Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution

Arthur Chun-Chieh Shih*, Tzu-Chang Hsiao*, Mei-Shang Ho†, and Wen-Hsiung Li‡§¶

*Institute of Information Science, †Institute of Biomedical Sciences, and ‡Genomics Research Center, Academia Sinica, Nankang, Taipei 115, Taiwan; and §Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637

Contributed by Wen-Hsiung Li, February 15, 2007 (sent for review January 30, 2007)

The HA1 domain of HA, the major antigenic protein of influenza A viruses, contains all of the antigenic sites of HA and is under continual immune-driven selection. To resolve controversies on whether only a few or many residue sites of HA1 have undergone positive selection, whether positive selection at HA1 is continual or punctuated, and whether antigenic change is punctuated, we introduce an approach to analyze 2,248 HA1 sequences collected from 1968 to 2005. We identify 95 substitutions at 63 sites from 1968 to 2005 and show that each substitution occurred very rapidly. The rapid substitution and the fact that 57 of the 63 sites are antigenic sites indicate that hitchhiking plays a minor role and that most of these sites, many more than previously found, have undergone positive selection. Strikingly, 88 of the 95 substitutions occurred in groups, and multiple mutations at antigenic sites sped up the fixation process. Our results suggest that positive selection has been ongoing most of the time, not sporadic, and that multiple mutations at antigenic sites cumulatively enhance antigenic drift, indicating that antigenic change is less punctuated than recently proposed.

amino acid switch | influenza virus | positive selection | virus evolution

Influenza virus A causes flu epidemics or even pandemics that can kill millions of people in 1 year (1–3). Based on the antigenic specificities of the HA or neuraminidase (NA) protein, the influenza A viruses have been divided into 16 HA (H1–H16) and nine neuraminidase (N1–N9) subtypes, respectively. The HA protein consists of two domains, HA1 and HA2, and HA1 contains all of the antigenic sites of HA. All subtypes of type A viruses are maintained in aquatic bird populations (2), but only H1N1 and H3N2 have been circular in human populations. In circulating influenza viruses, antigenic drift is a major process that accumulates mutations at the antibody binding sites in the HA protein, enabling the virus to evade recognition by hosts' antibodies (4). Because such mutations in H3 HA occur often and new variants tend to replace older ones quickly, the evolution of the HA gene of H3 is much faster than that of other subtypes (4, 5).

In the study of influenza evolution, it is important to know what kind of selection pressure operates on HA1 because it can enhance our understanding of influenza virus evolution as well as vaccine strain prediction. This is usually inferred by comparing the rates of nonsynonymous (K_a) and synonymous (K_s) substitutions in the sequences under study. Conventionally, $K_a > K_s$ suggests positive selection, $K_a < K_s$ suggests purifying (negative) selection, and $K_a = K_s$ means no selection [e.g., Li *et al.* (6)]. To avoid a large variance in estimation, the traditional methods for identifying positively selected sites require the use of many codon sites to compute average K_a and K_s, and, consequently, the results have usually been assigned to an amino acid residue region (7). But if positive selection had operated at only a few amino acid sites, these sites would not be identified if the average number of nonsynonymous substitutions was smaller than that of synonymous substitutions in the analyzed region (7–9). Therefore, a number of methods have been proposed to study selection on a site-by-site basis (7–14).

Using the K_a/K_s method and a method to determine positive selection along individual branches of a phylogenetic tree, Fitch *et al.* (9) and Bush *et al.* (8) identified, respectively, 14 and 18 amino acid sites in H3 HA1 as having undergone positive selection. On the other hand, without assuming that the K_a/K_s ratio was the same for all positively selected codon sites, Suzuki and Gojobori (7) detected only three positively selected codon sites and argued that the inferences of Fitch *et al.* (9) might have contained some false positives. However, this large difference could be because the method of Suzuki and Gojobori is less powerful than that of Fitch *et al*. More recently, using codon usage bias to distinguish between diversifying and purifying selection at a codon site, Plotkin and Dushoff (15) identified 25 codon sites as diverse codons from a comparison of 525 viruses isolated from 1968 to 2000. Thus, it remains unsettled as to how many amino acid sites in HA1 have undergone positive selection in the recent past. In this study we try to resolve this issue by developing a new approach.

We shall also address the issues of whether positive selection on HA1 is absent most of the time during evolution, that is, it is punctuated, and whether antigenic change in HA1 is punctuated (epochal) as recently proposed (16–18). As will be seen below, our analyses suggest that positive selection has been ongoing most of the time and that antigenic change accumulates over time with occasional large changes due to multiple mutations at antigenic sites in HA1.

Results

Frequency Diagrams of Amino Acid Residues. The frequency diagram at an amino acid residue site shows the frequency changes of the amino acids at the site over years or flu seasons. It can readily reveal the temporal dynamics of mutations at any specific site. Such diagrams clearly demonstrate that sites 156 and 145 of HA1 have undergone multiple substitutions from 1968 to 2005 (Fig. 1 *a* and *b*). In contrast, there was no significant frequency change at site 138, excepting one change from 1981 to 1982, and there was no significant change at site 194 during the entire period (Fig. 1 *c* and *d*), contradicting the inference derived from the ratio of nonsynonymous rate to synonymous rate (K_a/K_s) that these two sites were under positive selection (8). On the other hand, sites 83, 155, 172, and 189 were not identified previously (8, 9) as positively selected sites, but positive selection is clearly implicated because each of these sites is an antigenic site and underwent three or more substitutions (Fig. 1 *e–h*). These observations highlight the power of the frequency diagram approach for revealing the footprint of natural selection. [For a

Author contributions: A.C.-C.S., M.-S.H., and W.-H.L. designed research; A.C.-C.S., T.-C.H., and W.-H.L. performed research; A.C.-C.S. and T.-C.H. analyzed data; and A.C.-C.S., M.-S.H., and W.-H.L. wrote the paper.

The authors declare no conflict of interest.

¶To whom correspondence should be addressed. E-mail: whli@uchicago.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701396104/DC1.

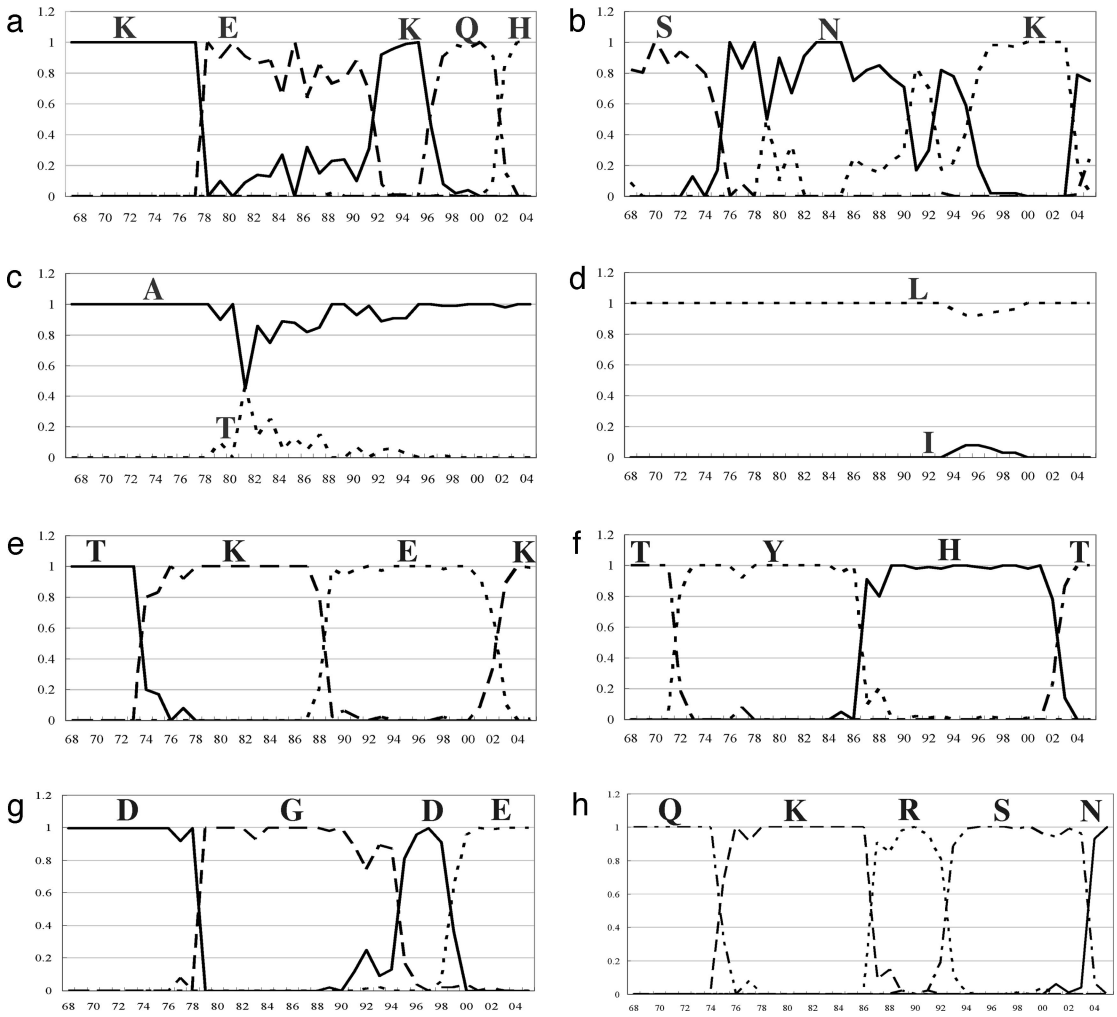

Fig. 1. Frequency diagrams of six sites. (*a* and *b*) Frequency changes at residue sites 156 (*a*) and 145 (*b*) were highly dynamic. (*c* and *d*) Sites 138 (*c*) and 194 (*d*) were identified as under positive selection by Bush *et al.* (8) but did not undergo major frequency change over time. (*e–h*) Sites 83 (*e*), 155 (*f*), 172 (*g*), and 189 (*h*) each underwent three or more substitutions but were not identified by Fitch *et al.* (9) or Bush *et al.* (8) as positively selected sites.

full list of sites 1–312 of HA1, see supporting information (SI) Fig. 5.]

Frequency Switches. Frequency diagrams also accentuate drastic changes in mutant frequencies, such as when a new allele quickly predominates over the other alleles at a residue site. We define a "frequency switch" as the replacement of one major amino acid by another between successive years. We further define an "effective switch" as a frequency switch in which the new dominant amino acid at the site later became fixed or almost fixed in the population for at least 1 year; otherwise, the switch is said to be an "ineffective switch." We identified 95 effective switches and 35 ineffective switches (Fig. 2). The 95 effective switches occurred at 63 residue sites, all but six of which are located in known antigenic epitopes; eight of them are also in receptor binding sites (SI Table 1). Thus, immune selection and structural constraints seem to be the major factors that determine the evolutionary fate of a mutation.

Transition Times. Information on how fast the 95 substitutions occurred can be enlightening. For this purpose we computed the transition time from the first appearance of a mutation in the sample to its fixation (or almost fixation) in the population (19). The transition time highlights the dynamics of a sweeping allele. Because the yearly sample size before 1987 was comparatively small, for this part of the analysis we consider only the 51 fixations that occurred after 1986. The shortest transition time was 4 years, and the longest one was 32 years (Fig. 3). There are 38 transition times $\leq$15 years and only 13 transition times $>$15 years. These transition times are far shorter than the conditional fixation time (T) expected for a neutral mutation, which would be 80,000 days or 219 years, if we use a conservative estimate of 10,000 for the effective population size (N_e) of the virus and an interhost generation time of 4 days ($T = 2\ N_e$ generations; see ref. 20).

Comparison with Previous Results. The facts that almost all of the 95 substitutions occurred at antigenic sites and that the transition time for each substitution was short suggest that most of the 63 sites were under positive selection; this should be especially true for the 23 sites that have undergone at least two substitutions during the period. It is informative to compare these 63 sites with those of previous inferences. Twelve of the 14 positively selected sites inferred by using the K_a/K_s test of Fitch *et al.* (9) and 15 of the 18 positively selected sites inferred in Bush *et al.* (8) overlap with our 63 sites. These agreements between their results and ours suggest that the methods they used are considerably more powerful than the method used by Suzuki and Gojobori (7),

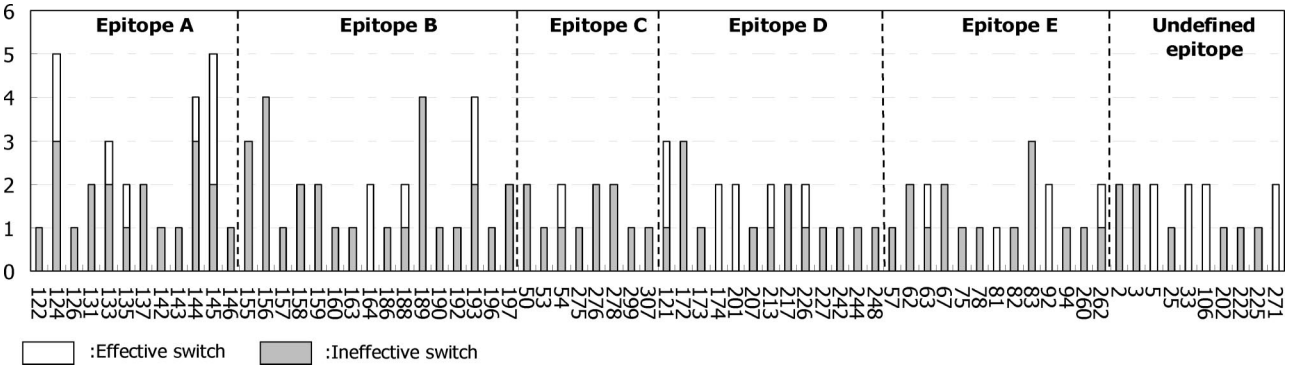

Fig. 2. Frequency switches. Numbers of effective and ineffective residue frequency switches are shown. Gray bars represent the effective switches, and open bars represent the ineffective ones.

which detected only three positively selected sites. Additionally, all of the 25 sites inferred from codon usage bias (15) overlap with our 63 sites. Furthermore, the 11 parallel replacements in HA1 identified by Wolf *et al.* (16) were identified by our analysis, except for V546A, which is in HA2 and thus not included in our analysis. These comparisons show that our method has a higher detection power than previous methods.

Simultaneous Multiple Fixations. A most striking observation from the frequency diagram analysis is the propensity of simultaneous (parallel) multiple amino acid fixations to occur in the same year (Fig. 4). We found 7 years (1973, 1976, 1978, 1979, 1984, 2001, and 2004) each with seven or more fixations, 1 year (2005) with four fixations, 5 years (1990, 1992, 1995, 1997, and 1999) with three fixations, and 3 years (1989, 1998, and 2002) with two fixations. Several cases of simultaneous substitutions have previously been noted (16), and, remarkably, we found that simultaneous substitutions account for 88 of the 95 substitutions we observed (Fig. 4). The remaining seven fixations were dispersed over 22 years. Note also that no year with more than four fixation events was found from 1985 to 2000, whereas from 2000 to 2005 there were 2 years each with more than nine fixations. Thus, the fixation events were highly unevenly distributed over time, and this may imply that genetic evolution is not continual, but punctuated. However, a comprehensive temporal map of the amino acid frequency (Fig. 4) reveals that the amino acid substitution process was ongoing in nearly all of the years under study, suggesting that positive selection was ongoing continuously. For example, positive selection seems to have occurred shortly after H3N2 influenza virus entered the human population, so that the frequency of the mutation V78G suddenly increased from absence in the sample in 1969 to 0.33 in 1970. This positive selection was then enhanced by mutations at other antigenic sites, leading to the fixation of nine mutations in 1973.

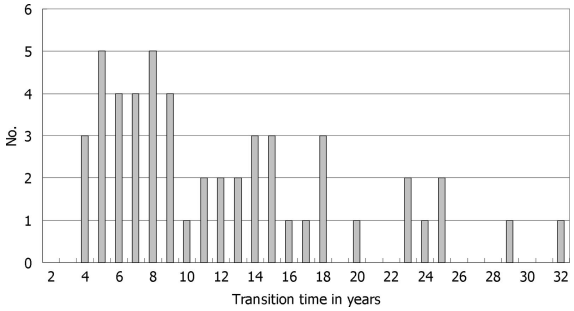

Fig. 3. Transition times. Shown is the distribution of transition times for substitutions from 1987 to 2005.

This was then immediately followed by positive selection for the mutation D63N in 1974, and positive selection was apparently enhanced by other mutations in subsequent years, leading to eight simultaneous fixations in 1976. Such contiguous intervals of positive selection can be seen in the rest of Fig. 4. Therefore, positive selection is evidently not punctuated. In summary, the results of our analysis on the yearly amino acid frequency do not support the scenario of long intervals of stasis (absence of positive selection) punctuated by positive selection (16).

Discussion

Hitchhiking or Positive Selection? Simultaneous fixations can derive from enhancement of antigenic drift, compensatory mutations to retain function, or hitchhiking (17). Note that antigenic enhancement and compensatory mutation both lead to positive selection. Compensatory mutations have been found to occur at receptor binding sites: subneutralizing level of monoclonal antibody specifically against residue 155 resulted in resistant mutants with amino acid substitutions occurring at sites 190 and 226; all three sites are receptor binding sites (SI Fig. 6 and ref. 21). Hitchhiking may explain some, but not the majority, of the fixations because only eight of the 95 fixations were not in the known antigenic epitopes. If hitchhiking was prevalent, it should have occurred at many nonantigenic residue sites, because the number of nonantigenic residue sites is larger than that of antigenic ones (eight substitutions among 181 nonantigenic sites vs. 87 substitutions among 131 antigenic sites; $P < 10^{-19}$).

A close inspection of the frequency changes for each case of multiple fixations gives insight into the mechanisms that govern their dynamics. First, we use the simultaneous fixations in 1973 as a case study. The first observed mutation was 144D, which appeared at a frequency of 0.09 in 1968. Its frequency increased to 0.2 in the next year, but it disappeared from the sample (sample frequency 0) in the next 2 years. So 144D might first have a selective advantage for its frequency to increase to 0.2, but the advantage might not have been sufficient to carry it to fixation. Alternatively, 144D might have been out-competed by mutations at other sites, e.g., 63D, which was on the rise from 1968 to 1970. Interestingly, in 1972 the frequency of 144D suddenly increased to 0.94. This was unlikely to result from hitchhiking, because there was no other mutant with a frequency >0.88. It would be more likely that a combination of 144D with mutants at other antigenic sites conferred a strong antigenic drift, possibly of epistatic nature, so that the frequency increased rapidly. The second mutation to appear was 78G, which increased in sample frequency from 0 in 1969 to 0.33 in 1970 and then to 0.86 in 1971. Because there was no other mutant on the rise from 1969 to 1971 with such a high frequency, we invoke selective advantage as the most likely explanation for the rapid frequency increase of 78G.

EVOLUTION

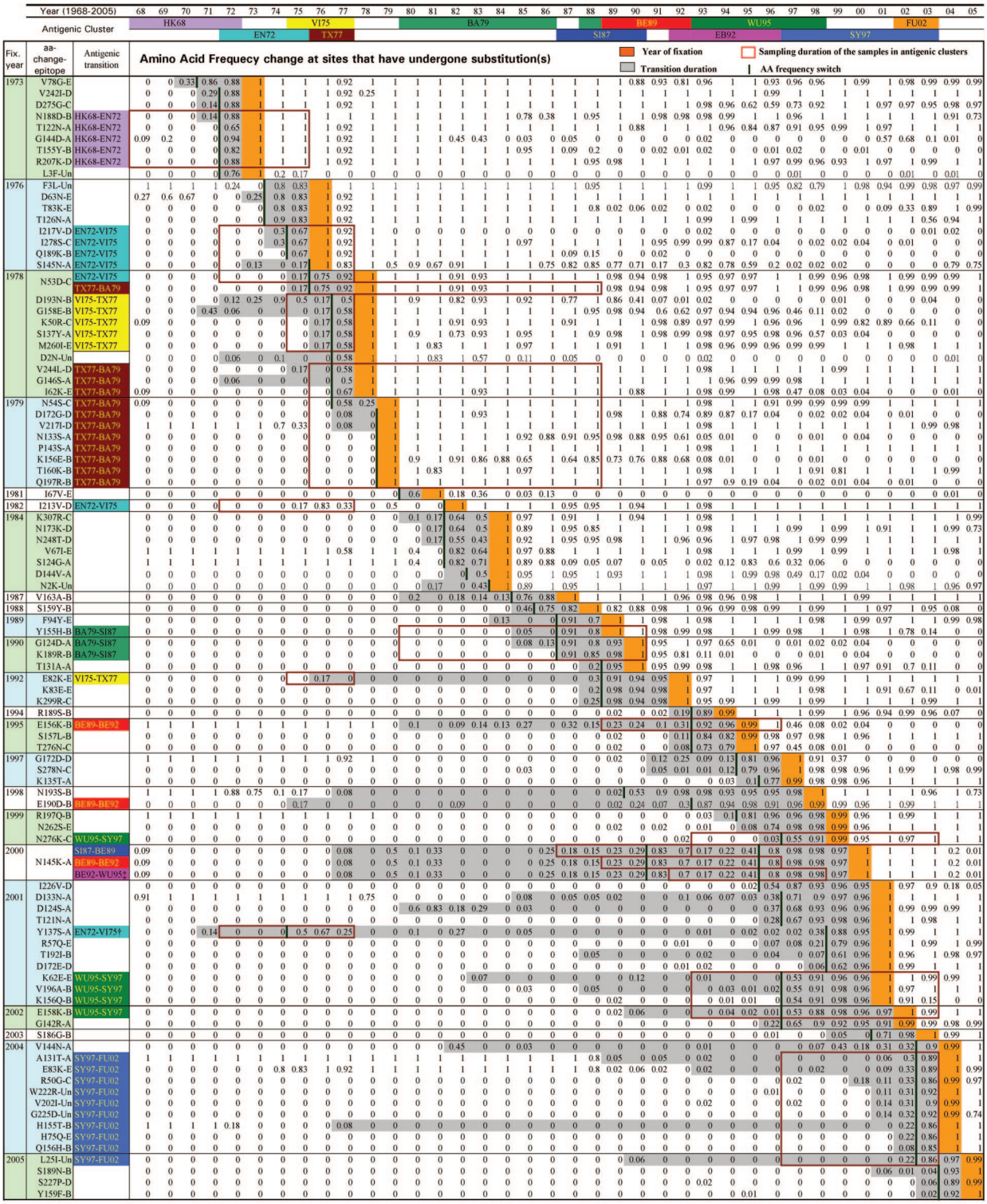

Fig. 4. A temporal map of amino acid frequencies on sites of the detected substitutions. The detected substitutions are grouped in rows by the year of fixation; the transition of each site is given in the second column. Information on antigenic clusters, extracted from Smith *et al.* (17), is included as marked on the top

Similar comments apply to the other mutations that became fixed in 1973. One important feature to note is that when several mutations appeared together they quickly became fixed in the population, suggesting a cumulative or epistatic effect.

Second, for a case of two simultaneous fixations, it is relatively simple to see what happened, so we consider the two fixations in 1989. Mutant 94Y was first observed in 1984 at a sample frequency of 0.13 but disappeared from the sample in 1985 and 1986, and 155H was first observed in 1985 at a sample frequency of 0.05 but disappeared from the sample in 1986. Apparently, individually each mutant did not have a selective advantage large enough to carry them individually to a high frequency. Surprisingly, they reappeared together in 1987 at a sample frequency of 91% and quickly went to fixation together in 1989. A simple explanation is that the two mutations enhanced each other's antigenic effect to increase the selective advantage, so that they could go to fixation together. Another cluster of two fixations occurring in 1998 assumed a similar pattern of fixation process as described above.

One can examine the other cases of multiple fixations in a similar manner and find that in most cases the simplest explanation is that multiple mutations at antigenic sites increase antigenic effects, speeding up the fixation process and increasing the fixation probability. We note that, until recently, many isolates of HA1 were selected for analysis because of their antigenic diversity and so do not represent an unbiased sample of the influenza A virus. However, the substitution patterns in the earlier years (e.g., 1970s) are similar to those in recent years (Fig. 4), suggesting that this bias has had no strong affect on our results.

Antigenic Change. We now consider amino acid replacements together with data on antigenic changes. Smith *et al.* (17) used 273 HA1 sequences (isolates) and hemagglutination inhibition data to infer antigenic clusters and amino acid differences between clusters. We incorporate their data into the temporal map of amino acid frequency (Fig. 4) and note that the vast majority of the amino acid changes found between antigenic clusters eventually went to fixation in the population. For example, this is the case for the five amino acid changes, N188D, T122N, G144D, T155Y, and R207K, between the HK68 and EN72 clusters. However, there were cases where some of the amino acid changes between clusters did not become fixed in the population (bottom of Fig. 4). For example, four of the 10 amino acid changes between EN72 and VI75 (indicated by dark cyan in Fig. 4) failed to become fixed. These four changes, I230V, L164Q, F174S, and N193D, disappeared in 1978 because the antigenic cluster VI75 was rapidly replaced by TX77 (upper part of Fig. 4). Note, however, that a single mutation with a large antigenic change may not lead to a quick fixation. For example, the mutation N145K was estimated to have a very large antigenic effect (4 units) and to be responsible for the BE92–WU95 transition (17), but this mutation appeared in the sample as early as 1968, disappeared from 1969 to 1976, reappeared in 1977, disappeared in 1978, reappeared in 1979, disappeared in 1982, and reappeared in 1985 but did not become fixed in the population until 2000. Apparently, a single mutation may not be effective in evading herd immunity. These observations show that mapping the antigenic data into the amino acid frequency diagram can be highly informative.

The antigenic cluster model has led to the interesting inference that genetic change is gradual, whereas antigenic change is punctuated or epochal (17, 18). However, there is considerable variation in cross-immunity among isolates within a cluster, and the antigenic distance between two isolates in the same cluster can be considerably larger than that between two isolates from two adjacent clusters. For example, the three largest pairwise antigenic distances within the BE92 cluster are all >5 units (≈5.6, 5.5, and 5.5 units), whereas the shortest pairwise distance between BE92 and WU95 is only ≈1.2 units (17). Moreover, because only 273 HA1 sequences (isolates) were used in the study, the data did not include the antigenic changes in other available sequences. For these reasons, the current antigenic cluster model may imply a rather discrete nature of antigenic evolution. Indeed, if the transition from HK68 to EN72 occurred abruptly in 1972 as the model implies, then it is difficult to explain the rapid increases in frequency before 1972 for the mutations V78G, V242I, D275G, and N188D. It is likely that the antigenic drift had already accumulated to a significant extent before 1972 because of the mutations V78G, V242I, D275G, and N188D, which were all at antigenic sites, and then the new mutations T122N, possibly G144D, T155Y, and R207K in 1972 together added a large antigenic change, completing the transition from HK68 to EN72 (Fig. 4). As another example, the proposal of a sudden transition from EN72 to VI75 in 1975 cannot explain why the frequencies of F3L, D63N, T83K, and T126N had rapidly increased to 0.88 or higher before 1975 (Fig. 4). Again, it is likely that the antigenic drift had already accumulated to a significant extent before 1975. Similar comments apply to the other cases of antigenic cluster transitions (Fig. 4). In conclusion, antigenic change seems to be less punctuated than proposed (17, 18), and we propose instead that antigenic change accumulates most of the time with occasional large changes.

Concluding Remarks. From the analyses in this study, we propose the following scenario for the evolution of HA1. Positive selection appears to be ongoing most of the time. Occasionally, a single mutation at an antigenic site may confer sufficient antigenic drift for the virus to escape the existing herd immunity and become fixed in the population, e.g., mutants 67V and 213V became fixed in 1981 and 1982, respectively (Fig. 4). In the majority of cases, however, a mutation confers limited antigenic drift and its frequency increases only to a low or intermediate level. Subsequent mutations at other antigenic sites often significantly contribute additional effects, so that multiple fixations, possibly including hitchhikers, occur rapidly and (almost) simultaneously. However, we caution that we have not considered the effects of reassortment; to do this requires a different analysis (see ref. 22).

The present study provides some guidance for rational design of seasonal vaccines. If no new mutation occurs in a season and the frequency of a combination of certain existing mutations (e.g., 307R, 173K, 248T, 67I, 124G, 144V, and 2K in 1983) is on the rise, the combination is likely to be the prevailing strain in the next season (Fig. 4). On the other hand, if one or more mutations occur at antigenic sites and are in linkage with some existing mutations, the linked mutations will likely become the dominant strain in the next season. For example, mutations T122N, G144D, T155Y, and R207K, together with existing mutations V78G, V242I, D275G, and N188D, caused the HK68–EN72 antigenic cluster transition. In some cases, however,

of the columns according to the year of isolation, and the antigenic transition between two clusters is marked in the third column; the unmarked areas in the third column represent isolates devoid of antigenic cluster information in the Smith *et al.* study (17). The yearly frequency (year on top of each column) of the emerging amino acids in transit (amino acid residues of each site are shown on the left of each row in the second column) is presented as proportional frequency of 0~1. The original or transited amino acids of three antigenic transitions are different between our samples and those in the Smith *et al.* study (17): N137S, K145N, and S193D, as indicated by †, ‡, and * in the figure, respectively.

multiple mutations occur together and sweep the population in one season. For example, N133S, P143S, K156E, T160K, and Q197R were not seen in 1977 or earlier but suddenly became fixed simultaneously in 1978. In such a case, predicting the prevailing strain in the next season is very difficult. In any event, however, having a geographically broader coverage of virus surveillance would provide a better picture of the evolutionary dynamics of the influenza A virus and enhance our ability to select antigenically matched vaccine, although events of reassortment will remain major hurdles to optimizing the efficacy of influenza vaccine targeting HA1 antigens.

Materials and Methods

Data Collection. All sequences of the H3N2 HA1 domain were downloaded from National Center for Biotechnology Information on March 14, 2006. After removing those with a length shorter than 315 codons or without the record of the year of isolation, the number of the sequences became 2,248; their years of isolation were from 1968 to 2005. We used Muscle (23) to align the amino acid sequences. After eliminating the gap-rich regions at the carboxyl end of the alignment, the range of the final alignment is from position 1 to position 312. We clustered the sequences from the same year into one group and obtained 38 groups.

Amino Acid Frequency Diagram. If the number of sequences with amino acid a_k at the jth position at year t is $n(t, j, a_k)$, the amino acid frequency $f(t, j, a_k)$ at the jth position at year t is given by $f(t, j, a_k) = n(t, j, a_k)/N(t)$. Therefore, for site j and amino acid a_k we can show the amino acid frequency as a function of t in a diagram.

Frequency Switch. For the frequencies of two different amino acids at a given site j at year t, $f(t, j, a_k)$ and $f(t, j, a_m)$, we say there was a frequency switch between ja_k and ja_m between years t and $t+1$ if the following two conditions hold: (*i*) $f(t, j, a_k) + f(t, j, a_m) > 0.7$ and $[f(t+1, j, a_k) + f(t+1, j, a_m)] > 0.7$; and (*ii*) $[n(t, j, a_k) - n(t, j, a_m)]$ and $[n(t+1, j, a_k) - n(t+1, j, a_m)]$ have opposite signs and differ significantly in absolute value. The first condition checks whether these two amino acid residues formed the majority of amino acids at site j in years t and $t+1$. That is, either a_k or a_m has a frequency >0.35 and is the major allele in year t or year $t+1$ because any other amino acid at the site has a frequency <0.3 ($=1-0.7$). The second condition checks whether the change in frequency from year t to $t+1$ was statistically significant. Thus, when the two conditions hold, a switch in the major amino acid at the site has occurred.

To test the second condition above, we used a 1-year contingency table. The values in the cells of the table are the numbers of the residues at that site at year t and $t+1$. To deal with small sample sizes, we used the one-tailed Fisher exact test to examine whether there is a positive association of the values in the 1-year contingency table. Moreover, because the numbers of sequences in some years were very small, we also used a 2-year contingency table for the test.

When a frequency switch between ja_k and ja_m is identified between t and $t+1$, we do not know whether the new dominant amino acid would be fixed or almost fixed in the population at a later time. Thus, we call a frequency switch from residue ja_k to ja_m an effective switch, if the following condition holds: $f(t+\tau, j, a_m) < 0.99$, $0 \le \tau \le T_f - 1$ and $f(t+T_f, j, a_m) \ge 0.99$, where $T_f > 0$. We say that ja_k underwent an effective switch at t and became fixed at $t+T_f$.

Transition Time. In population genetics, the time required for the fixation of an allele depends on the initial frequency of the allele, its selective advantage or disadvantage, and the size of the population (20, 24). Because we do not know the time the mutation occurred, we cannot calculate the precise fixation time of an amino acid substitution. Instead, we consider the transition time, which is defined as the time period from the first time the mutant amino acid was observed in the sample to the first time the frequency reached $\geq$99%. However, we add 1 year to this time because when the mutant was observed in the sample its frequency was usually not low because of a small sample size. Moreover, in some cases, when a mutant amino acid was first observed, it was almost fixed in the sample. In this case, we assume that the transition time was 2 years; such cases occurred only before 1987. Because the yearly sample size before 1987 was <20, we calculate the transition times of the effective switches only from 1987 to 2005.

We thank Eddie Holmes, J. J. Emerson, Walter Fitch, and Feng-Chin Chen for suggestions. This work was supported by Taiwan Pandemic Influenza Vaccine Research and Development Program, Taiwan Centers for Disease Control; Thematic Project of Academia Sinica, Taiwan; the National Science Council of Taiwan; the Institute of Information Science and the Genomics Research Center, Academia Sinica, Taiwan; and the National Institutes of Health.

1. Cox NJ, Subbarao K (2000) *Annu Rev Med* 51:407–421.
2. Horimoto T, Kawaoka Y (2005) *Nat Rev Microbiol* 3:591–600.
3. Hilleman MR (2002) *Vaccine* 20:3068–3087.
4. Treanor J (2004) *N Engl J Med* 350:218–220.
5. Hay AJ, Gregory V, Douglas AR, Lin YP (2001) *Philos Trans R Soc London Ser B* 356:1861–1870.
6. Li WH, Wu CI, Luo CC (1985) *Mol Biol Evol* 2:150–174.
7. Suzuki Y, Gojobori T (1999) *Mol Biol Evol* 16:1315–1328.
8. Bush RM, Fitch WM, Bender CA, Cox NJ (1999) *Mol Biol Evol* 16:1457–1465.
9. Fitch WM, Bush RM, Bender CA, Cox NJ (1997) *Proc Natl Acad Sci USA* 94:7712–7718.
10. Suzuki Y (2004) *J Mol Evol* 59:11–19.
11. Nielsen R, Yang Z (1998) *Genetics* 148:929–936.
12. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) *Genetics* 155:431–449.
13. Huelsenbeck JP, Dyer KA (2004) *J Mol Evol* 58:661–672.
14. Yang Z, Swanson WJ (2002) *Mol Biol Evol* 19:49–57.
15. Plotkin JB, Dushoff J (2003) *Proc Natl Acad Sci USA* 100:7152–7157.
16. Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ (2006) *Biol Direct* 1:34.
17. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA (2004) *Science* 305:371–376.
18. Koelle K, Cobey S, Grenfell B, Pascual M (2006) *Science* 314:1898–1903.
19. Zanotto PM, Kallas EG, de Souza RF, Holmes EC (1999) *Genetics* 153:1077–1089.
20. Li W-H (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
21. Temoltzin-Palacios F, Thomas DB (1994) *J Exp Med* 179:1719–1724.
22. Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK (2005) *PLoS Biol* 3:e300.
23. Edgar RC (2004) *BMC Bioinformatics* 5:113.
24. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York).