Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A

Robin M. Bush, Walter M. Fitch,* Catherine A. Bender,† and Nancy J. Cox†*

*Department of Ecology and Evolutionary Biology, University of California at Irvine; and †Influenza Branch, Centers for Disease Control and Prevention, Atlanta, Georgia

The hemagglutinin (HA) gene of influenza viruses encodes the major surface antigen against which neutralizing antibodies are produced during infection or vaccination. We examined temporal variation in the HA1 domain of HA genes of human influenza A (H3N2) viruses in order to identify positively selected codons. Positive selection is defined for our purposes as a significant excess of nonsilent over silent nucleotide substitutions. If past mutations at positively selected codons conferred a selective advantage on the virus, then additional changes at these positions may predict which emerging strains will predominate and cause epidemics. We previously reported that a 38% excess of mutations occurred on the tip or terminal branches of the phylogenetic tree of 254 HA genes of influenza A (H3N2) viruses. Possible explanations for this excess include processes other than viral evolution during replication in human hosts. Of particular concern are mutations that occur during adaptation of viruses for growth in embryonated chicken eggs in the laboratory. Because the present study includes 357 HA sequences (a 40% increase), we were able to separately analyze those mutations assigned to internal branches. This allowed us to determine whether mutations on terminal and internal branches exhibit different patterns of selection at the level of individual codons. Additional improvements over our previous analysis include correction for a skew in the distribution of amino acid replacements across codons and analysis of a population of phylogenetic trees rather than a single tree. The latter improvement allowed us to ascertain whether minor variation in tree structure had a significant effect on our estimate of the codons under positive selection. This method also estimates that 75.6% of the nonsilent mutations are deleterious and have been removed by selection prior to sampling. Using the larger data set and the modified methods, we confirmed a large (40%) excess of changes on the terminal branches. We also found an excess of changes on branches leading to egg-grown isolates. Furthermore, 9 of the 18 amino acid codons, identified as being under positive selection to change when we used only mutations assigned to internal branches, were not under positive selection on the terminal branches. Thus, although there is overlap between the selected codons on terminal and internal branches, the codons under positive selection on the terminal branches differ from those on the internal branches. We also observed that there is an excess of positively selected codons associated with the receptor-binding site and with the antibody-combining sites. This association may explain why the positively selected codons are restricted in their distribution along the sequence. Our results suggest that future studies of positive selection should focus on changes assigned to the internal branches, as certain of these changes may have predictive value for identifying future successful epidemic variants.

Introduction

Several recent studies have analyzed molecular sequence data to identify genes or parts of genes that are targets of natural selection (Hughes and Nei 1988, 1989; Fitch et al. 1991; Hughes 1992; Ina and Gojobori 1994). Of particular interest is the hemagglutinin (HA) gene of the influenza virus. The HA gene, which encodes the major surface antigen of the virus, exhibits a rapid rate of change, presumably in response to human immune surveillance in a partially immune human population. The first molecular evidence that the HA gene of the type A (H3 subtype) human influenza virus was under positive selection to change was the finding that for codons in antibody combining sites, more amino acid replacements were fixed on the trunk of the evolutionary tree than on its side branches. This suggested that replacements in antibody-combining sites enhance survival of the strain (Fitch et al. 1991). Ina and Gojobori (1994) later showed positive selection occurring in the HA gene of the H1 subtype, as evidenced by a significantly elevated ratio of nonsilent to silent mutations in antibody-combining sites.

The method of testing for positive evolution by comparing the relative frequencies of nonsilent and silent substitutions in two gene sequences was developed by Li, Wu, and Luo (1985). They looked for a significant increase in nonsilent changes under a null hypothesis of neutral drift and equiprobable nucleotide mutations given the specific codon at each position. To get a large enough sample of substitutions, one had to compare whole regions of the gene, such as, for example, the V3 region of HIV. Recent studies have broadened the method's utility. Messier and Stewart (1997) have used the method to determine positive selection along individual branches of a phylogenetic tree. The method has also been broadened to utilize maximum likelihood (Yang 1998). In addition, Golding (1987) and Golding and Felsenstein (1990) have expanded the utility to negative selection. We have further broadened the method by examining sufficient numbers of sequences in a phylogenetic tree to obtain enough substitutions to determine whether positive selection has acted upon individual codons. Moreover, we have developed an alternative null model by defining the expectations using the observed changes in the tree (Fitch et al. 1997).

Key words: influenza, virus, positive selection, hemagglutinin, evolution.

Address for correspondence and reprints: Robin M. Bush, Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697. E-mail: rmbush@uci.edu.

Mol. Biol. Evol. 16(11):1457–1465. 1999

The ultimate goal of our work is to understand the underlying patterns in the evolution of the HA of influenza viruses during replication in humans. By doing so, we hope to be able to predict the survival of individual strains arising in currently circulating influenza strains and thus predict which lineages, among those sampled through surveillance, are most likely to survive. Thus, we have focused our efforts on identifying the individual codons under positive selection.

We recently developed a technique for analyzing selection on individual codons, rather than pools of codons, by testing whether the numbers of silent and nonsilent mutations at each codon deviated from binomial expectations. These expectations were generated using the total number of silent and nonsilent mutations across codons (Fitch et al. 1997). In that study, we found excess changes on the terminal branches of the HA phylogenetic tree. Terminal branches are those that join the strain isolates to the tree. Although the source of these excess changes is unknown, they could easily include several classes of changes that do not involve viral evolution in the human population. The first class consists of host-mediated changes. These are amino acid replacements that become fixed during growth of isolates in the laboratory, particularly in embryonated chicken eggs. We know of 22 codons for which host-mediated change has been reported in the literature (Nakajima, Nakajima, and Kendal 1983; Robertson 1993; Rocha et al. 1993; Gubareva et al. 1994; Hardy et al. 1995). However, there is variation in how frequently these codons have been observed to undergo host-mediated change. Some changes that were observed very infrequently may have been due to chance rather than selection for growth in eggs. Positions in which host-mediated changes occur have altered over time, and there may be additional codons subject to host-mediated change that have yet to be detected or reported in the literature. We have examined data for codons for which there are multiple reports of host-mediated change and have found that these codons may show as many replacements on branches leading to cell-cultured isolates as on branches leading to egg-grown isolates (unpublished data). Thus, we can neither discard nor include these codons as a separate group.

In addition to host-mediated changes, other potential sources of extra changes on terminal branches include (1) investigator bias in choosing to sequence more atypical (rather than typical) isolates that are potential antigenic variants; (2) errors inherent in sequencing; and, as Golding, Aquadro, and Langley (1986) have suggested, (3) deleterious mutations, which are more likely than favorable mutations to be sampled only once before they are removed by selection. By definition, the nodes on the terminal branches of the tree leave no descendants. Whether the HA mutations assigned to the terminal branches were responsible for this lack of evolutionary success is, of course, unknown. In this study, we devised techniques to remove the variation on terminal branches as a way to determine whether it was affecting our analysis.

To determine the extent to which our results are affected by chance amino acid replacements, we ran separate analyses using only terminal and only internal changes. Such separation was not possible in our previous study because the sample size was too small. The test for selection requires a minimum of at least four nucleotide substitutions per codon to permit a statistically significant result ($P < 0.05$). The current data set of 357 sequences is 40% larger than the one we previously analyzed, primarily due to inclusion of HA sequences from isolates obtained in 1997.

We improved our analysis in two additional ways. First, in our test for positive selection, we reduced the effects of the extreme right skew in the distribution of nonsilent substitutions per codon. In our previous analysis of a subset of these data (Fitch et al. 1997), we found that six codons (amino acid positions 138, 145, 156, 186, 193, and 226) had 18 or more substitutions each, while no other codon had more than 12 substitutions. These six positions represented less than 2% of the total codons but contained 13% of all nucleotide substitutions. Change at these positions was strikingly nonrandom; together, they had 9 silent and 123 nonsilent substitutions, a 1:14 ratio. The ratio for the rest of the positions was 1 silent to 0.9 nonsilent changes. This observation suggests that the ''hypervariable'' codons were subject to a different set of selective forces than the rest of the codons. Thus, it is not appropriate to include these substitutions when testing whether the ratio of nonsilent to silent mutations at individual codons differs from random. In the current study, codons with a significant excess of silent or nonsilent substitutions were eliminated from the analysis prior to calculating the mean number of silent and nonsilent mutations across the gene. Second, we studied a population of parsimony trees rather than a single tree to be sure that our test for positive selection is robust against minor variation in tree structure. Maximum-likelihood routines do not yet exist to analyze data sets of this size, and the complexity of the HA data set precludes obtaining a single representation of the phylogeny using maximum parsimony.

Materials and Methods

Sequencing and Phylogenetic Analysis

The data set includes 357 nucleotide sequences for the HA1 domain of the HA gene from the human influenza A (H3N2) viruses. These sequences were generated at the Centers for Disease Control and Prevention using previously described methods (Fitch et al. 1997). They have been deposited in GenBank (accession numbers AF008656–AF008909 and AF180564–AF180666). All sequences are 329 codons and 987 nt long.

Preliminary analyses showed that the method by which we detect positive selection is sensitive to variation in tree structure. The codons affected are those that are near the cutoff in the positive-selection test. For these codons, subtle alterations in tree structure can change the assignment of a nucleotide substitution from an internal branch to a terminal branch, or vice versa.

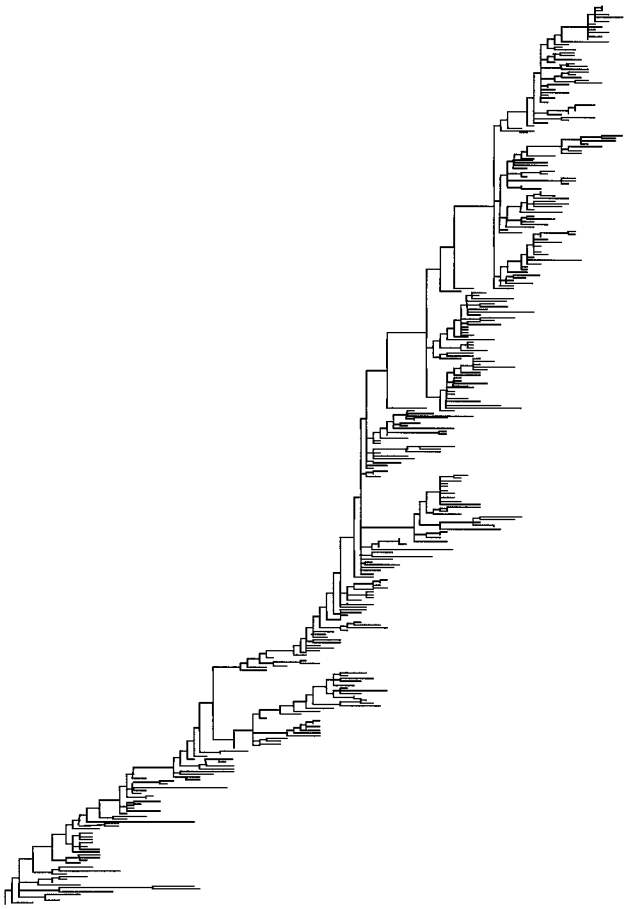

FIG. 1.—Maximum-parsimony tree constructed from 357 HA1 genes of the human influenza virus type A subtype H3.

Table 1
Distribution of Amino Acid Replacements by Branch Type

Branch Type	No. of Branches	Replacements		χ^2
		Observed	Expected	
Trunk.....	60	40	62.78	8.27
Twig.....	295	181	308.67	52.81
Tip.....	357	524	373.55	60.60
Total.....	712	745	745.00	121.68

NOTE.—$P < 10^{-26}$; df = 2.

This can affect whether a position is found to be under selection in a particular branch category. To examine whether our positive-selection test is robust with respect to minor variation in tree structure, we analyzed a population of 100 trees generated using the tree bisection-reconnection branch-swapping option of the heuristic search option of the maximum parsimony routine of PAUP, version 4.0.0d60 (Swofford 1993). The order in which sequences were input to the program were randomized for each run.

The number of nucleotide substitutions required to construct the 100 trees ranged from 1,347 to 1,355. The number of codons found to be under selection increased slightly with the total number of nucleotide changes that occurred on a tree. The additional selected codons obtained from analysis of the longer trees result from extra substitutions that would have been eliminated in an exhaustive search of all possible tree structures. Given the size of this data set, it was not possible to do an exhaustive search. Thus, it is important to analyze only the shorter trees that were generated: we found only four trees in the shortest category, those trees requiring 1,347 nucleotide substitutions. The four shortest trees produced slightly different sets of selected positions, indicating the need to sample a variety of tree structures. By starting with one of the four shortest trees in memory, additional branch swapping could be used to produce a large number of different trees of equal length. However, these additional trees always produced the same sets of selected positions as the tree from which they were generated. The only way we were able to successfully generate a range of substantially different tree structures was to randomize the order in which sequences were read into PAUP; however, PAUP rarely found a tree requiring less than 1,350 substitutions. We therefore identified the codons obtained from analysis of the 11 trees of length 1,347 or 1,348 as our set of positively selected codons. These 11 trees, which make up the shortest 11% of the total sample of trees, are very similar to one another, differing only slightly in branching order. The standard deviations associated with the mean number of substitutions and with the mean number of nonsilent and silent changes on the internal and terminal branches are generally smaller than 1% of the associated means. We will describe only one of these 11 trees in this paper (fig. 1). This tree was chosen because it generated the entire sample of positively selected positions produced from analysis of all 11 smallest trees. When amino acid sequences were examined, we assumed that this nucleotide tree was the correct topology. All analyses other than construction of phylogenetic trees were performed using programs written by R.M.B.

The Distribution of Replacements on Terminal and Internal Branches

In our previous work we divided the branches of the influenza HA phylogenetic tree into three categories; trunk, twig, and tip branches. The trunk was defined as the set of interior branches leading from the root to the most distal tip. Tip branches (called terminal branches in this paper) joined isolates to the tree, and all other branches were twigs. If there is no a priori expectation concerning where amino acid replacements may occur in the tree, then replacements should be distributed in trunk, twig, and tip branches in proportion to the number of such branches. Table 1 shows that this null hypothesis is not true, principally because there are about 40% more replacements on the terminal branches than expected. After removing the terminal branches, a test of the distribution of replacements in the remaining trunk and twig branches shows that we cannot reject the null hypothesis that replacements are distributed with equal probability to all internal branch-

es ($X^2 = 0.23$; df = 1; $P = 0.7$). Thus, trunk and twig branches were pooled for analysis here and are referred to as internal branches.

As reviewed above, several processes other than selective pressure exerted by the human immune system could produce changes on the terminal branches of a phylogenetic tree. To ensure that such nonselected changes were assigned to terminal branches when there was a choice, we used a method of assignment similar to the DELTRAN method of character optimization in PAUP. Under this option, changes are assigned as far from the root as possible. This increased the number of substitutions assigned to terminal branches by 3% (from 510 to 524) over the alternative (ACCTRAN-like) assignment.

Detecting Codons Under Positive or Negative Selection

We calculated the frequencies of nonsilent and silent substitutions as $p = n/(n + s)$ and $q = s/(n + s)$, respectively, where n and s are the numbers of nonsilent and silent nucleotide substitutions, respectively, and $n + s = S$ is the total number of substitutions summed over all sites. We then use the binomial equation to compute the probability that the observed numbers of silent and nonsilent changes in any given codon would occur by chance. For example, the probability of a codon having six nonsilent and two silent substitutions is $P = 8!p^6q^2/(6!2!)$. To obtain a statistically significant result ($P < 0.05$), this test requires a minimum of $M = 4$ changes in a codon unless the smaller of p and q is greater than 0.473 ($0.473^4 \approx 0.05$). In that case, a minimum M of 5 changes in a codon is needed.

Different codons have different fractions of possible silent changes. Incorporating codon-specific expectations into our estimation of q might have produced a different set of positively selected codons. However, all possible mutations are equally probable only under the assumption of neutrality. Under neutrality, q should have averaged 0.25 rather than 0.577, the value we observed. Our result is more consistent with the selective removal of non-silent mutations prior to sampling. As shown below (in the most robust subset of our data), we observed 191 silent mutations. If the assumption of neutrality applied, these 191 would represent 25% of all mutations. In that case the original number of mutations would have been 764 and the number of non-silent mutations (75% of 764) would have been 573. The numbers we observed were quite different; of a total of 331 mutations, 191 (58%) were silent and 140 (42%) were nonsilent. We observed 433 fewer non-silent mutations than expected under neutrality. Since 75.6% of the non-silent mutations were, under this model, inferred to be deleterious, we did not consider it appropriate to assume neutrality for mutations affecting this gene.

In our previous analysis of a subset of these data (Fitch et al. 1997), we identified a set of six ''hypervariable'' codons. Our results suggested that these codons might be subject to a different set of selective forces than the rest of the codons. If this is true, then it would not be appropriate to include them when calculating p and q values. In the present analysis, we eliminated the bias introduced by these ''hypervariable'' codons by estimating p and q, the frequencies of nonsilent and silent substitutions, respectively, only after first removing from the data set those codons that showed significant excesses of either silent or nonsilent substitutions. We did this by first calculating values of p and q from their observed frequencies in the total data set, and then using these values to identify an initial list of N codons with a significant excess of silent or nonsilent mutations. As multiple testing may have produced some significant results by chance, we removed from this list of N significant codons those 5% of codons with P values closest to 0.05. Half, or 2.5% of the N significant codons, were removed from each end of the distribution. When that 5% of N was fractional, we rounded up to the next whole number, E. When E was an even number, we excluded $E/2$ of the codons from each tail of the distribution. When E was an odd number, the extra codon was excluded from the end of the distribution with the larger number of codons in N.

The remaining codons, which did not show an excess of nonsilent (n) or silent (s) substitutions, were used to calculate new values of p and q. The new values were used to identify the set of codons with a significant excess (or deficit) of nonsilent nucleotide substitutions compared with binomial expectations. These are the codons we designate as being under positive (a significant excess of nonsilent changes) or negative (a significant excess of silent changes) selection to change.

Results

Distribution of Changes on the Phylogenetic Tree

The tree in figure 1 has 1,348 nucleotide substitutions distributed over 329 codons; an average of 4.1 substitutions per codon occurred across the gene. Fifty-one percent of the nucleotide positions (508 of 987) are unvaried, as are 46% (152/329) of the codons. Of the 1,348 total nucleotide substitutions, 745 (55%) were nonsilent and 603 (45%) were silent. Nine hundred twenty-eight substitutions were located on the 357 terminal branches, which is more than twice the 420 changes found on the 355 internal branches. The terminal branches show a 40% excess of mutations over expectations (table 1). The average numbers of nonsilent (NS) and silent (S) nucleotide substitutions on the internal branches are 221 and 199, respectively. For the terminal branches, NS = 524 and S = 404. There are more nonsilent than silent substitutions on both terminal and internal branches. However, the excess is larger on the terminal branches (NS/S = 1.30) than on the internal branches (NS/S = 1.11). Furthermore, there is an excess of changes on terminal branches leading to isolates grown in embryonated eggs, and there is a deficit of changes in isolates grown in cell culture (table 2).

Selection on Internal Branches

Of the 420 substitutions on the internal branches, 221 were nonsilent and 199 were silent. The resulting frequencies ($p = 0.526$, $q = 0.474$) require that a codon

Table 2
The Distribution of Amino Acid Replacements on Terminal Branches According to the Method of Laboratory Propagation Prior to Sequencing

Propagation Method	No. of Isolates	No. of Replacements		χ^2
		Observed	Expected	
Egg	152	256	223.1	4.85
Cell	148	186	217.2	4.49
PCR.....	4	9	5.9	1.67
Other.....	53	73	77.8	0.30
Total	357	524	524.0	11.31

Note.—"Egg" identifies any isolate propagated in egg culture for at least part of its history. "Cell" indicates isolates passaged only in cell culture. "PCR" indicates isolates sequenced directly by polymerase chain reaction. "Other" indicates any isolates not fitting the above three categories. Across propagation methods, the probability of a worse fit occurring by chance is <0.025 with three degrees of freedom. There is a significant excess of replacements in egg-passaged samples and a significant deficit of replacements in the cell category (both $P < 0.05$).

must have at least five substitutions for a statistically significant deviation from the expected number of silent and nonsilent substitutions, because $0.474^4 = 0.0505$. There were $S = 17$ codons with five or more substitutions on the internal branches. Twelve of the 17 codons showed significant deviations from expectations, and 11 of the 12 had excesses of nonsilent substitutions. The exception was codon 315, with five silent substitutions. Because we performed the test for significance on each of the 17 codons having at least five substitutions, 5% of 17, or 1 of the 12 significant results, would be expected to occur by chance alone. Therefore, we removed 1 codon from the class of 12 significant codons, leaving 11 whose significance could not be attributed to multiple testing. Codon 315 was removed, as it had the largest *P* value among the 12 significant codons.

We then recalculated *p* and *q*, excluding the 11 codons with significant excesses of nonsilent changes. The remaining 318 codons had a total of 331 nucleotide substitutions: 140 nonsilent and 191 silent. Using only these 318 codons, $p = 0.423$ and $q = 0.577$, a considerable change from the values of *p* and *q* estimated using all codons. As a result, the number of substitutions per codon needed in order to obtain a significant result was now four instead of five. Thirty-eight codons had four or more substitutions and 18 of these codons showed significant excesses of nonsilent substitutions. Of the 18 codons identified as being under positive selection using only changes on internal branches, three had a single silent substitution; the rest had none.

The above procedure removed the effect of a small number of codons with extreme nonsilent-to-silent ratios from the group of codons used to calculate our expected values. This procedure can be repeated by eliminating 5% of *S* from the new pool of significant codons and recalculating *S, p, q,* and *P* over and over until the system converges. We performed a third round of analyses on these data and identified the same positively selected codons as before. However, because the number of changes, *M*, needed in order to get a significant result was now 3 and $S = 59$, four additional codons were significant, with NS/S being 3/0. We did not include them in our set of selected codons, because approximately this number of significant results is expected to occur by chance ($0.05 \times 59 = 3$) as a result of performing multiple tests. The resulting pool of positively selected codons is shown in table 3. None of the codons showed evidence for negative selection on the internal branches.

Selection on Terminal Branches

Of the 928 changes occurring on terminal branches, 524 were nonsilent ($p = 0.565$) and 404 were silent ($q = 0.435$). The number of changes, *M*, needed for a codon to deviate significantly from expectations using these frequencies is four. Of the 92 codons with four or more substitutions, 34 showed significant deviations from the expected distribution of nonsilent to silent changes. Sixteen of the 34 significant codons had excess nonsilent changes, and the other 18 had excess silent changes. We removed 5% of the 92 codons from the list of significant codons, two with excess nonsilent substitutions and three with excess silent changes. Thus, there were 29 codons whose significant excesses of silent or nonsilent changes could not be attributed to multiple testing. We then calculated new values of *p* and *q* excluding the mutations in the 29 significant codons, which together accounted for 164 nonsilent and 85 silent substitutions. The remaining pool of 300 codons had 675 substitutions. Because 360 substitutions were nonsilent and 315 were silent, $p = 0.533$ and $q = 0.467$. We used these new frequencies to examine those codons with four or more substitutions and found that 39 codons had values that significantly differed from expectation. Of these, 21 had significant excesses of nonsilent changes and 18 had significant excesses of silent changes on terminal branches. None of the 18 codons were located in the central third of the sequence, the region in which all of the positively selected codons found using the internal branches occur (table 3). A significant excess of silent substitutions could arise from the suppression of replacements at a position.

Discussion

To identify the key amino acid changes in the influenza HA, we examined the pattern of change across a phylogenetic tree composed of HA gene sequences of 357 influenza A (H3N2) field strains collected between 1983 and 1997. Codons in the influenza HA gene identified here as being under positive selection are presumably encoding amino acid replacements that allow the virus to evade existing immunity in the population. Prospectively monitoring changes in these key positions in the HA may be useful in identifying the precursors of new genetic variants that may cause future influenza epidemics. Indeed, we have evidence that this is so (unpublished data).

We improved our test for positive selection by removing the extreme right skew in the distribution of nonsilent substitutions. As a result of this change, we

Table 3
The Codons of HA1 Hemagglutinin from Human Influenza Virus A (H3N2) Under Positive or Negative Selection

Codon	Internal NS/S	Terminal NS/S	Sets	Codon	Internal NS/S	Terminal NS/S	Sets
12.....	0/0	0/7−		186.....	9/1+	14/4+	B
15.....	0/1	1/8−		190.....	4/0+	9/0+	B, R
16.....	0/2	0/5−		193.....	4/0+	20/1+	B
33.....	0/1	0/4−		194.....	4/0+	9/2+	B, R
42.....	0/1	0/4−		196.....	3/1	9/0+	B
61.....	0/3	0/9−		197.....	4/0+	1/3	B
80.....	2/0	5/0+	E	201.....	4/0+	2/1	D
109.....	0/0	0/5−	E	219.....	5/3	8/0+	D
115.....	0/1	0/4−		220.....	0/0	5/0+	
118.....	0/1	0/6−		226.....	20/1+	18/2+	D, R
119.....	0/1	0/4−		227.....	1/4	0/4−	D, R
121.....	5/0+	5/1	D	239.....	0/0	0/5−	
124.....	5/0+	1/2	A	244.....	0/2	0/5−	D
128.....	0/0	8/0+	B	246.....	3/2	13/2+	D
133.....	8/0+	4/0	A	261.....	1/1	0/5−	E
135.....	5/0+	7/1+	A, R	262.....	4/0+	3/2	E
137.....	0/0	12/1+	A, R	275.....	6/0+	4/0	C
138.....	6/0+	16/0+	A, R	276.....	2/0	9/0+	C
142.....	4/0+	4/1	A	293.....	0/2	1/6−	
145.....	8/0+	11/0+	A	307.....	1/2	0/4−	C
156.....	9/1+	15/1+	B	310.....	2/2	7/0+	C
158.....	5/0+	3/0	B	312.....	1/0	7/0+	C
159.....	3/1	11/2+	B	321.....	0/0	0/5−	
182.....	0/0	7/1+	D	323.....	1/1	0/4−	

NOTE.—"NS/S" is the number of nonsilent substitutions over the number of silent substitutions. A plus or minus sign indicates that the probability that NS/S would be observed by chance is less than 0.05. A plus sign indicates positive selection on the internal (n = 18) and terminal (n = 21) branches. A minus sign indicates negative selection on the terminal branches (n = 18). "Sets" indicates membership in the following codon sets: A–E = antibody combining sites A–E; R = receptor binding site.

found 18, rather than 11, codons on internal branches to be under positive selection. The seven additional codons all have NS/S ratios of 4/0. These codons are those that may or may not pass the test for positive selection, depending on minor variation in tree structure in our collection of parsimony trees. Sensitivity to tree structure and to the method of analysis indicates that additional data may be required to firmly place these codons in the class of positively selected codons.

Eliminating changes on terminal branches from the analysis produced an effect similar to that produced by removing the extreme right skew in the distribution of nonsilent substitutions per codon, because much of that right skew arose from changes on terminal branches. With these changes removed, only position 226 remains an outlier. It undergoes 20 replacements (out of 21 total substitutions) on the internal branches, with the rest of the codons having at most 9 replacements. It also has 18 replacements (out of 20 total substitutions) in the terminal branches. It is thus a hot spot for mutations and/or selection. Selection is very likely the cause, because residue 226 is located in the receptor-binding pocket of the HA, and whether HA binds preferentially to the 2→3gal form of sialic acid (the human structure) or to the 2→6gal form (the avian structure) is determined by whether the residue at position 226 is, respectively, glutamine or leucine (Rogers et al. 1983; Ito et al. 1997).

Excess Replacements on the Terminal Branches

We previously observed a large (38%) excess of amino acid replacements on the terminal branches of the HA tree, and we were concerned that this excess represented genetic variation generated in response to pressures other than immune selection (Fitch et al. 1997). Analysis of the new data set containing 40% more HA sequences resulted in a similar 40% excess of amino acid replacements on terminal branches. A partial cause of this excess may be host-mediated change, because branches joining sequences from egg-grown isolates to the tree have significantly more changes than expected. Our findings also support results in the literature (Nakajima, Nakajima, and Kendal 1983; Robertson 1993; Rocha et al. 1993; Gubareva et al. 1994; Hardy et al. 1995) indicating that codons vary in their propensity to undergo host-mediated change. Furthermore, amino acid changes that were associated only with growth of influenza A viruses in eggs during circulation of a particular antigenic variant may subsequently be "fixed" into the virus population as evolution continues. This likely occurs due to antibody pressure when host-mediated changes occur in or adjacent to antibody-combining sites (Rocha et al. 1993).

At least three other processes may contribute to the excess of changes on terminal branches. The first is investigator bias. The majority of viruses selected for HA sequencing exhibit at least minor antigenic variation based on hemagglutinin inhibition tests. We can show through simulation that such a bias produces excess changes on terminal branches. However, we are not able to quantify or remove such a bias from our data set. A second possible source of excess changes on terminal branches is the errors that occur during sequencing. This

source of error (0.002 per nucleotide per cycle; Saiki et al. 1988) is far too small to explain the terminal branches having more than twice as many replacements per branch as the interior branches. The third source of excess changes arises from the fact that deleterious mutations are more likely than favorable mutations to be sampled only once before they are removed by selection. This would produce excess changes on terminal branches as well. However, we do not currently have methods to evaluate or remove these biases.

Contrasting Results from Terminal and Internal Branches

Nine codons showed evidence for positive selection in both terminal and internal branches. This suggests the possibility that the codons found to be under selection on the terminal branches and on the internal branches might be two samples from the same underlying distributional process. This does not seem to be the case for either positively selected codons or the set of negatively selected codons for the following reasons. Twenty-one codons were found to be under positive selection in only one of the two branch type groups. Twelve codons were positively selected only using changes on terminal branches. The nine codons that were positively selected only on internal branches have significantly more nonsilent changes (mean = 5.0, SD = 1.3 vs. mean = 3.0, SD = 1.4; *t*-test: $P < 0.007$) and significantly fewer silent changes (mean = 0, SD = 0 vs. mean = 1.1, SD = 1.05; *t*-test: $P < 0.01$) on the internal branches than on the terminal branches. Eighteen codons have significant excesses of silent substitutions on the terminal branches; no codon has an excess of silent changes on the internal branches. Only 3 silent substitutions across the 18 positively selected codons were identified on internal branches, but 17 silent substitutions across the 21 codons under positive selection were identified on terminal branches. Two of the codons under positive selection on internal branches (124 and 197) show more silent than nonsilent changes on the terminal branches. These results would not be expected if the two codon sets were simply a subset of a single larger group responding to the same selective forces.

Thus, although there is overlap between the selected codons found using changes on terminal and internal branches, the codons under positive selection using changes on the terminal branches appear to contain information which is, at least in part, different from that conveyed by changes on the internal branches. We have evidence that host-mediated change and other causes not related to evolution during replication in a partially immune population may be responsible for additional changes occurring on terminal branches. Thus, the adaptive importance of positively or negatively selected codons determined using changes on the terminal branches is unknown.

What Is the Target of Selection?

Why are these 18 codons under positive selection to change? The answer may lie in the contribution of these codons to the receptor-binding site and the antibody-combing sites of the HA. Weis et al. (1988) listed 16 codons that are involved in the HA receptor-binding site. This site enables the virus to bind to sialic acid on the host cells. One might expect these codons to show few or no replacements in order to preserve the specificity of the receptor site. This may be true for codons such as positions 227 and 228, which have more silent than nonsilent changes. However, we found that five of the codons associated with the receptor-binding site are under positive selection to change (135, 138, 190, 194, and 226). Thus, some of these codons are flexible with regard to which amino acid they encode. We are currently investigating whether covariation among codons compensates for the resulting changes in protein structure.

All 18 of the positively selected codons are associated with the antibody-combining sites on the HA. By examining mutation in the HA of laboratory-selected monoclonal antibody escape mutants of the influenza A/HongKong/68 (H3N2) isolate and taking into account changes that had occurred in the major epidemic strains circulating between 1968 and 1975, researchers initially identified four antigenic sites (A–D) in which amino acid replacements appeared to affect antibody binding (Wiley, Wilson, and Skehel 1981). The boundaries of each site on the HA molecule were not known because of the limited number of escape mutants available. Abundant sequence data have been available for HA genes of field strains. However, the disadvantage of relying solely on these data is that not all amino acid replacements detected are antigenic. Many of the replacements may reflect a neutral polymorphism present in circulating viral genes. The most definitive studies involving antibody escape mutants were performed by Caton et al. (1982), who used an extensive panel of monoclonal antibodies to demonstrate that many of the external residues of the globular region of the HA can affect the ability of antibodies to bind and neutralize influenza viruses of the H1N1 subtype. These authors demonstrated that amino acid changes in the monoclonal antibody escape mutants in their study clustered into five antigenic regions that exhibited various degrees of overlap. Although studies of monoclonal antibody escape mutants of the H3 subtype have not been as extensive, the original proposal that there were four antigenic regions has been modified to include a fifth (A–E) (Wiley and Skehel 1987; Wilson and Cox 1990), and it is apparent that much of the surface of the globular head of the HA molecule has been altered by amino acid substitutions during the circulation of the H3 subtype since 1968.

Two hundred fifty-seven naturally occurring field strains obtained and sequenced since 1968 were used to further characterize the extent and nature of the previously described antigenic sites by locating their positions on the three-dimensional structure of the A/Aichi/2/68 HA using the MidasPlus (Ferrin et al. 1995) software system. The water-accessible surfaces of residues identified as sites A–E were displayed, and a variant residue was assigned to an antigenic site when it was located directly in the area of an existing antigenic site

and had more than 80% of its surface exposed (not buried). Variant residues that appeared to belong to more than one antigenic site were assigned a site based on proximity. The assignment of the positively and negatively selected codons to antibody-combining sites A–E is shown in table 3.

The fact that many positively selected codons are in the receptor-binding site, and that all of them are in antibody combing sites, helps to explain their uneven distribution on the gene. Sixteen of the 18 (89%) codons under positive selection on internal branches are located between codon 121 and codon 226, a region comprising 106 codons out of 329 (32%). The proximity of the receptor-binding site to antibody-combining sites, along with the observation that viruses with changes in receptor specificity may exhibit corresponding antigenic changes, may help to explain this distribution. It has been reported that amino acid changes that occur during antigenic drift of the HA of influenza viruses of the H1 and H3 subtypes are concentrated in antibody-combining sites A and B and the residues between them that make up the receptor-binding pocket; this suggests that change in this area is favored by selection in nature (Cox and Bender 1995). Our results here are consistent with those previous observations.

Significance

We identified 18 codons in the HA gene that appear to be under positive selection to change. All 18 codons are associated with antibody-combining sites, and five help form the receptor-binding site of the HA. This result is consistent with the hypothesis that mutation in the HA gene is a process by which the human influenza virus can escape immune surveillance. We are now in a position to determine whether mutation in this set of codons is useful for predicting future epidemics.

Acknowledgments

R.M.B. is supported by a University of California Directed Research and Development grant in collaboration with the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory. We gratefully acknowledge the technical expertise of Huang Jing and the discussion and critical comments of Dr. Kanta Subbarao.

LITERATURE CITED

CATON, A. J., G. G. BROWNLEE, J. W. YEWDELL, and W. GERHARD. 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell **31**:417–427.

COX, N. J., and C. A. BENDER. 1995. The molecular epidemiolgy of influenza viruses. Semin. Virol. **6**:359–370.

FERRIN, T. E., C. C. JUANG, L. E. JARVIS, and J. LANGRIDGE. 1995. The MIDAS display system. J. Mol. Graph. **6**:13–27.

FITCH, W. M., R. M. BUSH, C. A. BENDER, and N. J. COX. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA **94**:7712–7718.

FITCH, W. M., J. M. E. LEITER, X. LI, and P. PALESE. 1991. Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. USA **88**:4270–4274.

GOLDING, B., and J. FELSENSTEIN. 1990. A maximum likelihood approach to the detection of selection from a phylogeny. J. Mol. Evol. **31**:511–523.

GOLDING, G. B. 1987. The detection of deleterious selection using ancestors inferred from a phylogenetic history. Genet. Res. Camb. **49**:71–82.

GOLDING, G. B., C. F. AQUADRO, and C. H. LANGLEY. 1986. Sequence evolution within populations under multiple types of mutation. Proc. Natl. Acad. Sci. USA **83**:427–431.

GUBAREVA, L. V., J. M. WOOD, W. J. MEYER, J. M. KATZ, J. S. ROBERTSON, D. MAJOR, and R. G. WEBSTER. 1994. Codominant mixtures of viruses in reference strains of influenza virus due to host cell variation. Virology **199**:89–97.

HARDY, C. T., S. A. YOUNG, R. G. WEBSTER, C. J. NAEVE, and R. J. OWENS. 1995. Egg fluids and cells of the chorioallantoic membrane of embryonated chicken eggs can select different variants of influenza A (H3N2) viruses. Virology **211**:302–306.

HUGHES, A. L. 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of Plasmodium falciparum. Mol. Biol. Evol. **9**:381–393.

HUGHES, A. L., and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167–170.

———. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA **86**:958–962.

INA, Y., and T. GOJOBORI. 1994. Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. Proc. Natl. Acad. Sci. USA **91**:8388–8392.

ITO, T., Y. SUZUKI, A. TAKADA, A. KAWAMOTO, K. OTSUKI, H. MASUDA, M. YAMADA, T. SUZUKI, H. KIDA, and Y. KAWAOKA. 1997. Differences in sialic acid–galactose linkages in the chicken egg amnion and allantois influence human influenza virus receptor specificity and variant selection. J. Virol. **71**:3357–3362.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MESSIER, W., and C.-B. STEWART. 1997. Episodic adaptive evolution of primate lysosymes. Nature **385**:151–154.

NAKAJIMA, S., K. NAKAJIMA, and A. P. KENDAL. 1983. Identification of the binding sites to monoclonal antibodies on A/USSR/90/77 (H1N1) hemagglutinin and their involvement in antigenic drift in H1N1 influenza viruses. Virology **131**:116–127.

ROBERTSON, J. S. 1993. Clinical influenza virus and the embryonated hen's egg. Rev. Med. Virol. **3**:97–106.

ROCHA, E. P., X. XU, H. E. HALL, J. R. ALLEN, H. L. REGNERY, and N. COX. 1993. Comparison of 10 influenza A (H1N1 and H3N2) haemagglutinin sequences obtained directly from clinical specimens to those of MDCK cell- and egg-grown viruses. J. Gen. Virol. **74**:2513–2518.

ROGERS, G. N., J. C. PAULSON, R. S. DANIELS, J. J. SKEHEL, I. A. WILSON, and D. C. WILEY. 1983. Single amino acid substitutions in influenza haemagglutinin change receptor binding specificity. Nature **304**:76–78.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, G. T. HORN, K. B. MULLIS, and H. A. ERLICH. 1988. Primer-directed enzymatic amplification of DNA with a thermostabile DNA polymerase. Science **239**:487–491.

SWOFFORD, D. L. 1993. PAUP (phylogenetic analysis using parsimony). Version 4.0.0d60. Illinois Natural History Survey, Champaign.

WEIS, W., J. H. BROWN, S. CUSACK, J. C. PAULSON, J. J. SKEHEL, and D. C. WILEY. 1988. Structure of the influenza virus

haemagglutinin complexed with its receptor, sialic acid. Nature **333**:426–431.

WILEY, D. C., and J. J. SKEHEL. 1987. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu. Rev. Biochem. **56**:365–394.

WILEY, D. C., I. A. WILSON, and J. J. SKEHEL. 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature **289**:373–378.

WILSON, I. A., and N. J. COX. 1990. Structural basis of immune recognition of influenza virus hemagglutinin. Annu. Rev. Immunol. **8**:737–771.

YANG, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

YUN-XIN FU, reviewing editor

Accepted July 7, 1999