INTRODUCTION

The influenza A virus infects a variety of animals including birds, swine, horses, and humans. Two surface proteins of the virus, hemagglutinin and neuraminidase, are important in determining host specificity because they are responsible for binding to and release from host cell receptors. In mammals, especially humans, these proteins are subject to considerable immune selection. The high mutation rate of the viral RNA genome leads to amino acid replacements that make possible both adaptation to different hosts and escape from previously established immunity. Antigenic and functional sites can be identified by analyzing mutants that “escape” a panel of antibodies or that demonstrate different functional characteristics (Aytay and Schulze, 1991; Gubareva et al., 1996; Kilbourne et al., 1988; Luoh et al., 1992; Webster et al., 1982). An alternative method is to determine which amino acid positions in the protein are phylogenetically informative (Fanning and Taubenberger, 1999). These positions, which show lineage-specific changes, are likely to be involved in virus–host interactions and may include, in addition to antigenic sites, receptor-binding sites and sites involved in protein conformation, host–protein interaction, and/or posttranslational modification.

Neuraminidase cleaves sialic acid linkages, including those between hemagglutinin and cell surface receptors, thus releasing newly formed virus from the cell surface (Palese and Comans, 1976). If a virus’s hemagglutinin binding and neuraminidase cleavage activities are not balanced, productive rounds of infection cannot be maintained (Kaverin et al., 1998; Rudneva et al., 1996). Similarly, if neuraminidase is neutralized by antibody, influenza infections are mitigated (Schulman et al., 1965; Webster et al., 1988). Therefore, after initial infection by a new influenza variant, and the establishment of immunity in a majority of the host population, viruses with neuraminidases containing alterations in specific regions of the protein are selected since they allow the virus to evade the host’s immune system.

There are nine different influenza A neuraminidase subtypes circulating in birds, some of which also circulate in mammals (Murphy and Webster, 1996). Two subtypes, N1 and N2, have been found to circulate in human populations. The N1 subtype predominated in humans from at least 1918 until 1957, when it was replaced by the N2 subtype. In 1977 the N1 subtype reemerged and the two subtypes have cocirculated in the human population ever since (Webster et al., 1992). N1 and N2 are quite dissimilar, with only 42% identity at the amino acid level. Nevertheless, each successfully adapted to long-term circulation in humans. Homologous regions shown to have phylogenetic importance in both subtypes, therefore, are very likely to be important in host adaptation.

There are several methods used to identify phylogenetically important amino acid positions in a protein. One method employs a comparative approach in which parsimony trees are compared with random trees (Maddison and Maddison, 1992). Parsimony attempts to minimize the total number of steps (e.g., amino acid substitutions) required to construct a phylogenetic tree. With influenza viral proteins, some amino acid positions (e.g., invariant positions) will not be phylogenetically inform
tive, while other positions can be very informative (e.g.,
those in which all members of one group have a particular
amino acid, while all members of another group
have a different amino acid). Phylogenetically informative
amino acid positions can be identified by comparing the
shortest parsimony tree with a collection of much longer
random trees. At each amino acid position there will be
a certain number of amino acid changes as one moves
through the tree from the deepest ancestral node to the
terminal branches. Random trees, in which the se-
quencies are arranged randomly on the terminal
branches, inevitably have many changes. The parsimony
tree must have fewer changes at some positions be-
cause it is the tree that, by definition, has the least
number of changes.

RESULTS

Phylogenetic analyses of N1 and N2 neuraminidases

Because of our interest in influenza–human interac-
tions, we have restricted our analyses to the N1 and N2
neuraminidase subtypes. Parsimony analysis of 32 N1
neuraminidase protein sequences produced four trees of
621 steps each. Three major clades—human, avian, and
swine—were differentiated, with swine being split be-
tween new isolates and an isolate from 1930 (Fig. 1A). A
parsimony analysis of 25 N2 neuraminidase sequences
produced a single tree of 573 steps with two prominent
clades, mammalian and avian (Fig. 1B).

Phylogenetically informative amino acid positions in
N1 and N2 neuraminidases

We compared the four most parsimonious N1 PAUP
trees with a set of 500 randomly generated N1 trees to
determine which amino acid positions in N1 were phy-
genetically informative, i.e., made the parsimony trees
shorter (Maddison and Maddison, 1992). Fifteen phylo-
genetically important regions (PIRs) of N1, defined as
several phylogenetically important amino acid positions
(PIPs) near one another, were observed. The 15 regions
are labeled A–O in Fig. 2A. Table 1 lists the regions,
together with their possible roles in NA function. Five
regions, I, J, and L–N, are homologous to sites shown to
be antigenic on N2 (Colman et al., 1983; Air et al., 1985,
1989). Region A is part of the signal-anchor domain of the
protein (Brown et al., 1988). Region B contains a glyco-
sylation site in human isolates, while region C, which
overlaps with the NA stalk, contains both glycosylation
sites and deletions in the N1 genes of many influenza
strains (Blok and Air, 1980; Luo et al., 1993). To our
knowledge, the other seven regions (D–H, K, and O) have
not been associated with any function of the N1 protein.

In a manner similar to that for N1, we compared
the single most parsimonious N2 tree with 100 randomly
generated N2 trees. Twelve PIRs were identified (Fig. 2B,
Table 2. Regions D', F', G', I', and K' are antigenic sites (Colman et al., 1983; Air et al., 1985, 1989). Regions C' and K' contain glycosylation sites in nearly all N2 neuraminidases sequenced to date. Regions A', B', E', H', J', and L', to our knowledge, have not been associated with any function of the N2 protein.

FIG. 2. PIPs in N1 and N2 influenza A neuraminidases. The histograms show positions in the protein at which parsimony trees are shorter than random trees. Scale on the left is the number of steps shorter. Regions discussed in the text are indicated over the histograms.
To visualize these regions in greater detail we have aligned two neuraminidase proteins to highlight the relationships of their PIRs (Fig. 3). Figure 3 shows the position of PIPs in the PIRs, the positions of the PIPs in the amino acid chain, and the relation of N1 neuraminidase PIPs to those of the N2 subtype. Actual amino acid sequences have not been used since these change from clade to clade. When analyzed by the RasMol program (Sayle, 1994), using the N9 neuraminidase crystal structure (PDB identifier: 7NN9), all of the PIRs beyond D/B are either partially or completely on the surface of the NA protein.

**DISCUSSION**

**Regions identifying antigenic sites**

We have identified phylogenetically informative amino acid positions (PIPs) in the N1 and N2 subtype influenza neuraminidases. We define two or more closely spaced positions as a phylogenetically important region (PIR). Because of the nature of the virus–host interaction, it seems reasonable to suppose that the data would support several predictions. First, many PIRs shared by N1 and N2 (coincident regions) might be expected to identify antigenic sites, since N1 and N2 have both been subject to immune pressure over several decades. This appears to be the case. There are seven PIRs on N1 that have corresponding PIRs on N2. These N1–N2 coincident pairs are: B’–A’, C’–B’, I’–F’, J’–G’, K’–H’, L’–I’, and M’–K’ (see Fig. 3). Of these, four pairs (I’–F’, J’–G’, L’–I’, and M’–K’) have N2 regions that are antigenic sites. In fact, of the 16 (73%) are in PIRs identified here; other antigenic amino acid residues are.

**TABLE 1**

Phylogenetically Important Regions (PIRs) in N1 Subtype Neuraminidases

<table>
<thead>
<tr>
<th>N1 region</th>
<th>N1 amino acids</th>
<th>Properties (references)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15–17</td>
<td>Signal-anchor region (Brown et al., 1988)</td>
</tr>
<tr>
<td>B</td>
<td>41–48</td>
<td>Glycosylation in humans</td>
</tr>
<tr>
<td>C</td>
<td>67–86</td>
<td>NA stalk, glycosylation/deletions (Blok and Air, 1982; Luo et al., 1993)</td>
</tr>
<tr>
<td>D</td>
<td>188–189</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>221–222</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>248–250</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>263–264</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>285–289</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>328–332</td>
<td>Antigenic site in N2 (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>J</td>
<td>339–344</td>
<td>Antigenic site in N2 (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>K</td>
<td>351–352</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>365–369</td>
<td>Antigenic site in N2 (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>M</td>
<td>430–434</td>
<td>Antigenic site in N2 (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>N</td>
<td>454–455</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2**

Phylogenetically Important Regions (PIRs) in N2 Subtype Neuraminidases

<table>
<thead>
<tr>
<th>N2 region</th>
<th>N2 amino acids</th>
<th>Properties (references)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A’</td>
<td>40–46</td>
<td>Stalk (Blok and Air, 1980)</td>
</tr>
<tr>
<td>B’</td>
<td>81–82</td>
<td>Glycosylation site</td>
</tr>
<tr>
<td>C’</td>
<td>141–149</td>
<td>Antigenic site (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>D’</td>
<td>197–199</td>
<td></td>
</tr>
<tr>
<td>E’</td>
<td>302–308</td>
<td>Antigenic site (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>F’</td>
<td>328–339</td>
<td>Antigenic site (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>G’</td>
<td>344–347</td>
<td>Antigenic site (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>H’</td>
<td>356–358</td>
<td></td>
</tr>
<tr>
<td>I’</td>
<td>368–369</td>
<td>Antigenic site (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>J’</td>
<td>384–386</td>
<td></td>
</tr>
<tr>
<td>K’</td>
<td>400–403</td>
<td>Antigenic site (Colman et al., 1983; Air et al., 1985, 1989)</td>
</tr>
<tr>
<td>L’</td>
<td>463–468</td>
<td></td>
</tr>
</tbody>
</table>

**FIG. 3.** Locations of PIPs within PIRs of N1 and N2 neuraminidases. A representative example of each neuraminidase subtype was chosen and the two were aligned. The regions corresponding to A–O (N1) and A’–L’ (N2) are shown. In each case a [0] represents a phylogenetically informative position, whereas a [+] is not informative. PIRs are boxed and in bold. Numbers refer to the initial amino acid positions of the regions.

**DISCUSSION**

Regions identifying antigenic sites

We have identified phylogenetically informative amino acid positions (PIPs) in the N1 and N2 subtype influenza neuraminidases. We define two or more closely spaced positions as a phylogenetically important region (PIR). Because of the nature of the virus–host interaction, it seems reasonable to suppose that the data would support several predictions. First, many PIRs shared by N1 and N2 (coincident regions) might be expected to identify antigenic sites, since N1 and N2 have both been subject to immune pressure over several decades. This appears to be the case. There are seven PIRs on N1 that have corresponding PIRs on N2. These N1–N2 coincident pairs are: B’–A’, C’–B’, I’–F’, J’–G’, K’–H’, L’–I’, and M’–K’ (see Fig. 3). Of these, four pairs (I’–F’, J’–G’, L’–I’, and M’–K’) have N2 regions that are antigenic sites. In fact, of the 22 antigenic positions identified by Colman et al. (1983) and Air et al. (1985, 1989), 16 (73%) are in PIRs identified here; other antigenic amino acid residues are.
not in PIRs, but are found as isolated PIPs (see below). Second, many PIPs in antigenic sites might be expected to identify the human clade. This is because the human immune system should exert considerable pressure on the virus, resulting in selection of new viral variants that can evade immune surveillance. This expectation appears to be borne out from a total of 23 PIPs in predicted antigenic sites of N1, 15 (65%) identify the human clade. From a total of 18 PIPs in predicted antigenic sites of N2, 14 (78%) identify the human clade.

Regions of undetermined importance

There are a number of PIPs that are unique to a particular neuraminidase subtype, or that are coincident on N1 and N2 but have not been identified as antigenic sites. Coincident regions B–A’ are in a section of the neuraminidase protein buried in the viral membrane, as is the N1-specific A region. How these regions might be involved in virus–host interactions is unknown. Coincident regions C–B’ are in the stalk region of the neuraminidase protein, which is quite variable in N1 proteins, both in amino acid sequence and because several N1 subtype strains have small overlapping, nonidentical deletions in the stalk (Blobk and Air, 1980, 1982). In all cases, however, the stalk regions retain a cysteine residue and at least one glycosylation site. Members of a third coincident pair, K–H’, are situated between, and are possibly part of, the predicted antigenic sites J and L on N1, and G’ and I’ on N2. N1-specific regions D, E, G, and O, and N2-specific region D’ consist of only two amino acid positions and may be spurious or represent culture artifacts. On the other hand, enzymatic active site residues 225 and 228 in N1 are near region E, and the active site residue 198 in N2 is between the two positions making up region D’. Thus, changes in E and D’ may affect protein conformation around some active site residues and possibly alter enzyme activity or specificity, which might be important for growth in certain hosts. PIRs H in N1 and C’, E’, J’, and L’ in N2 consist of more than two amino acid positions. They may be antigenic sites unidentified by other methods, or they may identify regions of the proteins involved in some other aspect of virus–host interaction.

Positions that may be involved in mammalian adaptation

Thirteen PIPs in N1 are potentially interesting because they distinguish N1 avian isolates from N1 swine and human isolates. Five of these PIPs are of particular interest since they place both SwEngland92 and Brevig Mission18 outside the avian clade (Reid et al., 2000). SwEngland92 is the descendant of an avian virus introduced into European swine in the late 1970s (Brown et al., 1997). Although present in swine for about 15 years, the virus is still clearly avian-like (Fig. 1A). Brevig Mission18 is the virus responsible for the 1918 “Spanish” influenza pandemic that killed over 20 million people worldwide and probably entered the human population from an avian source shortly before 1918 (Reid et al., 1999; Reid and Taubenberger, 1999). Three of the potentially interesting PIPs (79, 81, and 84) are in a highly variable PIR of the N1 protein, region C, which overlaps with the NA stalk. Although sequence variation in the stalk, including small deletions, appears to have little effect on function (Luo et al., 1993), the presence of PIRs in the stalk regions of both N1 and N2 suggests that some features of this area may nevertheless play a role in host adaptation. Another PIP, 285, is part of a six-amino-acid region lying between two blocks of sequence that are highly conserved in all NA proteins. These two conserved blocks of 9 and 13 amino acids contain 4 of the 18 enzymatic active site residues. Residue 285 probably lies near the surface of the NA protein as seen by examining the crystal structure of the N9 protein using the RasMol program (Sayle, 1994). Changes in this region could conceivably lead to alterations in the conformation of the active site that, in turn, might modify enzymatic activity or specificity. The fifth interesting PIP is 344, which lies within the N1–N2 coincident PIR pair J–G’. No function has yet been assigned to either of the J or G’ PIRs. Because both SwEngland92 and Brevig Mission18 are nonavian at these five positions it is quite possible that changes at one or more of them is involved in the ability of avian viruses to circulate in mammals.

Conclusions

By comparing the shortest, most parsimonious phylogenetic tree(s) with a large set of random trees we have been able to identify a number of phylogenetically “information-rich” regions (PIRs) of the influenza A neuraminidase protein. As in the previous study, which looked at the influenza A hemagglutinin protein (Fanning and Taubenberger, 1999), many of the PIRs are coincident with known antigenic and/or glycosylation sites on the neuraminidase protein. Other PIRs, however, have not been previously identified by traditional methods and a knowledge of their existence and locations may prove useful in interpreting future research.

A number of functional amino acid positions would be missed using the parsimony tree(s) versus random trees approach described here. First, all invariant positions will be completely ignored by the method. Thus, neuraminidase amino acids involved in enzymatic activity were not identified since they are conserved in all subtypes (Colman et al., 1983). Second, since a PIR is required to have several PIPs near one another (Fig. 3), isolated PIPs will not be counted. For example, position 431 of N2 was identified as an antigenic residue (Colman et al., 1983; Air et al., 1985, 1989), but was not within an N2 PIR. Position 431 is indeed a PIP in N2, but it is
surrounded by a number of invariant positions. In contrast, the corresponding region of N1 is quite variable (Fig. 3).

Our analyses suggest that in addition to known antigenic and glycosylation sites on N1 and N2 subtype neuraminidases, some PIRs may represent other functionally important sites. It seems likely that some of the short PIRs will turn out to be spurious. However, longer PIRs and PIRs shared between neuraminidase subtypes probably identify regions of the proteins genuinely involved in virus–host interactions. While most of these regions are likely to represent antigenic sites, there are other possibilities, e.g., regions involved in protein conformation, changes in which could alter the enzymatic properties of the protein.

MATERIALS AND METHODS

Influenza strains

Sequences were obtained from GenBank, EMBL, and DDBJ. Strain abbreviations are as found in Fanning and Taubenberger (1999), Reid et al. (1999, 2000), Taubenberger et al. (1997), or listed below.

N2 subtype. Lenin57, A/Leningrad/134/57 (H2N2); R157, A/R1/5–57 (H2N2); Niigata96, A/Niigata/137/96 (H3N2); Swiowa98, A/Swine/owa/8548–1/98 (H3N2); Johannesburg94, A/Johannesburg/33/94 (H3N2); Guangdong93, A/Guangdong/25/93 (H3N2); Shandong93, A/Shandong/9/93 (H3N2); Vermont94, A/Vermont/3/94 (H3N2); Beijing89, A/Beijing/353/89 (H3N2); England88, A/England/427/88 (H3N2); Mississippi86, A/Mississippi/1/86 (H3N2); SwKanagawa78, A/sw/Kanagawa/2/78 (H1N2); HK71, A/Hong Kong/107/71 (H3N2); HK68, A/Hong Kong/8/68 (H3N2); SwNagasaki90, A/swine/Nagasaki/1/90 (H1N2); DkHK97, A/Duck/Hong Kong/Y439/97 (H9N2); ChHK94, A/Chicken/Hong Kong/739/94 (H9N2); ChHK97, A/Chicken/Hong Kong/G9/97 (H9N2); QualiHK92, A/Quali/Hong Kong/AF157/92 (H9N2); ChHidalgo94, A/Chicken/Hidalgo/26654–1368/94 (H5N2); ChPuebla94, A/Chicken/Puebla/8623–607/94 (H5N2); TyCalifornia66, A/Turkey/California/189/66 (H9N2).

N7 subtype. N7Weybridge, A/FPV/Weybridge (H7N7).

N8 subtype. EqKen81, A/Ken/1/81 (H3N8).

N1 sequences

Thirty-two N1 neuraminidase amino acid sequences plus one N8 sequence (as outgroup) were analyzed by PAUP as above. A single tree of 573 steps was obtained (consistency index = 0.82).

Phylogenetically important positions in N1 and N2

MacClade (Maddison and Maddison, 1992) is a program that allows the evolution of characters to be traced throughout the phylogenetic tree. Viral sequences can be moved around in the parsimony tree to evaluate the effect of different tree topologies on branch length, tree length, and character transformation. MacClade was used to identify phylogenetically informative amino acid positions in the N1 and N2 proteins. For N1, one of the four most parsimonious trees determined by PAUP was imported into the MacClade program and the data were used to generate 500 random trees (range: 1171–1428 steps) using the random joining/splitting option. The four most parsimonious PAUP trees were then compared with the random N1 trees at each amino acid position to determine how important the position was in shortening the PAUP trees (Fanning and Taubenberger, 1999). For N2, the single most parsimonious tree was compared with 100 random N2 trees (range: 880–1041 steps). (A complete listing of N1 and N2 PIPs is available by contacting the corresponding author at his e-mail address.)

Sequence alignment

A diagrammatic representation of the phylogenetically important regions of N1 and N2 was constructed using two representative amino acid sequences, NJ76b (N1) and Lenin57 (N2). The sequences were aligned using Lasergene software (DNAsStar, Madison, WI) with a gap penalty of 20. The resulting alignment showed 41.8% sequence identity with only 10 short gaps, six in N1 and four in N2.

ACKNOWLEDGMENTS

This study was supported by grants from the American Registry of Pathology, the Department of Veterans Affairs, and intramural funds of the Armed Forces Institute of Pathology. The opinions contained herein are our own private views and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

REFERENCES

Blok, J., and Air, G. M. (1982). Block deletions in the neuraminidase
genes from some influenza A viruses of the N1 subtype. Virology 119, 229–234.