

## The Wilhelmine E. Key 1999 Invitational Lecture



Since 1989, Walter M. Fitch has been a professor of Ecology and Evolutionary Biology in the School of Biological Sciences at the University of California, and served as chairman of the department from 1990 to 1995. His undergraduate education was in Chemistry at the University of California, Berkeley, where he also earned his Ph.D. in Comparative Biochemistry. He did post-doctoral work at Stanford and at University College, London before going to the University of Wisconsin where he worked for 24 years. Dr. Fitch has published over 150 papers in over 40 different professional journals. These studies in molecular evolution have earned him a number of awards including election to the National Academy of Sciences, the American Academy of Arts and Sciences, and the Linnean Society (London). He founded the Society for Molecular Biology and Evolution with Professor Masatoshi Nei in 1992, was its first president, and was the editor in chief of that society's journal for its first ten years, raising it to international prominence. That society has named its annual award for best student paper presented at its annual meeting, The Fitch Prize. Dr. Fitch was the 1996 Penn State Marker Lecturer. Considered one of the founding fathers of the field of molecular evolution, Dr. Fitch developed two of the major methods for inferring evolutionary relationships from amino acid and nucleic acid sequences. He contributed immensely to the methods of inferring other relationships from the evolutionary trees including estimates of the rates of evolution, Darwinian selection in specific positions of molecules, and the rooting of the tree of life. He has made contributions to numerous areas of biology including bacteriology, virology, the origin of life, the genetic code, taxonomy, genetics, the genetic code, and molecular biology. Dr. Fitch is active today developing still better analytical methods and studying evolution, especially of the human influenza virus in an attempt to see if one can better determine which circulating strain of influenza is most likely to produce next year's epidemic. This could be of great benefit in better determining from which flu strains the flu vaccine should be made. This (Wilhelmine E. Key) lecture was delivered on June 12, 1999, at the American Genetic Association Symposium, "Genome Diversity and Evolution," at the Pennsylvania State University, University Park, PA. From the Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697 (Fitch and Bush) and Influenza Branch, Centers for Disease Control and Prevention, Atlanta, Georgia (Bender, Subbarao, and Cox). Address correspondence to Walter M. Fitch at the address above or e-mail: wfitch@uci.edu.

## Predicting the Evolution of Human Influenza A

W. M. Fitch, R. M. Bush, C. A. Bender, K. Subbarao, and N. J. Cox

**We studied the evolution of the HA1 domain of the H3 hemagglutinin gene from human influenza virus type A. The phylogeny of these genes showed a single dominant lineage persisting over time. We tested the hypothesis that the progenitors of this single evolutionarily successful lineage were viruses carrying mutations at codons at which prior mutations had helped the virus to avoid human immune surveillance. We found evidence that eighteen hemagglutinin codons appeared to have been under positive selection to change the amino acid they encoded in the past. Retrospective tests show that viral lineages undergoing the greatest number of mutations in the positively selected codons were the progenitors of future H3 lineages in nine of eleven recent influenza seasons. Codons under positive selection were associated with antibody combining sites A or B or the sialic acid receptor binding site. However, not all codons in these sites had predictive value. Monitoring new H3 isolates for additional changes in positively selected codons might help identify the most fit extant viral strains that arise during antigenic drift.**

I wish first to express how honored I feel at having been chosen by you to join such a list as those who have previously been honored as Wilhelmine Key lecturers. I thank you very much. And while I'm giving thanks, I wish to acknowledge the enormous importance of my co-authors in the work I am reporting. The material in this paper has been presented in greater detail in Fitch et al. (1997) and Bush et al. (1999a, 1999b).

The hemagglutinin gene of human influenza virus (type A, subtype H3) is one of the fastest evolving genes known. This rapid rate of evolution ( $5.7 \times 10^{-3}$  nucleotide substitutions/site/year for the HA1 domain) is believed to reflect selective pressure by the human immune system (Fitch et al. 1997). The evolutionary process behind this is thought to be as follows. After an influenza viral particle enters a lung cell, the human immune system detects the virus and begins to make antibodies. Antibodies bind to the virus, marking it for destruction by white blood cells. During this process, however, the virus has had time to replicate, and while doing so, has undergone mutation. Viral progeny containing mutations that interfere with antibody recognition will replicate until yet more antibodies that specifically recognize the new mutant

hemagglutinin can be produced. A race between the human host and the viral parasite ensues that favors mutant viral strains. This type of natural selection is called positive selection to change. In this presentation we show, first, that there are codons in the hemagglutinin gene of influenza A that have been under positive selection to change in the past. Second, we show that strains having undergone the greatest number of amino acid replacements in the positively selected hemagglutinin codons are more likely to be the progenitors of future generations of the influenza virus.

### Positive Selection

A codon is considered to have been under positive selection to change if it shows a significant excess of non-silent, as opposed to silent nucleotide substitutions. A non-silent mutation changes the encoded amino acid, whereas a silent mutation does not. We determined the number of non-silent ( $p$ ) and silent ( $q$ ) mutations that occurred in each codon in the HA1 domain of the hemagglutinin gene between 1983 and 1997 using a maximum parsimony tree constructed from 357 hemagglutinin sequences. Under binomial expectations, the probability of finding, for

**Table 1. Features of the 18 positively selected codons**

Codon	Features
121	D, F
124	A
133	A, F
135	A, F, R
138	A, F, R
142	A
145	A, F
156	B, F
158	B
186	B, F
190	B, F, R
193	B, F
194	B, F, R
197	B
201	D
226	D, F, R
262	E
275	CF

Column 1 shows the codon number for each of the 18 codons found to be under positive selection. Column 2 shows features of those codons, with the letters A–E indicating the regions that have been designated as in or adjacent to antibody combining sites, F indicating that it is among the fast-evolving positions, and R indicating that it is in the receptor binding site.

example, eight non-silent and only one silent mutation in a particular codon is  $9!p^8q^1/(8!1!)$ . If  $p = 0.45$ , the probability of obtaining the 8:1 ratio is 0.008. We consider a codon to show a significant deviation from binomial expectations if the probability of the observation is less than 0.05. Thus, our example codon shows evidence that it has been under positive selection to change in the past. This test can only be applied to codons that have undergone a total of at least four mutations, thus we were only able to test 38 of the 329 codons. Eighteen of these codons were under positive selection to change (Table 1). Details of this analyses are reported in detail in Bush et al., 1999b.

### Predicting Evolution

The ultimate goal of this work was to develop a method that will allow us to determine, at any given point in time, which, from among a collection of influenza strains, is most likely to be the progenitor of future influenza lineages. We demonstrate our method in Figure 1. Figure 1a is a maximum parsimony tree constructed using 173 influenza hemagglutinin sequences collected between 1983 and September 30, 1994, the date we use as the end of the 1993–1994 influenza season. This tree has a single main lineage, shown in bold, which we refer to as the trunk of the tree. The average life of a non-trunk lineage is only about 1.5 years. The trunk represents the only evolutionarily suc-

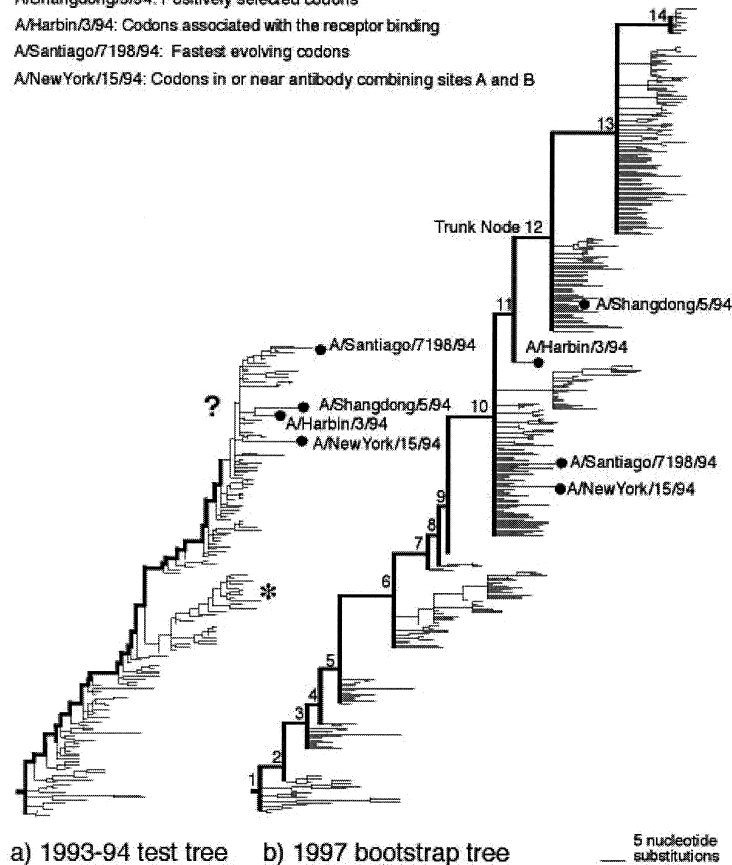
Predictive Isolate: Codon set

A/Shangdong/5/94: Positively selected codons

A/Harbin/3/94: Codons associated with the receptor binding

A/Santiago/7/198/94: Fastest evolving codons

A/NewYork/15/94: Codons in or near antibody combining sites A and B



**Figure 1.** Predicting the evolution of influenza A hemagglutinin. The tree on the left (Figure 1a) shows the evolution of the HA1 domain of the hemagglutinin gene of human influenza A from 1983 through the 1993–1994 influenza season. The tree on the right (Figure 1b), which contains isolates from 1983 through the 1996–97 influenza season, contains many fewer branches than the tree in Figure 1a because we collapsed nodes for which we lacked strong statistical support. Only one isolate from Figure 1a, A/Shangdong/5/94, descends from node 12 of Figure 1b, the uppermost node from which any of the isolates from Figure 1a descend. Results of tests of our main hypothesis and three alternative hypotheses are shown. Only the prediction test performed using the positively selected codons chose A/Shangdong/5/94, and thus was a successful prediction test. This figure is from Bush et al. (1999b).

cessful lineage because it is the only lineage that does not go extinct. Figure 1a was drawn as it would have appeared in 1994, with the trunk stopping short of the top of the tree. We would not be able to determine, until more time had past and all but one of the uppermost lineages on Figure 1a had gone extinct, where the trunk would emerge from the top of this tree. Our hypothesis is that the isolate on the tree in Figure 1a having undergone the greatest number of mutations in positively selected codons would be the progenitor of future lineages. We located this “predictive isolate”, A/Shangdong/5/94, by counting all of the amino acid replacements that occurred at the positively selected codons along each lineage on Figure 1a, from the root, or lower left end of the trunk, to the end of the 173 terminal branches. To determine whether future lineages, i.e., the trunk, eventually descended from A/Shangdong/5/94, we found the location of

all of the isolates from Figure 1a on Figure 1b. For a successful prediction test, the predictive isolate from Figure 1a must be as far up the trunk of Figure 1b as possible. Figure 1b. Figure 1b was constructed using sequences from 357 isolates collected between 1983 and 1997.

Taken together, Figures 1a and 1b illustrate a successful prediction test. A/Shangdong/5/94 was further up the trunk of Figure 1b than any of the other isolates from Figure 1a, and thus it was the isolate most closely related to future lineages. We performed such retrospective tests over an eleven year period, from the 1986–1987 through the 1996–1997 influenza seasons. Our prediction method was successful in nine of the eleven retrospective tests, and successful in every one of the last eight influenza seasons (Table 2).

We tested the probability that we could have obtained an equal or greater number of successful tests simply by chance in

**Table 2. Results of retrospective prediction tests for 11 recent influenza seasons using codons under positive selection and 7 alternative codon sets**

Codon sets	Number of codons	86–87	87–88	88–89	89–90	90–91	91–92	92–93	93–94	94–95	95–96	96–97	Successes
Positively selected	18	<b>5</b>	5	5	<b>8</b>	<b>8</b>	<b>10</b>	<b>10</b>	<b>12</b>	<b>13</b>	<b>13</b>	<b>14</b>	9
AB	41	<b>5</b>	5	<b>6</b>	<b>8</b>	7.0	<b>10</b>	<b>10</b>	10	12.0*	<b>13</b>	<b>14</b>	7
AB but not under positive selection	28	<b>5</b>	5	5.5	5.5	5.5	8.7	9.6	10	12	12	12	1
CDE	90	1	1	<b>6</b>	6	6	8	<b>10</b>	10	10	<b>13</b>	13	3
RBS	16	3.7	5	<b>6</b>	6	6	<b>10</b>	<b>10</b>	11	12.5	<b>13</b>	13.1*	4
RBS but not under positive selection	11	4.1	5	5	5	5	5	5	5	5	5	5	0
Fast	20	3	3	<b>6</b>	6.7*	7.0	<b>10</b>	<b>10</b>	10	<b>13</b>	<b>13</b>	<b>14</b>	6
Fast but not under positive selection	18	3	3	<b>6</b>	6	6	6	<b>10</b>	10	10	12	13	2
Top possible trunk node		5	6	6	8	8	10	10	12	13	13	14	

Positively selected: the set of 18 codons under positive selection.

AB: codons in or near antibody combining sites A and B.

CDE: codons in or near antibody combining sites C, D, and E.

RBS: codons associated with the sialic acid receptor binding site.

Fast: codons undergoing the greatest number of amino acid replacements.

The cells of the table indicate the trunk nodes on the 1997 bootstrap tree (Figure 1b) from which the predictive isolates resulting from each test descended. Successful tests (in bold) are those in which the predictive isolate descended from the uppermost possible node available on the 1997 tree (see bottom row). The right-hand column shows the total number of seasons in which each codon set produced a successful test. Cell entries with a decimal point or marked by an asterisk contain results that varied among replicate tests that were not described in this lecture, but which can be found in Bush et al. (1999b).

two different ways. First we determined the probability that a randomly chosen strain would produce a successful prediction test. To use Figure 1 as an illustration again, this probability is equal to the fraction of 173 isolates from Figure 1a that descend from the node marked 12 on Figure 1b. Node 12 is the uppermost trunk node from which any of the isolates from Figure 1b descend. Only one isolate from Figure 1a, A/Shangdong/5/94, descends from node 12 on Figure 1b, thus the probability of having obtained a successful result by chance is 1/173, or 0.6%. Across the eleven test years, this probability ranged from 0.6 to 19.7%, and averaged 6.8%. We also determined the probability that prediction tests done using 18 randomly chosen codons, rather than the 18 positively selected codons, would have produced a successful test. This probability, determined using one thousand randomly chosen codon sets, ranged from 0.6 to 49.0% across the 11 test years, with an average of 10.0%.

### Search for a Causal Explanation

A successful prediction test implies that changing the amino acid encoded by the 18 positively selected codons was adaptive, but it does not tell us why these mutations were selectively advantageous. We sought a causal explanation for our results by repeating the prediction tests using alternative codon sets that were of known function and that contained subsets of the

positively selected codons. If these alternative codon sets produced a greater number of successful prediction tests than the set of 18 positively selected codons, we might be able to discern the function under selection. This conclusion would only hold, however, if the alternative codon set also produced a successful prediction test when the subset of its codons that were also under positive selection was removed. We tested three alternative codon sets in this manner. 1) The set of 41 codons thought to lie in or near antibody combining sites A and B, sites that have been associated with antigenic drift; 2) the set of 90 codons in the other three antibody combining sites (C-E); and 3) the set of 16 codons that make up the sialic acid receptor binding site. The results of these tests (Table 2) reveal that none of the alternative codon sets of known function work as well as the positively selected codons in the prediction tests, and that they perform very poorly when the subset of positively selected codons was excluded from the alternative sets. Since the positively selected codons are among the most quickly evolving codons in the hemagglutinin, we also tested the hypothesis that any mutation is adaptive, and that positive selection is irrelevant, by repeating the prediction tests using the set of 20 most rapidly evolving codons that were not also under positive selection. This test was also less successful than the test using the 18 positively selected codons (Ta-

ble 2). Details of these prediction tests can be found in Bush et al., 1999b.

### Conclusions

We have identified a small set of rapidly evolving codons in the HA1 domain of the hemagglutinin gene of human influenza A subtype H3 in which non-silent mutations in the past appear to have been selectively advantageous. Strains with more mutations in these codons were more likely to be the progenitors of successful new lineages in nine of eleven influenza seasons. Although there is a significant overlap between the positively selected codons and the codons in or near antibody combining sites A or B and, to a lesser extent, codons associated with the sialic acid receptor binding site, codons associated with these sites of known function that are not under positive selection perform poorly in the prediction tests. Whether additional changes in these codons will confer a selective advantage in the future remains to be seen.

### References

- Bush RM, Fitch WM, Bender CA, and Cox NJ, 1999a. Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A. *Mol. Biol. and Evol.* 16:1457–1465.
- Bush RM, Bender CA, Subbarao K, Cox NJ, and Fitch WM, 1999b. Predicting the Evolution of Human Influenza A. *Science* 286:1921–1925.
- Fitch WM, Bush RM, Bender CA, and Cox NJ, 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Nat. Acad. Sci.* 94:7712–7718.