Avian Influenza Virus Exhibits Rapid Evolutionary Dynamics

Rubing Chen and Edward C. Holmes*†*

*Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University and
†Fogarty International Center, National Institutes of Health, Bethesda, Maryland

Influenza A viruses from wild aquatic birds, their natural reservoir species, are thought to have reached a form of stasis, characterized by low rates of evolutionary change. We tested this hypothesis by estimating rates of nucleotide substitution in a diverse array of avian influenza viruses (AIV) and allowing for rate variation among lineages. The rates observed were extremely high, at $>10^{-3}$ substitutions per site, per year, with little difference among wild and domestic host species or viral subtypes and were similar to those seen in mammalian influenza A viruses. Influenza A virus therefore exhibits rapid evolutionary dynamics across its host range, consistent with a high background mutation rate and rapid replication. Using the same approach, we also estimated that the common ancestors of the hemagglutinin and neuraminidase sequences of AIV arose within the last 3,000 years, with most intrasubtype diversity emerging within the last 100 years and suggestive of a continual selective turnover.

Introduction

Influenza A viruses have a wide host range and have been isolated from various animals, including humans, pigs, horses, and birds. However, ecological studies suggest that their natural reservoirs are wild aquatic birds, namely wild ducks, gulls, and shorebirds (Webster et al. 1992) and, with the exception of H5N1 highly pathogenic avian influenza virus, influenza viruses do not usually cause overt disease in these species. Influenza A viruses are categorized by their 2 surface antigens, the hemagglutinin (HA), of which there are 16 subtypes (H1–H16), and neuraminidase (NA), of which there are 9 (N1–N9) (Horimoto and Kawaoka 2005), all of which are maintained in aquatic bird populations. Along the migratory routes of these reservoir species, influenza A viruses can occasionally transmit to other bird species, creating self-sustaining epidemics and occasionally to mammalian hosts such as humans, although usually with little onward transmission.

A key concept in influenza virus evolution is that there is a marked difference in evolutionary dynamics between those viruses that infect aquatic birds and those from other host species. In particular, the viruses associated with wild aquatic birds are proposed to have reached an "evolutionary stasis" characterized by low rates of evolutionary change, particularly at amino acid–changing sites (Webster et al. 1992). According to this hypothesis, the evolutionary arms race between host and virus is less intense in avian than mammalian species, so that there is little selective requirement to repeatedly fix amino acid changes that evade host immune responses (Suarez 2000). It is this equilibrium that guarantees the perpetuation of influenza viruses in their natural hosts, such that they have reached a global fitness peak characterized by strong purifying selection. This model is supported by the low rates of amino acid change observed in nucleoprotein (NP) sequences from old world avian lineages sampled over 50 years (Gorman, Bean, et al. 1990). Similarly, low rates were observed in the avian PB2 protein (Gorman, Donis, et al. 1990), and previous studies have found that influenza viruses from wild aquatic birds are characterized by relatively low numbers of nonsynonymous to synonymous substitutions per site (d_N/d_S), indicative of strong selective constraints (Widjaja et al. 2004).

Although the hypothesis of evolutionary stasis has been central to studies of avian influenza virus (AIV), a comprehensive analysis of substitution rates, and the evolutionary processes that might determine these rates, is lacking. In particular, substitution rates have not been estimated using the full extent of available gene sequence data, and it is unclear whether rates differ among species or serotypes. Further, most previous studies have involved the linear regression of genetic distances against sampling time, therein providing an approximate estimate of substitution rates, and some have failed to distinguish between wild and domestic avian species (Suarez 2000). Similarly, the age of genetic diversity in AIV, as well as the timescale of viral evolution, remains obscure. Given the ongoing health risk posed by AIV, most notably H5N1, it is clearly important to accurately estimate the rates and dates of AIV evolution and determine what this means for the dynamics of AIV evolution. We therefore undertook an analysis of substitution rates and times of origin in a total of 3,147 sequences of AIV taken from 10 different viral subtypes (including H5N1) and a variety of avian species and employed a sophisticated Bayesian Markov chain Monte Carlo (MCMC) approach that allows for rate variation among lineages and a range of demographic histories (Drummond et al. 2006).

Materials and Methods

Data Preparation

Gene sequence data sets corresponding to the 8 genome segments of avian influenza A virus were compiled from GenBank. All laboratory recombinant or highly cultured sequences were excluded, as were those less than 500 bp in length. These data were further divided into subsets comprising different serotypes of AIV. Only serotypes with more than 20 sequences were subjected to further analysis. For each subtype, we also constructed data sets for individual avian hosts (chicken, domestic duck, and wild aquatic birds) when more than 20 sequences were available for these species. To avoid sampling bias, when multiple sequences were available from the same host in a particular location and a given year, a maximum of 4 sequences were included. Such sampling did not greatly affect parameter estimates. For computational tractability, all data sets were

Key words: avian influenza virus, relaxed molecular clock, emerging virus, H5N1, coalescent.

E-mail: ech15@psu.edu.

Mol. Biol. Evol. 23(12):2336–2341. 2006
doi:10.1093/molbev/msl102
Advance Access publication August 31, 2006

limited to 120 sequences, with sequences randomly removed from larger data sets. In total, we examined 3,147 viral sequences from 10 viral subtypes.

For each data set, sequence alignments were created using the MUSCLE (Edgar 2004) and ClustalX1.83 (Chenna et al. 2003) programs and then adjusted manually using the Se–Al program according to both amino acid and nucleotide sequences (Rambaut 1996), although in most cases few gaps had to be incorporated. All sequence alignments are available from the authors on request.

Estimating Substitution Rates and the Age of Genetic Diversity

Overall rates of evolutionary change (nucleotide substitutions per site, per year) and the age of the most recent common ancestor (MRCA) were estimated using the BEAST program (http://evolve.zoo.ox.ac.uk/Beast/), which employs a Bayesian MCMC approach, utilizing the number and temporal distribution of genetic differences among viruses sampled at different times (Drummond et al. 2002, 2006). To determine substitution rates as accurately as possible, AIV sequences were analyzed under a variety of models of demographic history, namely, constant population size, exponential population growth, expansion population growth and Bayesian skyline reconstruction, and with both strict (constant) and relaxed (variable) molecular clocks. A previous analysis of human influenza A virus evolution found that the uncorrelated exponential relaxed clock model provided a better fit to the data than the uncorrelated lognormal model (Drummond et al. 2006), and this was confirmed in a preliminary analysis of a subset of the data collected here (data not shown), although the parameter values estimated were similar under both models. All estimates also incorporated the General Time Reversible (GTR) + Γ_4 model of DNA substitution as this model, or its close relatives, was consistently the best supported in Modeltest (Posada and Crandall 1998; data not shown). Finally, all models were compared using Akaike information criterion (AIC), with uncertainty in the estimates reflected in the 95% highest probability density (HPD), and in each case chain lengths were run for sufficient time to achieve convergence.

Measurement of Selection Pressures

To determine the overall selection pressures faced by each gene in each serotype of AIV, we estimated the mean numbers of nonsynonymous substitutions (d_N) and synonymous substitutions (d_S) per site (ratio d_N/d_S) using the SLAC method within the HYPHY package (Kosakovsky Pond et al. 2005) and accessed through the Datamonkey interface (http://www.datamonkey.org). 156 H3N2 human influenza A virus isolates sampled from New York State during 1999–2005 (Holmes et al. 2005) were used for comparison. In all cases, d_N/d_S estimates were based on Neighbor-Joining trees under the GTR substitution model.

Analysis of Combined HA and NA Subtypes

To infer the phylogenetic relationships of the different subtypes of HA and NA, and to estimate their substitution rates and divergence times with as much accuracy as possible, we compiled a representative data set of 1–5 sequences for each subtype. To reduce the impact of multiple substitutions at single nucleotide sites (see below), this analysis was based on 2nd codon positions only, and all highly divergent positions where the alignment was compromised by insertions and deletions were excluded. This resulted in data sets of 110 taxa, 445 bp for the HA gene and 78 taxa, 366 bp for NA gene. Maximum likelihood (ML) trees for these data were estimated under the GTR + I + Γ_4 substitution model with parameters optimized from the empirical data (parameter values available on request). The robustness of each node was estimated using bootstrap resampling (1,000 replications) under the Neighbor-Joining procedure, with input genetic distances determined under the ML substitution model. Maximum pairwise genetic distances estimated in this manner were approximately 1.0 for 2nd codon positions in both HA and NA (~40% uncorrected), suggesting that widespread site saturation had not occurred, in contrast to the 1st and 3rd codon positions where maximum pairwise distances were approximately 2.0 and 12, respectively. All phylogenetic analyses were undertaken using the PAUP* package (Swofford et al. 2003). Average rates of nucleotide substitution and the age of the MRCA for 2nd codon positions in the HA and NA genes were estimated using the BEAST package as described above, employing a model of exponential population growth (as this was generally the best-fit model) and a relaxed (uncorrelated exponential) molecular clock. ML phylogenetic trees for each of the 6 internal genes, with viruses from different HA and NA subtypes combined, are provided in the Supplementary Material online.

Results

Our analysis of rates of nucleotide substitution is summarized in figure 1 (mean and 95% HPD values shown; full results are available in the Supplementary Material online). In all but one data set, the variable rate relaxed molecular clock (uncorrelated exponential) model was a significantly better fit to the data than that of a constant rate of evolutionary change across lineages. The one exception was the NA gene for the H4N6 data in which the strict and relaxed molecular clocks gave very similar posterior probabilities. Further, in the majority of cases, the data supported a model of exponential population growth, as expected under epidemic dynamics, although broadly similar rates and dates were observed under different models of demographic history. However, the most notable aspect of these results was that the substitution rates estimated for AIV were very high, ranging from 1.8 to 8.4 $\times 10^{-3}$ substitutions per site, per year (subs/site/year). These rates are similar to those previously estimated in human (5.7×10^{-3} subs/site/year in the HA1 domain; Fitch et al. 1997), equine (5.4×10^{-4} and 5.1×10^{-4} subs/site/year for the M and NS genes, respectively; Lindstrom et al. 1998), and swine (1.30×10^{-3} subs/site/year for the M gene; Lindstrom et al. 1998) influenza viruses and therefore indicate that AIV does not evolve anomalously slowly.

Also of note was that the age of the MRCA of each gene and serotype was generally extremely recent (range 15.4–196.0 years, with most less then 100 years; fig. 2),

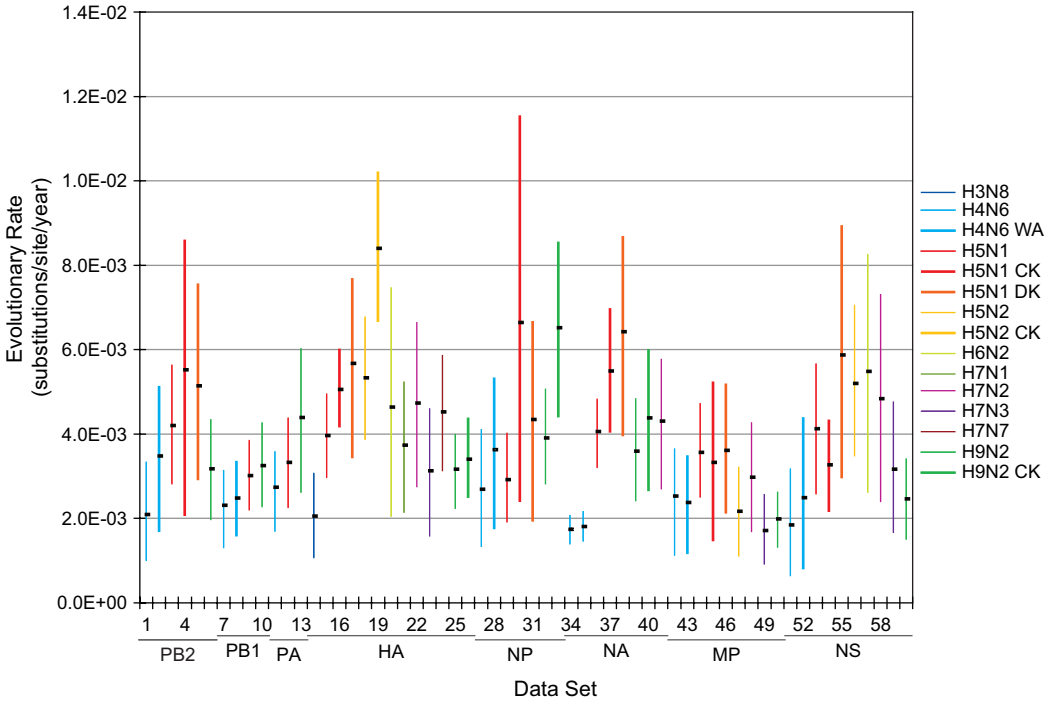

FIG. 1.—Rates of nucleotide substitution in avian influenza A viruses. Mean substitution rates are shown for each gene of each serotype, with 95% lower and upper HPD values shown as error bars. CK = chicken, DK = domestic duck, and WA = wild aquatic birds, shorebirds, and gulls. Data set numbers correspond to those listed in the Supplementary Material online.

with the exception of the NS gene (88.0–427.9 years), which exists as 2 divergent subtypes, A and B (Suarez and Perdue 1998). Segments with the estimated oldest MRCA values were always sampled from data sets containing bird species with different migration routes, such as East Asia/Australia and Atlantic/Americas. However, even in these cases, it is striking that the common ancestor of the American and Eurasian strains, which are usually

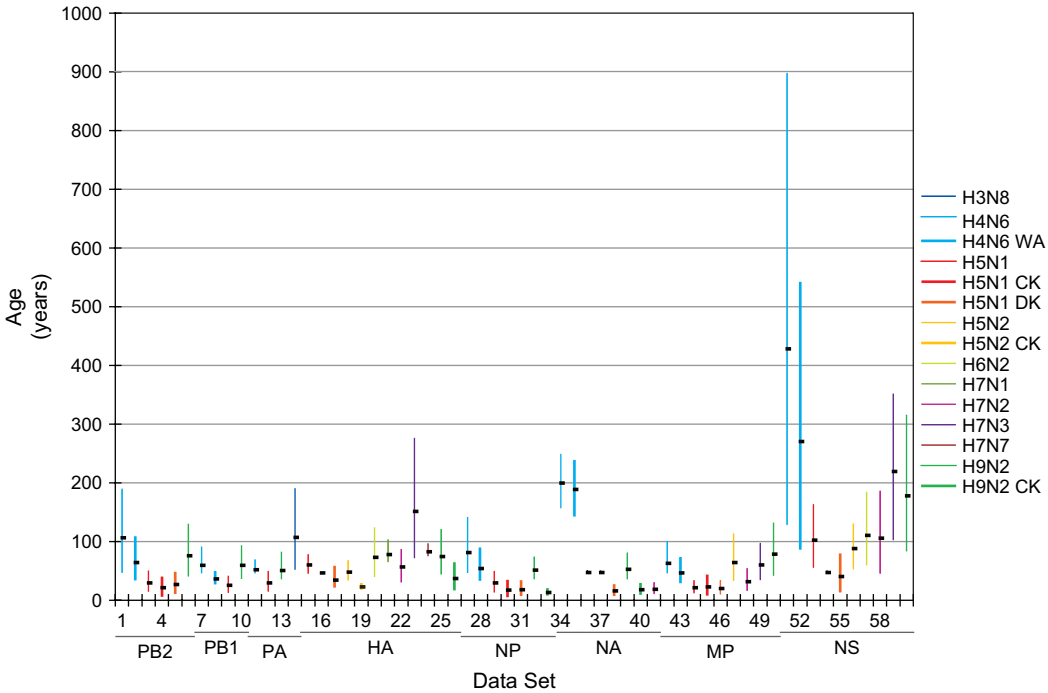

FIG. 2.—Age of the MRCAs of avian influenza A viruses. The estimated age for the MRCA of each gene from each serotype is given, with 95% lower and upper HPD values shown as error bars. CK = chicken, DK = domestic duck, and WA = wild aquatic birds, shorebirds, and gulls. Data set numbers correspond to those listed in the Supplementary Material online.

Table 1
Summary of Substitution Rates, Age of MRCA, and d_N/d_S for AIV

	Substitution Rate and 95% HPD (10^{-3} subs/site/year)			Range of Rate and Serotype (10^{-3} subs/site/year)		Range of MRCA Ages and Serotype (years)		
Gene	Mean	Lower	Upper	Lowest	Highest	Youngest	Oldest	d_N/d_S
Total	3.41	2.08	4.75	—[a]	—	—	—	—
CK, DK[b]	5.12	3.14	7.27	—	—	—	—	—
PB2	3.15	1.94	4.44	2.09 H4N6	4.20 H5N1	29.48 H5N1	106.03 H4N6	0.055
PB1	2.86	1.93	3.75	2.31 H4N6	3.25 H9N2	24.94 H5N1	59.29 H4N6	0.053
PA	3.48	2.20	4.66	2.73 H4N6	4.38 H9N2	29.43 H5N1	51.41 H4N6	0.062
HA	3.92	2.43	5.40	2.06 H3N8	5.33 H5N2	47.58 H5N2	151.1 H7N3	0.130
NP	3.17	2.03	4.39	2.69 H4N6	3.91 H9N2	29.05 H5N1	81.14 H4N6	0.055
NA	3.61	2.29	4.92	1.74 H4N6	4.05 H5N1	18.66 H7N2	196.03 H4N6	0.206
MP1	2.49	1.45	3.50	1.71 H7N3	3.56 H5N1	21.00 H5N1	78.49 H9N2	0.047
NS1	3.87	2.13	5.66	1.84 H4N6	5.48 H6N2	87.98 H5N2	427.9 H4N6	0.147

[a] —, not applicable.
[b] CK and DK refer to chicken and domestic duck, respectively. Total data sets (all avian species combined) were used in all other cases.

phylogenetically distinct, have recent MRCAs (47.6–196.0 years, excluding the NS gene). Finally, it is notable that segments taken from H5N1 viruses often have very shallow genetic diversity (21.0–101.9 years; table 2). However, these dates are still substantially older than the first appearance of H5N1 in humans in 1997, and although this subtype was first isolated in 1959, the majority of isolates analyzed were sampled after 1996.

To determine the factors that might affect the pattern of AIV evolution, we compared the average substitution rates, age of the MRCA, and d_N/d_S values for each gene (table 1). This revealed no significant difference in substitution rates among genes. Similarly, there was little difference in evolutionary rates among serotypes, including H5N1 viruses. The only notable difference in substitution rate was that average rates in poultry species (chickens and domestic ducks) were generally higher than those in all bird species combined, and in some cases (e.g., HA from H5N2), this difference is significant. However, the substitution rate in wild aquatic birds (in our study these comprise wild ducks, other wild waterfowl, shorebirds, and gulls) is not significantly different to that seen in other avian species, although sufficiently large data sets are only available for H3N8 and H4N6.

The selection pressures acting on AIV were also similar to those recorded in mammalian influenza A viruses (tables 1 and 2; Supplementary Material online) and to those previously reported in AIV (Obenauer et al. 2006) although lower values were seen in wild aquatic species (Widjaj et al. 2004). The highest d_N/d_S ratios were observed in the HA and NA (mean d_N/d_S 0.130 and 0.206, respectively), most likely reflecting immune selection pressure at a small number of amino acid sites (Horimoto and Kawaoka 2005), and also NS1 (mean d_N/d_S 0.147), which downregulates dsRNA-induced antiviral responses. In the case of H5N1, for which most data are available, selection pressures were very similar to those seen in human H3N2 viruses (table 2).

The relative consistency in substitution rates among genes, particularly in the HA and NA, allowed us to tentatively place these genes within an evolutionary time frame. The phylogenetic tree for the different subtypes of HA based on 2nd codon positions is shown in figure 3 (with the phylogeny of NA subtypes, as well as the internal genes, available in the Supplementary Material online). The mean substitution rates at the 2nd codon positions for these genes estimated using the same Bayesian MCMC approach as above were 4.565×10^{-4} subs/site/year for the HA gene and 4.073×10^{-4} subs/site/year for the NA gene. Under these rates, the estimated age of their common ancestors (i.e., the tree root) for both the HA and NA intersubtype phylogenies was surprisingly short and overlapping—1,415 years (95% HPD; 448–2,833 years) for the HA gene and 1,313 years (95% HPD; 391–2,782 years) for the NA gene—but compatible with some previous estimates (Suzuki and Nei 2002).

Discussion

The overall picture that arises from our analysis is that rather than forming a static gene pool, the evolution of AIV in all subtypes and species is characterized by the rapid accumulation of mutations, including those at nonsynonymous sites, typical of RNA viruses in general. Far from being in evolutionary stasis, AIV therefore evolves at a rate, $>1 \times 10^{-3}$ substitutions per site, per year, that is comparable to

Table 2
Comparison of d_N/d_S Values between Human H3N2 and Avian H5N1

Host	Subtype	PB2	PB1	PA	HA	NP	NA	MP	NS1
Human	H3N2	0.096	0.092	0.134	0.287	0.100	0.275	0.089	0.239
Avian	H5N1	0.102	0.088	0.103	0.179	0.076	0.243	0.116	0.202
	H5N1 CK[a]	0.085	—[b]	—	0.217	0.047	0.199	0.097	0.276
	H5N1 DK[a]	0.076	—	—	0.173	0.068	0.240	0.116	0.245

[a] CK and DK refer to chicken and domestic duck, respectively.
[b] —, not applicable.

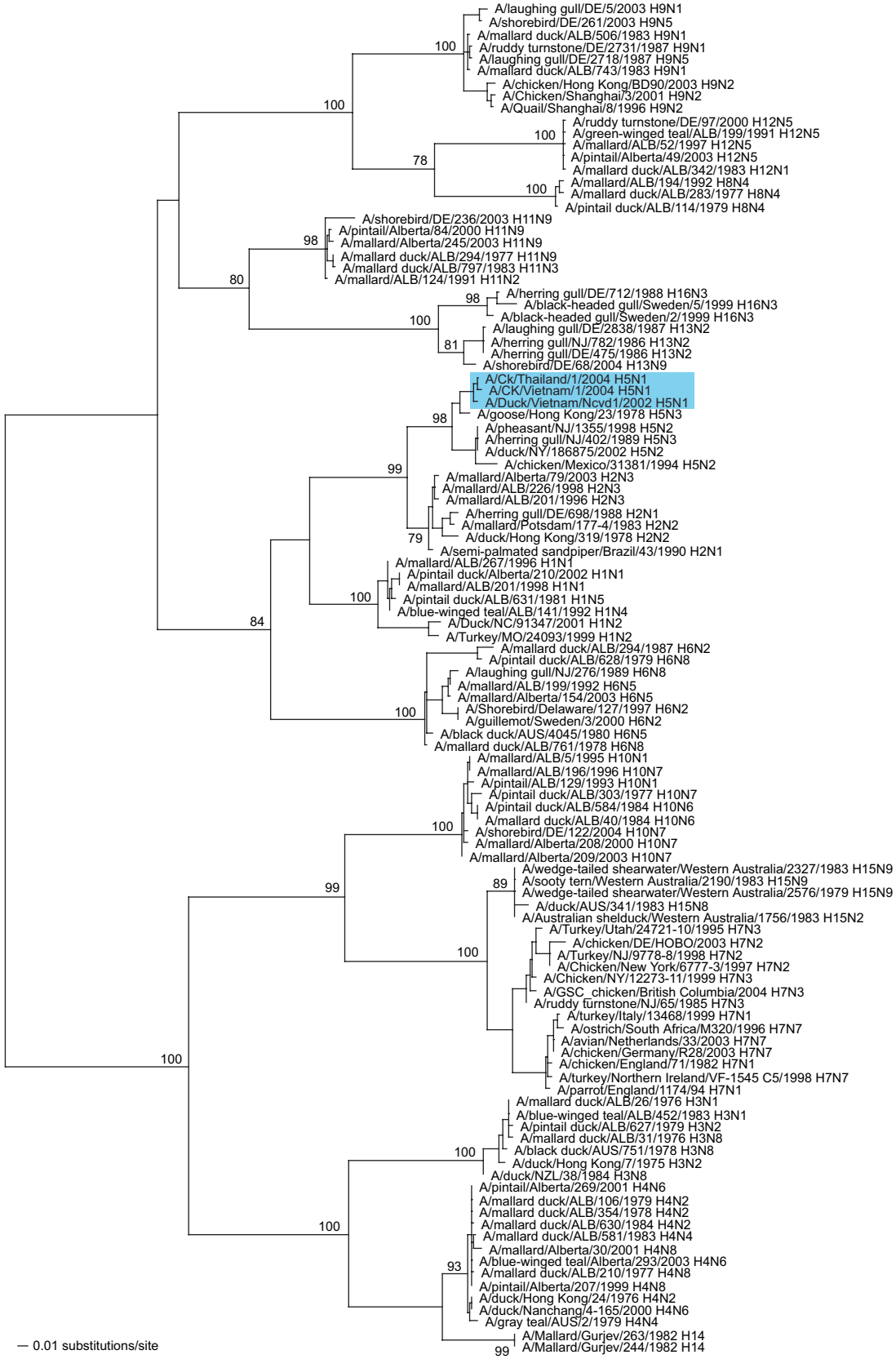

FIG. 3.—ML phylogenetic tree of different HA subtypes of AIV inferred using 2nd codon positions (110 taxa, 445 bp). Branch lengths are scaled according to the numbers of nucleotide substitutions per site. Bootstrap values (>75%) are shown for key nodes on the phylogeny. The grouping of H5N1 viruses is shaded.

that seen in other RNA viruses (Jenkins et al. 2002; Hanada et al. 2004), including those influenza A viruses isolated from mammals. Further, there is relatively little difference in substitution rates among genes or serotypes, indicating that their intrinsic dynamics of mutation and replication are similar among all species infected. That overall selection pressures are similar in the influenza A viruses sampled from birds and mammals also reveals that the former have not yet reached a global fitness peak characterized by little amino acid fixation. Finally, although there was some tentative evidence for higher substitution rates in those avian species characterized by the highest transmission rates—chickens and domestic ducks—rapid evolution was observed in all species and subtypes including those more often associated with wild aquatic species. Indeed, the branch lengths in the HA and NA trees, as well as those produced previously (Obenauer et al. 2006; Olsen et al. 2006), suggest that wild aquatic birds do not evolve at rates radically lower than those seen in poultry species, and it is likely that AIV regularly moves between wild aquatic and other bird species so that they do not constitute separate evolutionary cycles.

That the genetic diversity observed within individual subtypes is generally of an extremely recent origin, with the segments from most subtypes seemingly having arisen within the last 100 years, provides important clues to underlying evolutionary mechanisms. Given the much deeper divergence time among the subtypes, which may go back thousands of years, as well as the large number of infected individuals, such shallow diversity toward the tips of the tree is highly suggestive of population bottlenecks; these will have periodically purged genetic diversity, such that the variation currently observed has arisen since the last bottleneck event. As individual subtypes of AIV are also found in a variety of bird species, which would reduce the power of random evolutionary processes and extend neutral coalescent times, it seems likely that such bottlenecks are selective in nature, in which mutations of high fitness have periodically swept to fixation in each subtype eliminating preexisting genetic variation. In these circumstances, the different HA and NA subtypes are likely to represent only transient fitness peaks, which are subject to continual antigenic evolution but with little immunological interaction among them (Sharp et al. 1997). Under this model, rather than representing a pattern of slow endemicity, the evolution of AIV, of all subtypes and not only H5N1, is characterized by a dynamic epidemic turnover, manifest as extensive genetic diversity.

Supplementary Material

A supplementary table and 2 supplementary figures are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Literature Cited

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497–3500.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 161:1307–1320.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA. 94:7712–7718.

Gorman OT, Bean WJ, Kawaoka Y, Webster RG. 1990. Evolution of the nucleoprotein gene of influenza A virus. J Virol. 64:1487–1497.

Gorman OT, Donis RO, Kawaoka Y, Webster RG. 1990. Evolution of influenza A virus PB2 genes: implications for evolution of the ribonucleoprotein complex and origin of human influenza A virus. J Virol. 64:4893–4902.

Hanada K, Suzuki Y, Gojobori T. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol Biol Evol. 21:1074–1080.

Holmes EC, Ghedin E, Miller N, et al. (11 co-authors). 2005. Whole genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biol. 3:e300.

Horimoto T, Kawaoka Y. 2005. Influenza: lessons from past pandemics, warnings from current incidents. Nat Rev Microbiol. 3:591–600.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol. 54:152–161.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 21:2531–2533.

Lindstrom S, Endo A, Sugita S, Pecoraro M, Hiromoto Y, Kamada M, Takahashi T, Nerome K. 1998. Phylogenetic analyses of the matrix and non-structural genes of equine influenza viruses. Arch Virol. 143:1585–1598.

Obenauer JC, Denson J, Mehta PK, et al. (17 co-authors). 2006. Large-scale sequence analysis of avian influenza isolates. Science. 311:1576–1580.

Olsen B, Munster VJ, Wallensten A, Waldenstrom J, Osterhaus AD, Fouchier RA. 2006. Global patterns of influenza A virus in wild birds. Science. 312:384–388.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Rambaut A. 1996. Se-Al: sequence alignment editor [Internet]. Available from: http://evolve.zoo.ox.ac.uk/.

Sharp GB, Kawaoka Y, Jones DJ, Bean WJ, Pryor SP, Hinshaw V, Webster RG. 1997. Coinfection of wild ducks in influenza A viruses: distribution patterns and biological significance. J Virol. 71:6128–6135.

Suarez DL. 2000. Evolution of avian influenza viruses. Vet Microbiol. 74:15–27.

Suarez DL, Perdue ML. 1998. Multiple alignment comparison of the non-structural genes of influenza A viruses. Virus Res. 54:59–69.

Suzuki Y, Nei M. 2002. Origin and evolution of influenza virus hemagglutinin genes. Mol Biol Evol. 19:501–509.

Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. Microbiol Rev. 56:152–179.

Widjaja L, Krauss SL, Webby RJ, Xie T, Webster RG. 2004. Matrix gene of influenza A viruses isolated from wild aquatic birds: ecology and emergence of influenza A viruses. J Virol. 78:8771–8779.

William Martin, Associate Editor

Accepted August 24, 2006