

2004-06-11
Positive paper 2001

Evolution of the hemagglutinin of Type B Human Influenza Virus
27 August, 2001

Walter M. Fitch and Geoff Graham

Department of Ecology and Evolutionary Biology

University of California - Irvine

Irvine, California 92697

NOTE:

This paper has not been submitted to a journal with peer review because editors properly require that the sequences be deposited in a proper, accessible archive. Most of the sequences used here are from Dr. Nancy Cox, Centers for Disease Control, Atlanta, Georgia. As soon as Dr. Cox releases those sequences to such an archive, I will submit the paper to a journal for peer review. Until that time this draft will serve the needs of the research community in communicating my work on this subject. I apologize for the eccentricity of the procedure.

Looking forward to the time when this can be published normally, all critiques are welcome. Thank you very much.

Walter M. Fitch

wfitch@uci.edu

ABSTRACT

We have studied the evolution of hemagglutinin of the B type human influenza virus and compared it to the results of a similar study of the H3 hemagglutinin for the A type human influenza virus. The alignment of the two consensus hemagglutinins for the A and B types show the sequences are clearly homologous (84 identical, aligned residues, 34 very similar) but that since their common ancestor there have been 19 indels (insertion/deletion events) involving 43 amino acid positions.

The rate of evolution of the B hemagglutinin gene, is 3.8×10^{-3} nucleotide substitutions/nucleotide/year, about a fourth of the 16×10^{-3} nucleotide substitutions/nucleotide/year for the H3 gene of the A type. The B type tree has two major lineages, the Yamagata and the Victoria. The Victoria lineage has a clearly discernible trunk but the Yamagata lineage does not have an immediately obvious trunk after 1992. The average age of the branch tips is 3.0 years, just over twice as long as that for the A type (1.4 years).

There were only two sites where A and B type homologous positions were both positively selected. This implies that the positions under positive selection change with time. Similarly, the potential glycosylation positions have migrated between the two types. Thus hemagglutinin structures performing crucial functions readily shift their location on the surface of the hemagglutinin in relatively brief periods of time.

We determined that the number of codons under positive selection in the off-tip branches using our previous method was only nine. These proved too few positions to usefully determine any predictive isolates, sequences that would accurately predict where the trunk of the tree will go.

Thus, compared to the A type hemagglutinins, there is no immediately obvious trunk to the Yamagata tree after about 1992, the B sequences evolve only a fourth as fast, and B lineages have an average life span 2.1 times longer than the A lineages. This may explain why more time must elapse before the location of the trunk is obvious. Except for there being no pandemics in the B types, the B type appears to be very similar to the A type, just slower.

INTRODUCTION

Human influenza viruses cause significant economic loss through mortality and morbidity. We have tried to discover principles underlying their virulence using a phylogenetic approach. In three previous papers (Fitch et al. 1997); (Bush et al. 1999a; Bush et al. 1999b) we showed that evolution of the H3 hemagglutinin of human influenza virus (A type) produces a single successful lineage, which we call the "trunk." All other branches of the phylogenetic tree die out. The average age of non-trunk tips, measured from the trunk, is about 1.4 years. As in those papers, substitutions and replacements here refer to changes in nucleotides and amino acids respectively. We also showed that there were eighteen amino acid positions that were under selective pressure to replace their amino acid, presumably to attenuate the ability of immune surveillance to destroy the virus. This paper examines the evolution of the hemagglutinin gene from 386 B type sequences, spanning the years 1964 to 1999 to determine whether the B type strains behave similarly to type A strains. We find that they are broadly similar, differing primarily in their rate of change.

METHODS AND MATERIALS

Materials. We examined 386 non-redundant hemagglutinin HA1 sequences from B type, human influenza viruses obtained from the CDC (Centers for Disease Control and Prevention), GenBank and the Los Alamos Database. The sequences are all 345, 346 or 347 amino acids long. The missing residue in those that are only 344 amino acids long is at position 163. The missing pair of residues in those that are only 343 amino acids long are at positions 163 and 164. The CDC sequences were to have been deposited in GenBank and have accession numbers xxxx-yyyy (see note at beginning). Our amino acid sequence numbering is for the 347 amino acid sequence. Thus, after position 164, our number for a B type position will be one larger than the Berton and Webster number (Berton and Webster 1985). For example, in their table 5 they align antigenic B position 201 with A position 192. In this case we align the same two residues but our B position number is 202. There are 115 isolates that are known not to have been grown in eggs, 114 isolates that are known to have been grown in eggs and 157 isolates that are of uncertain provenance.

Alignment of types A and B protein sequences. A consensus sequence of 412 A type (H3) sequences was made and compared to the consensus of 395 sequences of the B type hemagglutinin. Consensus sequences have, at each amino acid position, the most frequent amino acid at that position. This is the best estimate of their ancestral sequences in the absence of any knowledge of their tree structure. The two consensus sequences were then aligned by the NCBI Blast 2/ClustalW (Thompson, Higgins, and Gibson 1994) and that alignment improved by eye.

Phylogeny reconstruction. 390 influenza B hemagglutinin sequences were subjected to a rapid batch tree-finding procedure. To create a tree, we entered the sequences into PAUP in random order, directed PAUP (Swofford 1993) to build a tree by the parsimony method, and then directed PAUP to find shorter trees using the bisection-reconnection procedure while holding only four trees in memory. 500 trees were examined of which 37 were of the minimal length, 1467 steps. Ten of these trees were examined closely and the tree that seemed most typical of the group was saved for further analysis.

The sequence set and the tree were later purged of four sequences that formed abnormally long tip branches to give the final sequence set and tree topology. This left a sequence set of 386 sequences.

Calculation of ancestral sequence was done under the DELTRAN (delayed transformation) assumption, where changes are treated as having occurred as far from the root as possible. However, the analysis of geographical spread was done under the ACCTRAN (accelerated transformation) assumption, where changes are treated as having occurred as close to the root and as far from the tips as possible.

Determining the codons in the positively selected codon set. We located the codons under positive selection in humans by the methods presented before (Bush et al. 1999a; Bush et al. 1999b) which involved ignoring all changes on the tip branches of the phylogenetic tree.

RESULTS AND DISCUSSION

Alignment of type A and B protein sequences. Figure 1 shows the alignment of the consensus A and B type sequences. There are 19 indels (insertions/deletions; represented by gaps in the sequence) in the alignment, 4 in type B hemagglutinin and fifteen in type A. There are nine gapped residues in the B sequence and 34 in the A sequence, leaving 309 pairs of ungapped alignment positions, of which 86 (27%) have identical amino acids and another 39 (another 19%) have very similar amino acids. These values are quite sufficient to conclude that the hemagglutinins of types A and B are homologous and that the alignment shows the homologous pairings of amino acids although there are a few positions bordering

indels where an indel might be shifted a residue or two. None of our conclusions are affected by this possibility.

The alignment largely speaks for itself but bears comparison to the alignment of Krystal et al. in their Figure 5 (Krystal et al. 1982) They have 16 indels involving 40 amino acid sites in the B sequence without corresponding amino acids in the A sequence. In addition they have six indels involving 12 amino acids in the A sequence without corresponding amino acids in the B sequence. That is a total of 22 indels involving 52 gapped amino acids. We obtained our alignment using only 19 indels involving 43 gapped amino acids. Both our alignments show 84 identical amino acids in homologous positions.

Figure 1 also shows with **bold** lettering the nine positions that were determined to be under positive selection to change their amino acids. These may be compared to the functionally equivalent positively selected codon set in H3. Also shown are the sialic acid binding residues of H3.

Only two of the positively selected codons of the B type map to the same homologous codon as a positively selected A type codon, a number not significantly different from chance (1.05 codons expected if they were randomly distributed among the 309 ungapped positions or 2.07 codons if randomly distributed over the 165 codons that have varied). Nevertheless, many of them cluster in two regions of the hemagglutinin molecule that are clearly parts of the A type antigenic regions A and B.

Glycosylation sites. A glycosylation site is one that is encoded as a tripeptide, N-X-S/T, where N is asparagine, X is any amino acid and S/T is either serine or threonine. The position of the tripeptide is designated by the position of the N. We found twelve sites in the set of B sequences. The positions are indicated in Figure 1. There were fourteen glycosylation sites found in the 357 A type sequences used in our previous study (Bush et al. 1999b). They too are denoted by a diamond in Figure 1.

There were twelve locations where triplets of amino acids had a potential glycosylation site in at least one sequence (see Figure 1). There were four triplet positions (25, 59, 304, and 333) that were glycosylation sites in all 386 B type sequences. That does not mean that the triplet is unvaried, only that the replacements do not affect the presence of a glycosylation site. These four retained sites are either in the first 59 or last 34 amino acid positions of the sequences, positions largely away from where most of the selection for change is occurring. An exception is the tripeptide at position 145 which is present in all but three sequences. It has had at least six different triplets that all preserve the glycosylation site. That position is in the middle of one of the two most antigenic regions of the A hemagglutinins. The triplet at position 166 is found in 354 of the sequences. The other positions with one or more potential glycosylation sites are 45, 148, 163, 168, 197 and 233.

Type A hemagglutinin had fourteen glycosylation sites also well spread over the sequence. Only two of them are in codons homologous to the twelve B type glycosylation sites, namely pairs 59-63 and 148-144 (the two numbers connected by the hyphen are the pair of positions in the B and A sequences, respectively, The first pair is an extraordinarily well conserved site, present in all 386 B sequences and all 349 A sequences. In an extreme contrast to the first homologous pair, the second homologous pair had only one sequence from among the 386 B hemagglutinin sequences and only one sequence from among the 349 A hemagglutinin sequences with a glycosylation site at that position. The occurrence of only two homologous positions with a glycosylation site in both the A and B type sequences is essentially the random expectation. That must mean that the positions providing this function can and do change over time.

Indel evolution. Figure 2 shows three possible subtrees near the root of the hemagglutinin tree illustrating alternative possibilities for the evolution of codons 163 and 164 (designated as A and B in this section) for these indels. The left tree is how parsimony on our most parsimonious tree (Figure 3) would explain the indel relationships. The ancestral form is as in B/Lee/40 which is missing codon 163 and for which the two codons are represented here as ΔB with the Δ denoting a missing codon. Around 1970 a second gap appeared with the loss of codon 164 as readily seen in B/Osaka/70 and represented as $\Delta\Delta$.

This indel is sometimes taken to occur at position 165 rather than at 164. By 1985 (B/Canada/3/85), a form with both codons present arose in the Victoria lineage and is represented by AB. Each arrowhead indicates a path on which an indel occurs so that on the most parsimonious tree, there are five indel events.

We think the AB form might well have derived from the $\Delta\Delta$ form in a single event. The reasons for believing this are (1) that a single event involving two adjacent codons is more likely than two separate events involving non-adjacent codons, and (2) that the DNA sequence of the ancestral Victorian strain, as determined by its parsimonious reconstruction, is AAC AAC AAC AAC, which is most easily explained as a slipped strand mispairing on an ancestral AAC AAC AA. As shown in the center tree of figure 2, this will save one indel event reducing their number from five to four. Nerome et al earlier suggested that this kind of mechanism may be operating here (Nerome et al. 1998). Krystal et al (Krystal et al. 1983), their figure 3) also put these two gapped positions together but placed them one position to the left relative to our alignment.

The tree does not have a structure that would account for the indels most parsimoniously. It is clear that the ancestral form ΔB changed to AB around the 1970s and reverted back to ΔB around 1992 (see B/Beijing/210/92). The branching structure in the 1970s region requires an extra indel event which would disappear, however, if B/Osaka/70 were moved one branch over at a cost of only two additional nucleotide substitutions. Thus, as shown in the right tree in figure 2, the indel events could be explained by $\Delta B \rightarrow \Delta\Delta$ and then $\Delta\Delta \rightarrow AB$ at the origin of the Victorian lineage and also reverting $\Delta\Delta \rightarrow \Delta B$ later in the Yamagata lineage. The total number of indel events is thus reduced to three. The very high ratio of substitutions to indels (1467:5 or roughly 300:1) might justify rearranging the branches in that area of the tree as suggested by the right hand tree. We did not make that change however. Indel events were not considered in searching for the most parsimonious tree.

The tree. Figure 3 shows the most parsimonious tree we found. Of the 771 branches on the tree, 538 have (nucleotide) substitutions on them, 233 have none. Of the 347 coding sequence positions, 165 have (amino acid) replacements, 137 others have substitutions but no replacements, and 45 positions have no substitutions, and hence no replacements either. The total tree has 1418 substitutions on it of which 782 are silent and 636 are non-silent and of which 939 occur on the tip branches and 479 occur on internal branches, also called off-tip branches. The fraction of substitutions that cause replacements is 44.9%, which compares to the 27.7% estimated by Air et al (Air et al. 1990). Because we are using a parsimony method, our estimate of change is a lower bound. Thus there is a considerable discrepancy between our result and Air's. A possible explanation for this discrepancy is that we detect more changes because our branches represent smaller time intervals and hence the parsimony algorithm can pick up changes missed by Air et al.

The two major lineages are the Victoria (lower) and the Yamagata (upper). A trunk can readily be drawn for the Victoria lineage (shown by the heavier line in Figure 3) but not for the combined lineages nor for the Yamagata lineage after 1992.

Forty-four tip branches (11.8% of them) have zero length, indicating that 46 hemagglutinin isolates have a sequence that is immediately ancestral to some other isolate's hemagglutinin. These branches are not randomly distributed across the tree but are more frequent in recent years when the sampling was more intensive. The number of interior branches with zero length is 189 which constitutes 48.5% of all internal branches. This is perceived in the tree by the large number of polytomies, ancestral nodes (vertical lines) with multiple horizontal lines deriving from them.

Tip branch bias. One can ask whether the replacements are distributed randomly between the tip and interior branches. The null hypothesis of a random distribution would give each branch the same expected number of substitutions. The result, seen in table 1, is a significant excess in the tip branches. This occurrence of about twice as many substitutions in the tip branches as in the off-tip branches is very

close in magnitude to that seen in the A type analysis. This excess is probably the result primarily of the non random selection of the viruses to be sequenced, as explained in (Bush et al. 2000).

Silent and non-silent changes on tip and non-tip branches. Table 2 shows that there is no significant difference between the way the silent and non-silent changes distribute themselves between branch types.

The positively selected codons. When the sequences were analyzed using the off-tip data, we found nine positively selected codons (codons 73, 75, 121, 137, 149, 150, 202, 230 and 293) In addition, there are two more codons (197 and 199) that are positively selected off-tips but are clearly positively selected in eggs because together they have five times more replacements in the tip than in the off-tip branches.. More importantly, if these codons, in all isolates except those known not to have been grown in eggs, are replaced by totally ambiguous codons (NNN) all but one of the 22 interior branch replacements in codons 197 and 199 disappear. Thus these two positions are not included in the positively selected set.

Rate of evolution. We plotted (Figure 4) the replacement distance of the tips from the root against the year of isolation and found the evolutionary rate to be 1.77 replacements/year which equals 5.1×10^{-3} replacements/codon/year or 3.8×10^{-3} substitutions/nucleotide/year ($=5.1 \times 10^{-3}$ replacements /yr/codon divided by .449 substitutions/codon divided by 3 nucleotides/codon). The value of 3.8×10^{-3} substitutions/nucleotide/year may be compared to the 2.0×10^{-3} observed by Cox et al (Cox et al. 1993) as well as by Rota et al (Rota et al. 1993). The rate of 5.1×10^{-3} replacements/codon/year for the B type hemagglutinin may be compared to the A type rate which we reported as 16×10^{-3} replacements/codon/year (Fitch et al. 1997; Bush et al. 2000). Thus the B type hemagglutinin's evolutionary rate is only one third that of the A type. Air et al's value for this rate (5.3×10^{-4} replacements/codon/year; (Table 3 of (Air et al. 1990) is only one tenth that of ours. Their very low value might arise because their sample size was small and covered a narrower range of time. Additionally, they did not exclude sequences isolated prior to 1978 as we did and their four earlier sequences constituted 40% of their sample.

Average age of tip sequences measured in years from the trunk. The rates are 1.77 replacements per gene per year and 3.97 substitutions /year, respectively. The average distance of the tips is 11.19 substitutions from the trunk. Thus the average age of the tips, measured in years, is $11.19/3.97 = 3.01$ years. This is twice the average tip see where the A type Yamagata hemagglutinin (1.42 years) (Fitch et al. 1997). As the previous study could know where the trunk went after about three years, one can expect that with an extended study, covering perhaps another three or four more years, we might easily discern the Yamagata trunk in this region of the tree.

The average age is probably a slight underestimate because of two processes occurring. One is that the location of the trunk is uncertain for the most recent years of isolates. This problem is solved by not using any isolates whose trunk ancestor is not defined. The second process is that some of the lineages may not yet be extinct causing their distance to the root to be less than the age at extinction. This answer here is not totally satisfactory. We think there should be a correlation between when lineages go extinct and when the trunk location is clear. Thus we arbitrarily employed the same criterion for the second problem as for the first problem. It is the same set of isolates removed in either case; the isolates with uncertain trunk ancestors are excluded.

Fraction of mutations that are deleterious. If the 279 silent changes in the off-tip branches are neutral and all mutations are equally probable (making the proportion of silent changes equal to 25% (Fitch 1972), then one may estimate the effective total number of substitutions to equal $279/0.25 = 1116$. That means there would have been 837 non-silent mutations of which 133 were fixed leaving $(837-133 =)$

704 deleterious mutations that were removed from the sample by negative selection. Thus $704/837 = .872$ or 87.25% of all non-silent mutations were disadvantageous. The number for the H3 hemagglutinin evolutionary process was 75.6% (Bush et al. 1999b) so the two Types, A and B, are behaving similarly except that a larger fraction of the mutations are deleterious in the B type sequences. Zou et al presented data which they interpreted as showing positive selection in HA1 and negative selection in HA2 (Zou, Prud'homme, and Weber 1997). We show here simultaneous positive and negative selection in HA1, a circumstance that surely applies to most genes that are still active.

Inter-regional movement of human influenza B. We employed the same method as previously (Fitch et al. 1997), counting how many times an ancestor was determined to be from a different geographical region than its immediate descendant, thus implying migration of the virus from its ancestral region to its descendant region. Monte Carlo trials were performed 1000 times by a random reassignment of the tip regional designations. Because the mean number of migrations of the Monte Carlo trials, 243.4, was nearly two-thirds greater than the mean of the real data, 148 migrations, we normalized the Monte Carlo trials to the same number of interregional migrations by multiplying the Monte Carlo results by 0.608 (= $148/243.4$).

The result is shown in Table 3. There are many cases where computing chi-squared would lead to a significant difference. Unfortunately, 21 of them involve expected values that are less than 1.0 and seven more have expectations below 5.0 so these should all be rejected in tests of significance.

There were fifteen migrants from China to Hong Kong when only 4.85 were expected which gives a chi-squared value of 21.2 (df = 1) leading to a probability of 4×10^{-5} of occurring by chance. It is a quite reasonable result.

There was only one migration from China to Europe when there were expected to be 13.3. This yields a chi-squared value of 11.3 and a probability of 8×10^{-4} of occurring by chance, another reasonable result.

There were only four migrants from China to Europe when there were expected to be 16.7. This yields a chi-squared value of 9.62 and a probability of 2×10^{-2} of occurring by chance.

There were eight migrants from North America to China when there were expected to be 17.88. This yields a chi-squared value of 5.5 and a probability of 0.02. All four of these cases seem in reasonable accordance with what the field feels to be the nature of viral dispersal.

There were only three instances in which the amount of inter-regional spread was significantly different between the two directions (see Fig.xxx). For example, the number of migrants from China to Hong Kong was 15 while that from Hong Kong to China was only one migrant. If the null hypothesis is that *a priori* the probability of a migrant going from one region to another is the same in both directions, then the binomial $p = 5 \times 10^{-5}$. The null hypothesis is probably incorrect because the number of isolates from China and Hong Kong are unequal 99 versus 21. Nevertheless, one must wonder if it can be related to the fact that thousands of chickens are daily shipped from China into Hong Kong. Similarly, the number of occurrences of spread from China to other parts of Asia was 15 while that from (non-China) Asia to China was only 3 ($p = 7.5 \times 10^{-3}$). The third case is between North America and (non-China) Asia where the spread was 14 from N. America to Asia but only three in the reverse direction ($p = 1.2 \times 10^{-2}$). Europe and North America were each other's donor and contributor equally with substantial spread (nine occurrences) in each direction. Other than these four examples, all other cases where the number of movements was at least 6 involved the spread of North American stocks to other regions causing there to be more occurrences of spread out of North America than out of China.

One cannot reasonably assert that the flow of influenza B virus has a demonstrable pattern outside of the three significant cases but the result shows promise. There is clearly not a mostly one-way flow of the virus in a river-like system. China appears to be the more probable source of new viruses but that could not be confidently asserted on the basis of these data alone, especially since there were more outward migrations from North America than from China. Moreover, there is a correlation between

sample size and the number of spread events, both as source and recipient which must be evaluated before accepting that the observations are not accounted for by the binomial computation. More data should be helpful.

Is B type influenza qualitatively different from A type influenza? Air et al (Air et al. 1990) conclude that there is no discernable pressure on the hemagglutinin gene to change; that there seems to be much less selection, if any, to change the B type protein. We disagree. The nine human selected codons, having a significant excess of non-silent changes, directly contradicts a lack of selection in B type influenza. For these nine codons, their collective NS/S ratio is 51/3, the NS numbers being 17-fold that of the S numbers.

In only two cases (positions 137 and 149) are the positions of the B type human set of positively selected codons homologous to a positively selected codon in the A type human set. Nevertheless there is a considerable clustering of positively selected codons in the region 137-149 (B type numbering). These associate with portions of the H3 antigenic region A. Three and five, respectively, of the B and A positively selected codon set positions occur in the space of the 14 positions of region A which is a major focus for immunological response. Moreover, the region also contains four of the potential glycosylation sites that varied between Type A and B.

Air et al (Air et al. 1990) believe that if there is any selection of influenza B or avian influenza A viruses at the protein level, it is not by antibodies. It is difficult to reconcile our results for the B type with their view. B is quite like A except slower. Its substitution and replacements rates of B type HA are a half to a third of that of the A type. Similarly the average age of an isolate is about three times that of a Type A isolate. It most certainly has positive selection occurring, concentrated in a region homologous to the antibody combining sites of the A type.

In summary, although the B type does not undergo reassortments, its hemagglutinin evolution resembles that of the A/H3 type. The B type differs from the A/H3 type by having a slower accumulation of beneficial replacements and a greater longevity of lineages.

ACKNOWLEDGMENTS

We thank Jing Huang for gathering an initial set of these sequences for us. This work was supported by NIH grant AI44474. and Nancy Cox and the members of the influenza branch of the CDC.

REFERENCES

- Air, G. M., A. J. Gibbs, W. G. Laver, and R. G. Webster. 1990. Evolutionary changes in influenza B are not primarily governed by antibody selection. *PNAS* **87**:3884-3888.
- Berton, M. T., and R. G. Webster. 1985. The antigenic structure of the influenza B virus hemagglutinin: Operational and topological mapping with monoclonal antibodies. *virology* **143**:583-394.
- Bush, R. M., C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. 1999a. Predicting the Evolution of Human Influenza A. *Science* **286**:1921-1925.
- Bush, R. M., W. M. Fitch, C. A. Bender, K. Subbarao, and N. J. Cox. 1999b. Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A. *MBE* **16**:1457-1465.
- Bush, R. M., C. A. Smith, K. Subbarao, W. M. Fitch, and N. J. Cox. 2000. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *PNAS* **97**:6974-6980.
- Cox, N. J., C. A. Bender, A. P. Kendal, H. L. Regnery, M. L. Hemphill, and P. A. Rota. 1993. Evolution of hemagglutinin in epidemic variants and selection of vaccine viruses. Pp. 223-229 *in* C. Hannoun, ed. *Options for the control of influenza*. Elsevier.
- Fitch, W. M. 1972. Evolutionary variability in hemoglobins. *Haematologie und Bluttransfusion* **10**:199-215.
- 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. *PNAS* **94**:7712-7718.
- Krystal, M., R. M. Elliott, E. W. J. Benz, J. F. Young, and P. Palese. 1982. Evolution of influenza A and B viruses: Conservation of structural features in the hemagglutinin genes. *J. Gen. Virology* **79**:4800-4804.
- Krystal, M., J. F. Young, P. Palese, I. A. Wilson, J. J. Skehel, and D. C. Wiley. 1983. Sequential mutations in hemagglutinins of influenza B virus isolates: Definition of antigenic domains. *PNAS* **80**:4527-4531.
- Nerome, R., Y. Hiromoto, S. Sugita, N. Tanabe, M. Ishida, M. Matsumoto, S. E. Lindstrom, T. Takahashi, and K. Nerome. 1998. Evolutionary characteristics of influenza B virus since its first isolation in 1940: dynamic circulation of deletion and insertion mechanism. *Arch. Virology* **143**:1569-1583.
- Rota, P. A., M. L. Hemphill, T. Whistler, H. L. Regnery, and A. P. Kendal. 1993. Antigenic and genetic characterization of the haemagglutinins of recent cocirculating strains of influenza B virus. *J. Gen. Virology* **73**:2737-2742.
- Swofford, D. L. 1993. PAUP (phylogenetic analysis using parsimony). Illinois Natural History Survey, Champaign, Illinois.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of multiple sequence alignment through sequence weighting, positions-specific pair penalties and weight matrix choice. *Nucleic Acid Reserach* **22**:4673-4680.
- Zou, S., I. Prud'homme, and Weber. 1997. Evolution of the hemagglutinin gene of influenza B virus was driven by both positive and negative selection pressures. *Virus Genes* **14**:181-185.

TABLES

Table 1. Do the substitutions distribute themselves randomly between tip and off-tip branches?

	Tips	Off-tips
Exp	709.9	708.1
Obs	939	479
X^2	73.9	74.1

Exp is expected, Obs is observed, X^2 is chi-squared for one degree of freedom, p value $\ll 10^{-18}$. Expected assumes that every branch has the same probability of receiving any substitution. This null hypothesis is rejected.

Table 2. Do the silent and non-silent substitutions distribute themselves randomly between tip and off-tip branches?

		Silent	Non-silent
Tip	Exp	517.9	421.1
	Obs	533	436
	X^2	0.43	0.52
Off-tip	Exp	264.2	214.8
	Obs	279	200
	X^2	0.83	1.02

Exp equals expected; Obs equals observed; X^2 equals Chi-squared.. The null hypothesis is that the distribution of changes onto tip and off-tip branches is independent of whether the change is silent or non-silent. The value of $X^2 = 2.80$ for two degrees of freedom gives a p-value = 0.25 and the null hypothesis is not rejected.

Table 3. Observed and expected number of the different regional migrations.

	A	C	E	H	N	S	Y	Σ	#	$\Sigma/\#$
A	0 0	0.60 1	0.63 2	0.25 0	0.54 2	0.29 0	0.59 2	2.92 7	13	0.22 0.53
C	2.81 1	0 0	13.26 1	4.85 15	16.66 4	4.74 1	11.86 15	54.30 37	99	0.54 0.37
E	.71 0	4.05 2	0 0	1.82 0	5.32 9	1.73 3	4.13 3	17.76 17	60	0.30 0.28
H	0.05 1	0.29 1	0.19 1	0 0	0.88 4	0.39 0	0.82 4	2.61 11	21	0.12 0.52
N	3.08 6	17.88 8	10.94 9	3.95 2	0 0	4.90 7	12.47 14	53.20 46	103	0.52 0.44
S	0.06 1	0.33 2	0.20 4	0.08 1	0.29 5	0 0	0.83 5	1.82 18	23	0.08 0.78
Y	0.91 3	4.86 3	3.07 3	1.13 0	4.14 3	1.15 0	0 0	15.26 12	67	0.23 0.17
Σ	7.60 12	28.03 17	28.34 20	12.10 18	27.84 27	13.19 11	30.70 43	148.0 148		
#	13	99	60	21	103	23	67			
$\Sigma/\#$	0.58 0.92	0.28 0.17	0.47 0.33	0.58 0.85	0.27 0.26	0.57 0.47	0.46 0.64			

The body of the table contains the number of times it is expected that an isolate will migrate from one of seven regions to another (upper number of a pair) and how many times it was so observed (lower number). Expected values are the average of 1000 Monte Carlo trials in which each trial was a different scrambling of the regional labels of the isolates. A = Australia; C = China; E = Europe; H = Hong Kong; N = North America; S = South America; Y = Asia (less C and H). Migration is **from** the region/row on the left **to** the region/column at the top. # = the number of isolates from that region. This does not change from trial to trial. Σ equals the sum of the row values to the left or the column values above. The mean number of expected migrations was normalized to the observed 148 migrations.

LEGENDS TO FIGURES

Figure 1. Alignment of type A and B consensus protein sequences and the positively selected codons.

The positions that show the same amino acid for each of the two sequences have that amino acid repeated in the space between the two amino acid sequences. If the pair of aligned amino acids is very similar but not identical (such as AG, DE, FY, IV, KR, ST), they have a plus sign (+) between the two amino acids. In addition, those amino acid positions that were designated as under positive selection in type A/H3 are shown in bold face with an asterisk beneath the amino acid (Bush et al. 1999b). Similarly, those designated in this report as under positive selection in the B type are shown in bold face with an asterisk above the amino acid. The o symbol indicates an amino acid position that is part of a sialic acid binding site. An = sign denotes one of the possible selective positions that was removed because the NS/S ratio was 3/0 or 4/1.

In addition, the two residues, 163 and 164, that are sometimes missing in type B sequences, are shown italicized and underlined. Finally, an open diamond at a residue indicates that, in at least one sequence, that position is the beginning of a three amino acid potential glycosylation site, N-X-(S/T). If the diamond symbol is solid, that glycosylation site was found in all 386 B sequences.

Figure 2. Alternative indel evolution. Codon positions 163 and 164 are sometimes not present in B type hemagglutinins. The codon pair is here represented as AB when the two codons are present, and either codon can be represented by a Δ when it is absent. The tree on the left is a most parsimonious solution for the data and the tree as given. It has five indel events. The tree in the middle shows an alternative that brings an indel event in both codons onto the same branch, the one descending to B/Canada/3/85. Because the two indel events could happen together in a single event, then the data can be explained by four indel events. The tree on the right shows how, by moving B/Osaka/70 one branch closer to the Yamagata lineage, we obtain a tree that requires only three indels. It comes at a cost of two extra nucleotide substitutions. Other B isolates are : Victoria, VC70; Aichi/7, AI76; Beijing/19, BJ93; Baylor/4, BL78.

Figure 3. Phylogeny of the B type human influenza viral hemagglutinin

Figure shows the most parsimonious tree found for 386 type B hemagglutinin sequences. The scale bar shows the horizontal scale in nucleotide substitutions. Horizontal distances are in nucleotide substitutions. If one treats the Victoria lineage separately from the Yamagata lineage, one may discern a trunk which is shown as a thicker line than that used for the rest of the tree. Also given is an arbitrary possible trunk shown by the dotted lines on the tree.

Figure 4. Rate of evolution of Human B type hemagglutinins. The year of isolation is plotted against the replacement distance from the root of the tree in figure 2. Each digit represents the number of isolates that map to that point. For points containing more than 9 isolates, A=10, B=11, C=12, D=16 and E=25. Points prior to 1978 were omitted in drawing the line.

Alignment of type A and B consensus sequences of Hemagglutinins

```

      ◆                               ◆                               ◆
B:   3 ICTGITSSNSPHVVKATQGEVNVTVGVIPLTTTPTKSHFANLKGTKTRGKLCPNCLNCTD 62
      +C G          +VKT T   + VT   L   + +          GK          NCT
A:  13 LCLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTGRICDSPHRILDGK-----NCTL 66
      ◆                               ◆           ◆                               ◆

      * *=      =
B:  63 LDVALGRPMCVGTTPSAKASILHEVRPVTSGCFPIIMHDRTKIRQLPNLLRGYENIRLSTQ 122
      +D LG P C G          + E          S C+P D          L L+          T
A:  67 IDALLGDPHCDG-FQNKEWDLFVERSKAYSNCYP--YDVPDYASLRSLVAS-----SGTL 118
      O

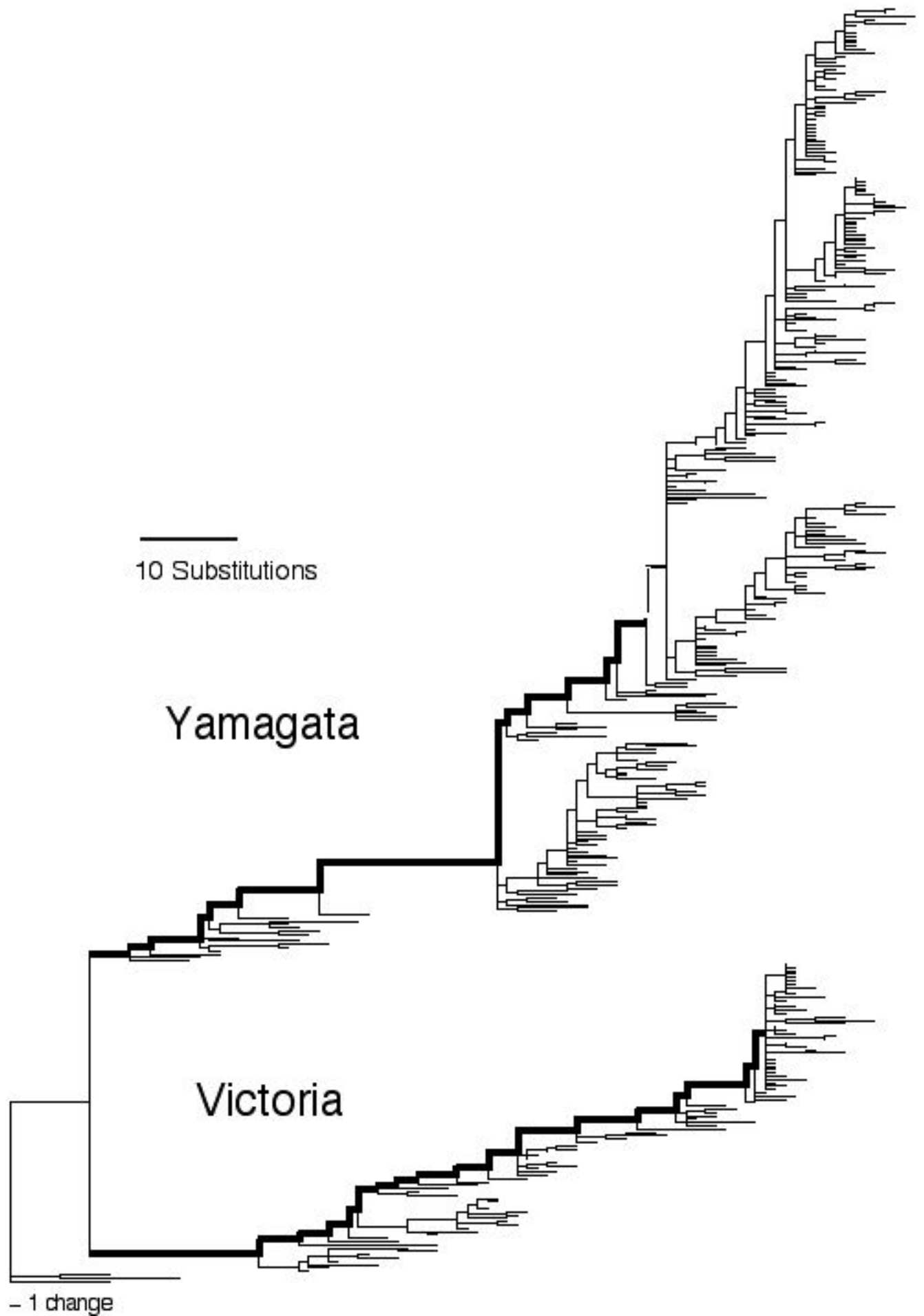
      =          *          ◆ ◆**          ◆ ◆ ◆          =
B: 123 NVINAEKAPGGPYRLLGTSGSCPNATSRSGFFATMAWAVPRNDNNKTATNPLTVEVPYICT 182
      IN          G          GTS C ++ +S FF + W + + +          A N +T+ P
A: 119 EFINEGFNWTGVAQDGTSYACKRGSVKS-FFSRLNW-LHKLEYKYPALN-VTM--PNN-- 171
      * ◆ * ◆          * ◆ * O O *          * ◆ *          O O * *          ◆          ◆
      O O O

      ◆          *          =          * ◆
B: 183 KGEDQITVWGFH-SDNKTQMKNLYGDSNPQKFTSSANGVTTHYVSQIGGFPDQTEDGGLP 241
      D + +WG H          + LY          S+          T + IG P          GL
A: 172 DKFDKLYIWGVHHPSTDSDQTSLYVQASGRVTVSTKRSQQT-VIPNIGSRPWVR---GL- 226
      O * * * * O * *          O * *          O * *          O *
      O O O          O

      *
B: 242 QSGRIVVDYMVQKPG-----KTGTIVYQRGILLPQKVWCASGRSKVIKGSLPLIG-EAD 294
      S RI + + + KPG          TG ++ RG          SG+S +++ P+          +
A: 227 -SSRISYWTIVKPGDILLINSTGNLIAPRGYFK-----IRSGKSSIMRSDAPIGCNSE 280
      OO          ◆          *          * ◆

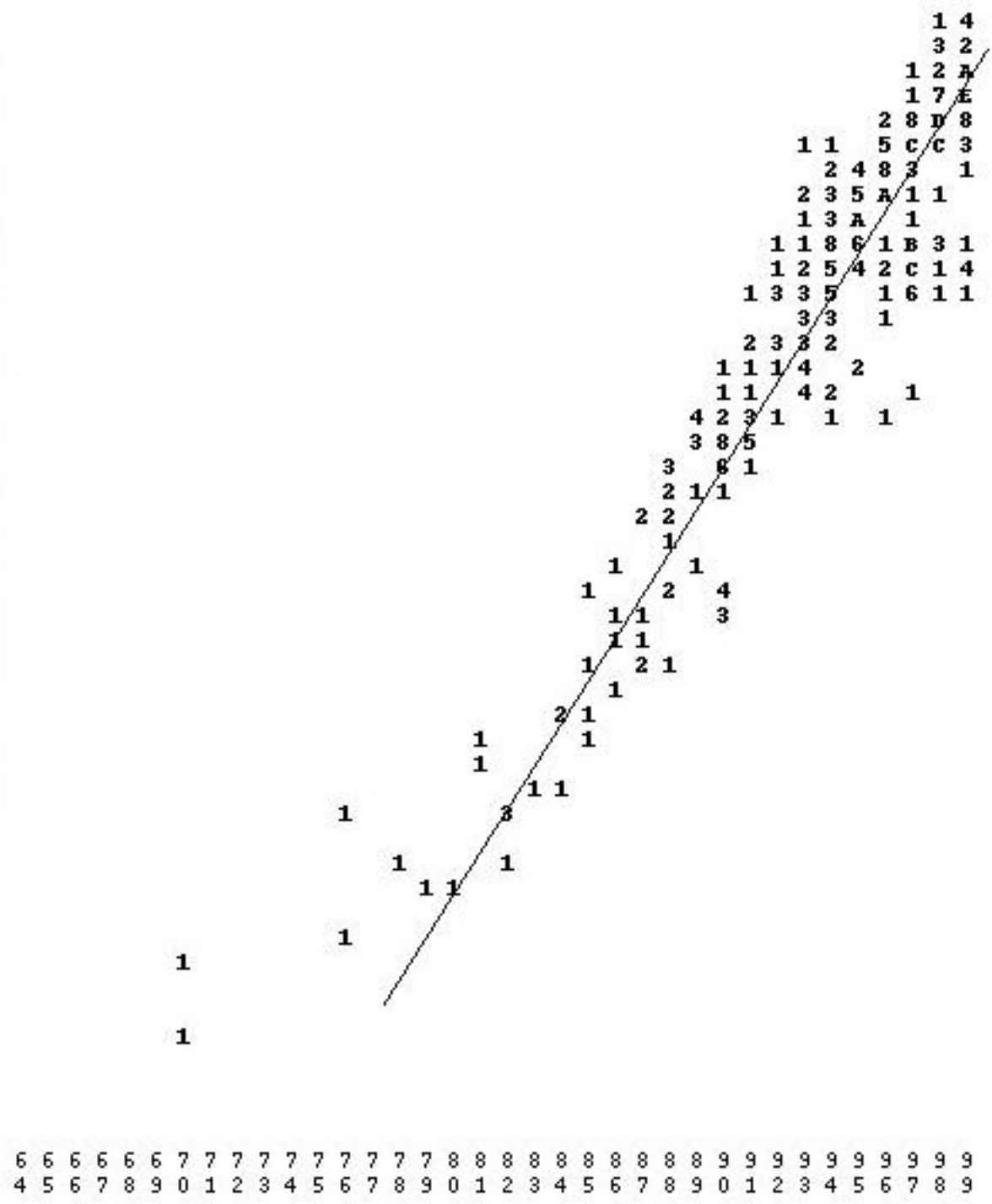
      ◆                               ◆
B: 295 CLHEKYGGLNKSKPYYTGEHAKAIGNCPIWVK-TPLKLANGTKYRPPAKLLKER 347
      C+          G + KP+          G CP +VK          LKLA G + P          K
A: 281 CITPN-GSIPNDKPFQNVNRI-TYGACPRYVKQNTLKLATGMRNVPEKQTRK-- 330
      ◆

```

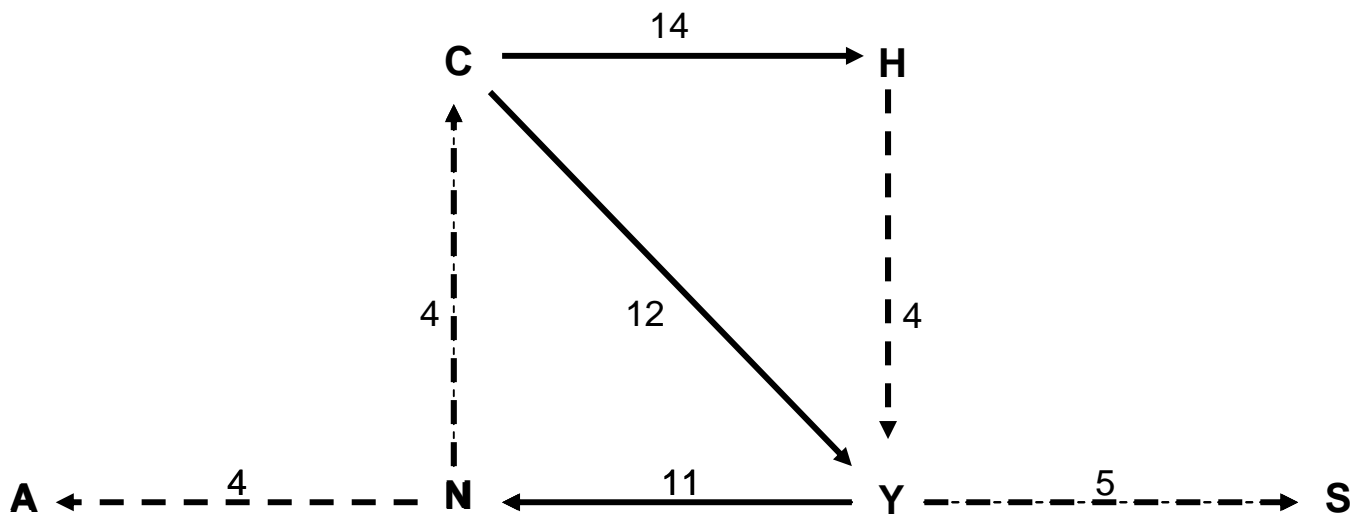



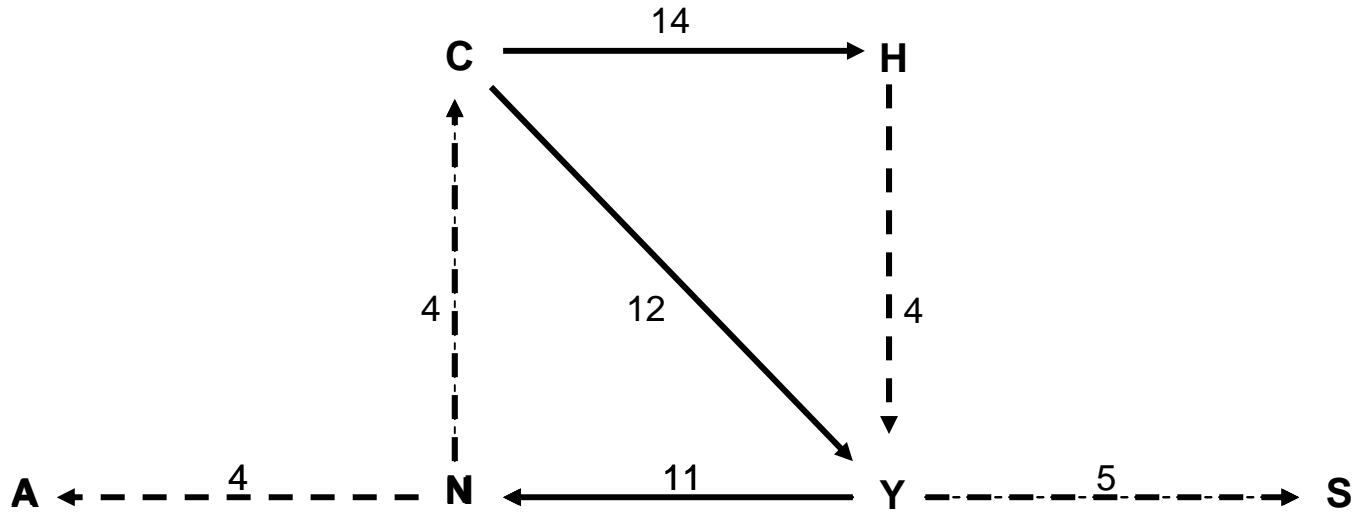
R
E
P
L
A
C
E
M
E
N
T
S

44
43
42
41
40
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0



Y E A R





020305