenetic analysis is now possible via synergistic development of heuristic tree search and parallel computing [4, 5, 6]. Large trees enable longitudinal analysis of the spread of strains of influenza, zoonotic events associated with outbreaks, and quality of surveillance as revealed by sequence databases.

*Department of Biomedical Informatics, The Ohio State University, Graves Hall, 333 W. 10th Av. Columbus, OH, 43210 USA.
† Instituto Superior de Entomología, Consejo Nacional de Investigaciones Científicas y Técnicas, Miguel Lillo 205, 4000 San Miguel de Tucumán, Argentina.
‡Museo Paleontologico Egidio Feruglio, Consejo Nacional de Investigaciones Cientificas y Téchnicas; Argentina.

TABLE 2.1

Estimated number of transformations between different states in the host character optimized using parsimony. The upper section of the table represents an estimate of minimum possible host shift events in a set of most parsimonious trees. The lower section of the table represents an estimate of the maximum possible host shift events in a set of most parsimonious trees.

Host shifts min.		TO						
		avian	human	swine	equine	leopard	tiger	seal
FROM	avian		18	16	2	0	2	1
	human	4		20	0	0	0	0
	swine	4	7		0	0	0	1
	equine	0	0	0		0	0	1
	mouse	0	0	0	0	0	0	0
	leopard	0	0	0	0		0	0
	tiger	1	0	0	0	0		0
	seal	0	0	0	0	0	0	
Host shifts max.		TO						
		avian	human	swine	equine	leopard	tiger	seal
FROM	avian		27	19	3	2	5	1
	human	13		26	0	0	0	0
	swine	6	12		0	0	0	1
	equine	0	0	0		0	0	1
	mouse	0	0	0	1	0	0	0
	leopard	1	0	0	0		1	0
	tiger	3	0	0	0	1		0
	seal	0	0	0	0	0	0	

2. Results and Discussion.

2.1. Host Shifts. Although rare, direct transmission of avian strains to which human populations have little protective immunity can lead to outbreaks. Such is the case of the 1997 Hong Kong and current H5N1 outbreak, which are strictly avian in origin [7]. After the discovery of receptors for both avian and mammalian strains of influenza in the trachea of pigs, it has been hypothesized that domestic swine act as intermediate hosts in which human and avian viruses can exchange gene segments to make a chimeric descendent [8]. Such reassortant viruses are thought to have lead to the pandemics of 1957 and 1968 [9]. However, little is known about the relative frequency of avian to human and swine to human host shifts. Using large-scale phylogenetic analysis, we show that avian to human shifts are more common than swine to human shifts. Moreover, we find multiple host shifts of all polarities among these three major host groups (table 1.1 and supplemental figure 1.1).

2.2. Surveillance Quality. Influenza surveillance has been characterized as biased towards sequencing of antigenically rare isolates that may signal the need to update the vaccine [9]. Here we adapt metrics [10] designed to measure the quality of the fossil record to assess influenza surveillance quality over time and geography. The metrics quantify discordance between dates of viral isolation and the branching pattern influenza A in the tree. In addition, we visualize the time between the date of sampling of an isolate and its minimum possible date of origin with branch lengths scaled to reflect the difference (supplemental figure 1). Our results show that surveillance is good overall in that it is significantly different from a random expectation (table 2.1). However, the relative quality of surveillance varies among geographic regions and periods of time. An interesting example of variable surveillance quality occurs in two sister groups of pathogenic avian influenza: H5N2 viruses that infect birds in the Americas and H5N1 viruses circulating in Eurasia that infect birds, humans, and other mammals. When the surveillance quality of these clades is compared, it becomes apparent that surveillance of the Eurasian clade is superior to surveillance of the Americas clade. These results imply that surveillance programs are failing to capture phylogenetic diversity of H5 sequences in the Americas when compared with the diversity of sequences surveyed for H5 in Eurasia.

TABLE 3.1

Results for Manhattan Stratigraphic Metric (MSM) (Pol and Norell, 2001) and Gap Excess Ratio (GER) (Wills, 1999) applied to H5 subtype hemagglutinin clades from Eurasia and the Americas sampled from two time periods (1975-2005 and 1990-2005). Cases in which there is poor correlation between the date of sampling of a given isolate and the inferred date of origin of the evolutionary lineage that underlies the isolate of interest indicates that the surveillance program is failing to closely monitor the origin and persistence of lineages of influenza. A score of 1 in these metrics reflects perfect correlation, and values < 1 to 0 represent diminishing surveillance quality.*

	MSM*	GER
Eurasia 1975-2005	0.53	0.98
Americas 1975-2005	0.12	0.85
Eurasia 1990-2005	0.60	0.94
Americas 1990-2005	0.30	0.79

3. Methods.

3.1. Sequence sampling. We sampled the public sequence databases with the following procedure. First we used representative hemagglutinin (HA) nucleotide sequences of each subtype as queries to BLAST [11] Genbank (`www.ncbi.nlm.nih.gov`) and the Influenza Sequence Database (`www.flu.lanl.gov`). Then we removed identical sequences, sequences with annotation indicating they were adapted to laboratory conditions, and sequences less than 987 nucleotides. Some short sequences between 287 and 987 nucleotides were included if they represented rare subtypes and historically important isolates.

Alignment was performed with CLUSTALW [12] with the following conditions GAPOPEN=3 GAPEXT=2. We then performed preliminary phylogenetic analyses using TNT and mapped host phenotypes and visualized surveillance gaps (search and visualizations described below).

Preliminary trees were inspected for subtrees that presented temporal gaps in surveillance or host shift activity, thus indicating that additional data was needed. Additional sequence data were acquired through searches of the public sequence databases via BLAST [11] using sequences drawn from the subtree of interest as queries.

Our final dataset consists of 2358 HA sequences of influenza A using an influenza B sequence (K00423) as an outgroup. Our dataset represents: worldwide influenza A sampling, isolates dating back to 1902, representatives from 16 HA subtypes and, isolates from 8 hosts, and the environment (supplemental figure 1).

3.2. Phylogenetic analyses. The aligned dataset of 2359 isolates was analyzed using the maximum parsimony criterion. The heuristic tree searches were conducted in TNT [6]. Due to the large size of the dataset, the analysis was conducted using parallel computing on a Beowulf cluster of 15 processors. The tree search strategy consisted of multiple replicates of the following procedure: random addition sequences followed by a combination of hill-climbing (tree bisection and reconnection), divide-and-conquer (sectorial searches), and simulated annealing (tree drifting) algorithms as described by Goloboff [13]. These replicates retrieved a set of trees that were subsequently input to a genetic algorithm (tree fusion). The entire process was replicated numerous times until a stable consensus tree was found. The minimum length found for this aligned dataset with equal weights for all substitutions and insertion deletion mutations is 39202 steps. This best-known minimum length score was achieved in 100 independent replicates. A strict consensus of this topology is presented in supplemental figure 1. The strict consensus tree has 1192 internal nodes.

Tree graphics were made with Mesquite (www.mesquiteproject.org) (figure 1.1). The average number of transformations between different states in the host character (table 1) was optimized using parsimony as described by Fitch [18].

In order to measure the robustness of clades we conducted jackknife analysis [14] assessing the frequency of recovery of every clade in 1000 resampling replicates. The results of the jackknife search are shown in supplemental figure 2. Numbers at nodes represent the percent of replicates that recovered that clade.

3.3. Phylogenetic metrics for surveillance quality. The public record of hemagglutinin sequences has been described as biased by efforts to characterize rare subtypes that may signal the need for a vaccine update [15]. However there may be additional sources of bias, such as the relative efforts of governments in various regions and in various time periods. In order to assess surveillance quality, here we apply measures entitled the Manhattan Stratagraphic Metric (MSM*) [16] and Gap Excess Ratio (GER) [17] to our tree.

Cases in which there is poor correlation between the date of sampling of a given isolate and the inferred date of origin of the evolutionary lineage that underlies the isolate of interest indicates that the surveillance program is failing to closely monitor the origin and persistence of lineages of influenza. A score of 1 in these metrics reflects perfect correlation, and values < 1 to 0 represent diminishing surveillance quality.

An extensive sampling of HA sequences, such as the one gathered for this study is therefore critical to comparatively assess quality of reporting in various regions and over periods of time. For example, the comprehensive phylogenetic tree we present (supplemental figure 1) allows us to comparatively evaluate how well the sampling of viruses of H5 subtype in Eurasia and the Americas reflects the phylogenetic diversification patterns.

In order to normalize the number of sequences evaluated in the measures of surveillance quality, we performed a 100 replicates of random deletion of terminal branches from the Eurasian H5 clade. In each replicate, the MSM* and GER were measured for the Eurasian H5 subtree derived from each of the most parsimonious trees (table 2).

4. Conclusions. Microorganisms that cause infectious diseases present critical issues of national security, public health, and economic welfare. For example, in recent years, highly pathogenic strains of avian influenza have emerged in Asia, spread through Eastern Europe and threaten to become pandemic. As demonstrated by the coordinated response to SARS and influenza, agents of infectious disease are being addressed via large-scale nucleotide sequencing projects such as the Influenza Genome Sequencing Project (`http://msc.tigr.org/infl_a_virus/index.shtml`). The goals of large collaborative sequencing projects are to rapidly put large amounts of data in the public domain to accelerate research on disease surveillance, treatment, and prevention. However, our ability to derive information from large comparative nucleotide sequence datasets lags far behind acquisition. Here we demonstrate that analysis of thousands of isolates is possible via synergistic application of heuristic tree search strategies and parallel computing. Among many uses of phylogenetic trees, here we demonstrate two new uses, longitudinal analyses of patterns of zoonotic transmission and assessment of surveillance quality.

REFERENCES

[1] WHO, *Avian influenza new areas with infection in birds update 34* `http://www.who.int/csr/don/2005_10_13/en/index.html` (2005).

[2] S. LAYNE, *Human influenza surveillance: the demand to expand*. Emerg. Infect. Dis. `http://www.cdc.gov/ncidod/EID/vol12no04/05-1198.htm` (2006).

[3] R. WEBSTER, W. BEAN, O. GORMAN, T. CHAMBERS, Y. KAWAOKA, *Evolution and ecology of influenza A viruses*. Microbiol. Rev. 56: 152–179. (1992).

[4] P. GOLOBOFF, *Analyzing large data sets in reasonable times: solutions for composite optima*. Cladistics. 15:415–428. (1999).

[5] D. JANIES, W. WHEELER, *Efficiency of parallel direct optimization*. Cladistics. 17: S71–S82. (2001).

[6] P. GOLOBOFF, J. FARRIS, K. NIXON, *T.N.T: Tree Analysis using New Technology*, `http://www.zmuc.dk/public/phylogeny` (2003).

[7] K. LI, Y. GUAN, J. WANG, G. SMITH, K. XU, L. DUAN, A. RAHARDJO, P. PUTHAVATHANA, C. BURANATHAI, T. NGUYEN, A. ESTOEPANGESTIE, A. CHAISINGH, P. AUEWARAKUL, H. LONG, N. HANH, R. WEBBY, L. POON, H. CHEN, K. SHORTRIDGE, K. YUEN, R. WEBSTER, J. PEIRIS, *Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia*. Nature. 430: 209-213. (2004).

[8] C. SCHOLTISSEK, Pigs as 'mixing vessels' for the creation of new pandemic influenza A viruses. Med. Principles Practice. 2: 65–71. (1990).

[9] R. BUSH, C. SMITH, N. COX, W. FITCH, *Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution.* Proceedings of the National Academy of Sciences U.S.A. 97: 6974–6980. (2000).

[10] D. POL AND M. NORELL, *Comments on the Manhattan Stratigraphic Measure.* Cladistics. 17: 285–289. (2001).

[11] S. ALTSCHUL, T. MADDEN, A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, D. LIPMAN, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res. 25: 3389–3402. (1997).

[12] J. THOMPSON, D. HIGGINS, T. GIBSON, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice*. Nucl. Acids Res. 22: 4673–4680. (1994).

[13] P. GOLOBOFF, *Analyzing large data sets in reasonable times: solutions for composite optima*. Cladistics. 15:415–428. (1999).

[14] S. FARRIS, V. ALBERT, M. KALLERSJÖ, D. LIPSCOMB, A. KLUGE, *Parsimony Jackknifing Outperforms Neighbor-Joining*. Cladistics. 12:99–124. (1996).

[15] R. BUSH, C. SMITH, N. COX, W. FITCH, *Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution.* Proceedings of the National Academy of Sciences U.S.A. 97: 6974–6980. (2000).

[16] D. POL, M. NORELL, *Comments on the Manhattan Stratigraphic Measure.* Cladistics. 17: 285–289. (2001).

[17] M. WILLIS, *Congruence between phylogeny and stratigraphy: Randomization tests and the Gap Excess Ratio.* Syst. Biol. 48:559–580. (1999).

[18] W. FITCH, *Toward defining the course of evolution: minimum change for a specific tree topology.* Systematic Zoology. 20: 406–416. (1971).

Edited by: Dazhang Gu
Received: Jan 16, 2007
Accepted: April 15, 2007