
Laura Campitelli,1,† Massimo Ciccozzi,1† Marco Salemi,2 Fabiana Taglia,1 Stefano Boros,1 Isabella Donatelli1 and Giovanni Rezza1

Correspondence Laura Campitelli campitel@iss.it Massimo Ciccozzi ciccozzi@iss.it

1Department of Infectious, Parasitic and Immune-Mediated Diseases, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy
2Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida, Gainesville, FL 32610, USA

The selection pressure acting along the entire genome sequence of H5N1 avian influenza viruses isolated from several bird species and humans infected in the 1997 and 2004 outbreaks, and on the HA1 genes from H5N1 viruses isolated during the entire study period, in eastern Asia was evaluated. According to maximum-likelihood analysis, viral genes appeared to be, in both epidemics, under strong purifying selection, with only the PB2, HA and NS1 genes under positive selection. Specific codons under positive selection were detected by using codon-based substitution models. Positive-selection analysis performed on single-codon sites might be helpful in clarifying the driving force of avian and human influenza virus evolution and in selecting specific targets for vaccines and antiviral drugs.

Starting in 2004, an epidemic of a highly pathogenic avian influenza A (H5N1) virus infection among domestic poultry affected eight Asian countries, causing human disease in 108 cases, 54 of whom died, in Thailand, Vietnam and Cambodia (WHO, 2005a, b). This was the largest outbreak since an H5N1 influenza epidemic in birds, causing multiple bird-to-human transmissions, was identified in 1997 in Hong Kong (Claas et al., 1998).

Most human cases represent rare events of direct transmission from infected domestic birds, whereas only one instance of probable human-to-human transmission has been documented so far (Liem et al., 2005; Ungchusak et al., 2005). To be transmitted efficiently from human to human, a zoonotic virus must overcome a series of obstacles (impairment of viral entry and/or virus replication, inhibitory immune response, etc.) that may lead interspecies passage to a dead end (Webby et al., 2004).

To evaluate the possible occurrence of adaptive changes that could favour interspecies transmission of avian influenza, monitoring mutation events among poultry and humans is essential. Reassortment events, which involved all but the HA gene and gave rise to a dominant H5N1 genotype (Z) in chickens and ducks that was responsible for the recent Asian epidemic, have been fully identified (Li et al., 2004), whereas there is no clear-cut evidence about the occurrence and the role of stepwise mutations; however, sequencing data of isolates from northern Vietnam and Thailand suggest recent genetic changes in the virus (Butler, 2005). To contribute to the comprehension of viral evolution, we evaluated H5N1 sequences from birds and humans infected between the 1997 and 2004 outbreaks.

All sequences analysed in this study were downloaded from the Influenza Sequence Database (ISD; Macken et al., 2001). GenBank accession numbers are available in the Supplementary Table in JGV Online. As the HA gene did not undergo reassortment, analysis of H5N1 HA1 sequences included viruses isolated in eastern Asia between 1997 and 2004. Due to reassortment events that occurred during the study period, the analysis of the remaining gene segments was limited to the 1997 and January to March 2004 (first-wave) epidemics.

Phylogenetic and positive-selection analyses were performed to compare the ten coding regions of H5N1 isolates. Separate maximum-likelihood (ML) phylogenies were estimated for each gene region. The best-fitting nucleotide-substitution models were chosen with the hierarchical likelihood-ratio test (LRT) strategy described by Swofford & Sullivan (2003), as implemented in the MODELTEST program (Posada & Crandall, 1998). Selected models are shown in

†These authors contributed equally to this work.

A supplementary table showing GenBank accession numbers and supplementary figures showing phylogenetic trees of the HA1 gene are available in JGV Online.
### Table 1. Nucleotide-substitution models for the gene regions of 1997 and 2004 H5N1 influenza viruses

<table>
<thead>
<tr>
<th>Year/gene region</th>
<th>n</th>
<th>Length of sequences (bp)</th>
<th>Selected model</th>
<th>( \alpha ) parameter of ( \Gamma ) distribution</th>
<th>lnLK (M7)*</th>
<th>lnLK (M8)*</th>
<th>( P ) value†</th>
<th>Positively selected sites‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PB2</td>
<td>16</td>
<td>2247</td>
<td>HKY + ( \Gamma )</td>
<td>0·288</td>
<td>−4677·91</td>
<td>−4672·38</td>
<td>0·004</td>
<td>17 (R, V); 82 (N, S, R); 199 (A, S); 334 (K, S, N); 336 (F, S); 355 (K, Q, R); 727 (G, R)</td>
</tr>
<tr>
<td>PB1</td>
<td>20</td>
<td>2226</td>
<td>HKY + ( \Gamma + I )</td>
<td>0·91</td>
<td>−4427·59</td>
<td>−4427·39</td>
<td>0·82</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>18</td>
<td>1731</td>
<td>TrN93 + ( \Gamma )</td>
<td>0·29</td>
<td>−3656·33</td>
<td>−3656·31</td>
<td>0·98</td>
<td></td>
</tr>
<tr>
<td>NP</td>
<td>19</td>
<td>1496</td>
<td>TrN93 + ( \Gamma )</td>
<td>0·0062</td>
<td>−3032·92</td>
<td>−3032·63</td>
<td>0·75</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>25</td>
<td>1308</td>
<td>HKY + ( \Gamma )</td>
<td>0·58</td>
<td>−3204·63</td>
<td>−3204·63</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>18</td>
<td>756</td>
<td>HKY + ( \Gamma )</td>
<td>0·198</td>
<td>−1401·31</td>
<td>−1397·77</td>
<td>0·03</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>18</td>
<td>285</td>
<td>K80</td>
<td>1·74</td>
<td>−554·030</td>
<td>−554·030</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NS1</td>
<td>17</td>
<td>690</td>
<td>HKY + I</td>
<td>0·277</td>
<td>−1811·83</td>
<td>−1810·43</td>
<td>0·24</td>
<td></td>
</tr>
<tr>
<td>NS2</td>
<td>17</td>
<td>363</td>
<td>F81 + ( \Gamma )</td>
<td>0·36</td>
<td>−807·98</td>
<td>−807·41</td>
<td>0·57</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PB2</td>
<td>23</td>
<td>1317</td>
<td>HKY</td>
<td>0·48</td>
<td>−2378·08</td>
<td>−2375·83</td>
<td>0·10</td>
<td></td>
</tr>
<tr>
<td>PB1</td>
<td>31</td>
<td>1461</td>
<td>HKY + ( \Gamma )</td>
<td>0·75</td>
<td>−2883·95</td>
<td>−2883·95</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>27</td>
<td>2133</td>
<td>TrN93 + ( \Gamma )</td>
<td>0·62</td>
<td>−4228·70</td>
<td>−4228·20</td>
<td>0·606</td>
<td></td>
</tr>
<tr>
<td>NP</td>
<td>32</td>
<td>1203</td>
<td>K81uf + ( \Gamma )</td>
<td>0·17</td>
<td>−2431·93</td>
<td>−2431·93</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td>42</td>
<td>1275</td>
<td>TIM + ( \Gamma )</td>
<td>0·58</td>
<td>−2913·75</td>
<td>−2911·92</td>
<td>0·160</td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>36</td>
<td>927</td>
<td>HKY</td>
<td>0·85</td>
<td>−1900·06</td>
<td>−1900·06</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>36</td>
<td>249</td>
<td>K80</td>
<td>0·74</td>
<td>−491·88</td>
<td>−491·87</td>
<td>0·990</td>
<td></td>
</tr>
<tr>
<td>NS1</td>
<td>31</td>
<td>801</td>
<td>HKY + ( \Gamma )</td>
<td>0·37</td>
<td>−1766·14</td>
<td>−1760·29</td>
<td>0·003</td>
<td>171 (D, G); 205 (S, N, G); 209 (D, N, G)</td>
</tr>
<tr>
<td>NS2</td>
<td>36</td>
<td>327</td>
<td>HKY + ( \Gamma )</td>
<td>0·28</td>
<td>−684·94</td>
<td>−682·21</td>
<td>0·065</td>
<td></td>
</tr>
</tbody>
</table>

*Natural logarithm of the likelihood for the codon-based substitution M7 and M8 models. Values in bold are the ones belonging to the best-fitting model according to an LRT.
†A \( P \) value of \(<0·01\) indicates that the more complex model, M8, fits the data significantly better.
‡Sites are numbered according to amino acid positions starting from the methionine codon at the first position of each protein. The different amino acids present at each site under positive selection are given in parentheses.

### Table 2. Nucleotide-substitution models for the HA1 gene region of H5N1 influenza viruses isolated from 1997 to 2004

1998 and 1999 viruses were not analysed due to the insufficient number of sequences available. The best-fitting nucleotide-substitution model was HKY + \( \Gamma \) for each year.

<table>
<thead>
<tr>
<th>Year</th>
<th>n</th>
<th>Length of sequences (bp)</th>
<th>( \alpha ) parameter of ( \Gamma ) distribution</th>
<th>( \omega )</th>
<th>lnLK (M7)*</th>
<th>lnLK (M8)*</th>
<th>( P ) value†</th>
<th>Positively selected sites‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>26</td>
<td>1617</td>
<td>0·13</td>
<td>0·19</td>
<td>−3063·64</td>
<td>−3062·84</td>
<td>0·45</td>
<td>138 (H, L, N, Q)</td>
</tr>
<tr>
<td>2000</td>
<td>16</td>
<td>1098</td>
<td>0·63</td>
<td>0·37</td>
<td>−1998·3</td>
<td>−1998·12</td>
<td>0·8352</td>
<td>138 (L, Q); 140 (K, R, S, N); 155 (S, D, N); 156 (S, T, A)</td>
</tr>
<tr>
<td>2001</td>
<td>37</td>
<td>1098</td>
<td>0·85</td>
<td>0·27</td>
<td>−2854·76</td>
<td>−2849·18</td>
<td>0·0037</td>
<td>140 (K, R, S, N); 155 (S, D, N); 156 (S, T, A)</td>
</tr>
<tr>
<td>2002</td>
<td>26</td>
<td>1098</td>
<td>0·38</td>
<td>0·28</td>
<td>−2421·55</td>
<td>−2418·03</td>
<td>0·0295</td>
<td>140 (K, R, S, N); 155 (S, D, N); 156 (S, T, A)</td>
</tr>
<tr>
<td>2003</td>
<td>20</td>
<td>1098</td>
<td>0·65</td>
<td>0·24</td>
<td>−2462·75</td>
<td>−2477·56</td>
<td>0·0001</td>
<td>140 (K, R, S, N); 155 (S, D, N); 156 (S, T, A)</td>
</tr>
<tr>
<td>2004</td>
<td>60</td>
<td>855</td>
<td>0·48</td>
<td>0·38</td>
<td>−2651·12</td>
<td>−2644·60</td>
<td>0·001</td>
<td>227 (E, D, G)</td>
</tr>
</tbody>
</table>

*Natural logarithm of the likelihood for the codon-based substitution M7 and M8 models. Values in bold are the ones belonging to the best-fitting model according to an LRT.
†A \( P \) value of \(<0·01\) indicates that the more complex model, M8, fits the data significantly better.
‡Sites are numbered according to amino acid positions starting immediately after the signal peptide. The different amino acids present at each site under positive selection are given in parentheses.
Tables 1 and 2. The K80 model assumes equal base frequencies and a different rate of transitions and transversions (transition/transversion bias) (Kimura, 1980); the K81 model makes the same assumption as K80, but also allows for two different transversion rates; the K81uf model assumes in addition unequal base frequencies compared with K81 (Kimura, 1981); whereas the F81 model assumes variable base frequencies and substitutions to be equally likely (Felsenstein, 1981). The HKY + \Gamma model assumes different nucleotide frequencies, a transition/transversion bias and heterogeneity of substitution rates over sites modelled by a \Gamma distribution (Hasegawa et al., 1985); the TrN93 + \Gamma model makes the same assumptions as HKY + \Gamma, but also allows purine (A\rightarrow G) and pyrimidine (C\rightarrow T) transitions to occur at different rates (Tamura & Nei, 1993); the TIM model makes the same assumption as TrN93 + \Gamma, allowing also for two different transversion rates (Posada & Crandall, 1998).

ML trees were then inferred with the best-fitting model and ML-estimated substitution parameters with a heuristic search performed by using a neighbour-joining tree as starting tree. Calculations were performed by using the Phylogenetic Analysis Using Parsimony (PAUP*) software package version 4.0b10 (Swofford, 1999).

In the positive-selection analysis, ML trees were used as input trees and probabilistic models of codon substitution that allow for variable non-synonymous/synonymous substitution-rate ratios (dN/dS or \omega) among sites were applied to identify positively selected sites (Nielsen & Yang, 1998). In particular, models M7 and M8, developed by Yang et al. (2000) and implemented in the CODEML program of the PAML package (Yang, 1997), were used: M7 assumes that codons are divided into ten classes, \beta-distributed, whose \omega values lie between 0 and 1 (neutral- and negative-selection model), whereas M8 includes the ten classes of M7 plus an additional codon class where \omega can be >1 (positive-selection model). Thus, M7 and M8 are nested (degrees of freedom = 2) and can be compared statistically by using an LRT. The LRT indicates whether the substitutions inferred from an alignment are best explained by either the negative/neutral- or positive-selection model. When the likelihood of the positive-selection model is significantly higher than that of the nested neutral-hypothesis model, the empirical Bayes procedure is used to predict which codons are under positive selection and to calculate their probabilities (Yang et al., 2000). It is worth noting that PAML LRTs are reported to be conservative for short sequences (e.g. positive selection could be underestimated), although the Bayesian prediction of sites under positive selection is largely unaffected by sequence length (Anisimova et al., 2001, 2002).

Overall, 684 human and avian sequences were analysed. Firstly, we examined the phylogenies of all gene segments in the years 1997 and 2004 by using the best-fitting evolutionary models. Table 1 shows the selected models for all segments, except HA, of the 1997 and 2004 human and avian influenza A virus datasets. In 1997, all but the M2 gene region (\alpha parameter = 1.74) showed nucleotide substitution-rate heterogeneity over sites (\Gamma models). In one case (NP gene), the \alpha parameter of the \Gamma distribution was extremely low (\alpha = 0.006).

ML-inferred trees for all gene segments showed essentially the same topology. Two main clades, supported by highly significant bootstrap and \P values (in the zero branch-length test), could be distinguished: a duck/goose/human host clade and a chicken/human host clade. In both clades, the sequences of viruses isolated from humans after zoonotic transmission appeared to be intermingled with the avian ones (see Supplementary Fig. S1, available in JGV Online).

In 2004, six gene regions showed nucleotide substitution-rate heterogeneity over sites (\alpha parameter of the \Gamma models < 0.7). The \alpha parameter of the \Gamma distribution ranged from 0.17 for the NP gene to 0.85 for the M1 gene.

When we considered phylogenetic analysis of 2004 viruses, all virus genes appeared to cluster into two or three main groups, based on the isolation region. In contrast to the situation observed in 1997, human sequences fell within one group only, comprising several different avian species from Vietnam and Thailand, as described previously by Li et al. (2004) (see Supplementary Fig. S2, available in JGV Online).

As regards the H5 HA1 gene, 192 sequences were analysed. The \alpha parameter of the \Gamma distribution ranged from 0.13 for 1997 to 0.85 for 2001 (Table 2). The ML tree for the entire HA sequence dataset from 1997 to 2004 showed a clear temporal evolutionary trend, with the human sequences intermingling with the avian ones (data not shown).

The inferred trees were then employed as input trees in the program PAML to investigate the presence of codons under positive selection in the viral genes. As indicated in Table 1, among the 1997 viruses, the neutral model M7 applied to all genes except PB2. For this gene, M7 was rejected in favour of the positive-selection model (M8) and seven sites appeared to be under significant positive selection (Table 1). In 2004, the neutral model M7 applied to almost all genes, except for NS1 (Table 1, positions 171, 205 and 209) and HA1. The analysis of the HA1 subunit over the years (Table 2) indicated that the mean non-synonymous/synonymous substitution-rate ratio ranged from 0.19 for the year 1997 to 0.38 for the year 2004, thus suggesting that, on average, the H5 HA protein was under strong purifying selection (\omega < 1). However, in the years 2001, 2003 and 2004, the positive-selection model M8 provided a significantly better fit to the data, as indicated by the LRT, than did M7. The proportion of all positively selected sites increased from 0.3% in 2001 to 6.7% in 2004 (data not shown). Amino acid positions under positive selection with 95% confidence level (138 in 2001; 138, 140, 155 and 156 in 2003; 138, 140, 156, 218 and 227 in 2004) were located on the globular head of the molecule, within or near two previously identified antigenic epitopes of the H5 molecule (Kaverin et al., 2002;
Within the 2004 H5N1 HA dataset, including the viruses responsible for the still-ongoing epidemic in South-East Asia, nine additional positions with a posterior probability above 0.5 were identified (Fig. 1). Of these, two amino acid positions (129 and 212) were under positive selection above the 80% confidence level, and the remainder (124, 137, 141, 189, 209, 269 and 326) were between the 50 and 80% levels.

In this study, we investigated the evolutionary dynamics of H5N1 strains by using ML techniques. Within the 1997 viral isolates, nine of the ten viral genes investigated showed strong nucleotide substitution-rate heterogeneity across sites. In particular, the NP gene had a very low $\alpha$ value, suggesting that whilst most of the sites along the gene may be invariable because of strong purifying selection, a few mutational hot-spots are also present that may accumulate mutation at a much faster rate because of positive selection and/or relaxed selective constraints. However, the mean ratio of non-synonymous to synonymous nucleotide substitutions in all gene segments did not exceed the threshold of $\omega = 1$ (data not shown), suggesting purifying selection along the sequences. On the contrary, the codon-based model showed an $\omega$ value significantly greater than 1, indicating positive selection, for seven sites in the PB2 gene (see Table 1). Substitution of Lys with Gln at residue 355 has been associated with a highly pathogenic phenotype in mammalian hosts (Katz et al., 2000). However, none of the residues under selection was found specifically in humans.

Three of the 2004 virus genes showed strong nucleotide substitution-rate heterogeneity across sites. As in 1997, NP had the lowest $\alpha$ value. Apart from HA, only NS1 appeared under positive selection, at positions 171, 205 and 209. The latter finding agrees in part with the results of another analysis of H5N1 influenza viruses circulating in South-East Asia in 2000–2004. Among the internal protein genes of genotype Z strains isolated between 2002 and 2004, the M2 gene (in 2002–2003) and NS1 and NS2 genes were under positive selection pressure (Li et al., 2004).

Of the eight gene segments originally present in the 1997 H5N1 virus from Hong Kong, seven have been replaced by reassortment over the years by segments belonging to different avian lineages. Only the H5 HA lineage has remained constant between 1997 and 2004 (Guan et al., 2002; Li et al., 2004). For this reason, we compared $\omega$ variation at single-codon sites of all the H5 HA1 sequences from H5N1 strains isolated between 1997 and 2004 and identified the positively selected residues. Our results indicate clearly that, despite an overall conservative evolution of the gene suggested by the mean $\omega$ values (ranging from 0.19 to 0.38), a small but increasing number of specific amino acid positions has been under positive selection pressure since 2001. Site 138 was a target of selection pressure in all 3 years considered (Table 2), whereas positions 140, 155, 156, 218 and 227 were positively selected in 2003 and/or 2004. Moreover, in 2004, several additional sites underwent amino acid replacements at varying degrees of probability (Fig. 1).

**Fig. 1.** Posterior probabilities of site classes along the 2004 HA1 gene region under the M8 model. Selected amino acid sites with posterior probabilities above 50% are indicated. Asterisks indicate sites with a posterior probability above 95%.
Very recent reports have characterized the antigenic mapping of the H5 molecule, either through the use of neutralizing mAbs or as a result of prolonged vaccine use (Lee et al., 2004; Li et al., 2004). Other authors have highlighted that receptor specificity of influenza viruses, including H5, is not limited to the differential recognition of Sia\textsubscript{2,3}-Gal (specific for avian influenza viruses) and Sia\textsubscript{2,6}-Gal (recognized by human viruses) moieties, but appears to be modulated more finely by several residues in the vicinity of the receptor-binding site (Ilyushina et al., 2004; Gambaryan et al., 2005). Most of the positively selected sites identified in our study are associated strongly with either mapped neutralizing epitopes or with regions involved in receptor-binding affinity modulation. Positions 124, 129, 137, 138, 140, 141 and 189 have previously been recognized as targets for neutralizing antibodies (Kaverin et al., 2002), whereas positions 124, 138 and 140 have been involved in the carbohydrate-recognition specificity of the HA towards the host-cell receptor (Ilyushina et al., 2004). Finally, we confirmed that position 156 of the HA was under positive selection for the acquisition of an additional glycosylation site near the receptor-binding and antigenic regions at the tip of the molecule (Li et al., 2004). Thus, positive selection at these sites, on one hand, may help the virus to evade the immune response and suggests the existence of an immunological pressure, possibly due to the use of poultry vaccines. On the other hand, the presence in the viral population of strains with differential affinity for the host-cell receptors could favour the emergence of avian viruses with a higher affinity for the epithelial respiratory tissues of mammals (Gambaryan et al., 2005).

The relatively limited evidence of positive selection in 1997 and 2004 H5N1 viruses could be explained in part by the short study period and by the fact that the human strains used in our study represent direct transmissions from birds to humans. As these strains did not undergo inter-human transmission, they may themselves be considered as avian viruses. This may account for the fact that the few sites under positive selection were not associated with a different amino acid pattern between human and animal influenza viruses.

The actual genes and residues involved in the adaptation of an avian influenza virus to a new host are mostly unknown. Of the three genes that we found to be under positive selection, HA is the major determinant of receptor specificity, virulence and antigenicity, whilst PB2 and NS1 have been involved in increase of pathogenicity in a mammalian host (Katz et al., 2000; Seo et al., 2002). Our study indicates that positive-selection analysis performed on single-codon sites could be helpful in clarifying some of these aspects. The occurrence of adaptive mutation in poultry, a non-natural host of avian virus, could be a crucial mechanism by which the virus could also improve its transmissibility among humans. The possibility of this event strengthens the importance of improving surveillance in both humans and poultry to control the spread and evolution of the avian influenza virus strains.

Acknowledgements

The authors would like to thank Ms Sabrina Tocchio for her editorial assistance. M. S. is supported by grants AI065265 and HD32259 and the Department of Pediatrics of the University of Florida, Gainesville, FL, USA.

References


