Computational analysis of adaptive antigenic mutations of the human influenza hemagglutinin for vaccine strain selection

Jared Rickert  
Department of Biology  
University of Nebraska-Omaha  
Omaha, NE 68182  
jdrickert@unomaha.edu

Todd Herpy  
Department of Chemistry  
University of Nebraska-Omaha  
Omaha, NE 68182  
thery@unomaha.edu

Haizhen Zhong  
Department of Chemistry  
University of Nebraska-Omaha  
Omaha, NE 68182  
hzhong@unomaha.edu

Guoqing Lu  
Department of Biology  
University of Nebraska-Omaha  
Omaha, NE 68182  
glu3@unomaha.edu

Abstract—The human influenza H1N1 and H3N2 viruses have been circulating over previous decades and are likely to continue due to antigenic mutations, with the hemagglutinin (HA) segment being the main target. However, adaptive evolution of the HA segment has not been fully understood because various selection pressures are involved. This paper aims to identify significant antigenic selection sites and link these sites to their respective amino acids to the structure changes over time. Strong positive selection was detected in the HA of both H1N1 and H3N2, with four positively selected sites for H1N1 and eight positively selected sites for H3N2. Most of the positively selected codons showed patterns of directional evolution; and certain lineages that arose from positively selected codons appear to have finite lifespans. Two lineages have existed in the past decade in H1N1, with positive selection pressure occurring most recently at position 206 for the 1A lineage (2009 swine influenza) and positions 144, 190, and 192 for the 1B lineage. For H3N2, the loop regions near 157 and 225 appear to be the latest regions under selection pressure that vaccines may be able to target. By monitoring and linking the sequence changes with the tertiary structural changes, researchers will be in better position in terms of picking appropriate vaccine strains and designing efficient vaccines and anti-viral drugs to combat this constantly mutating virus.

Index Terms—hemagglutinin, influenza, mutation, selection

I. INTRODUCTION

The H1N1 and H3N2 influenza viral strains that have been circulating within the past few decades have been mutating annually and these are primarily responsible for the seasonal flu epidemics. The mutation process responsible for these epidemics is known as antigenic drift, in which the virus inserts point mutations in order to evade the host immune response. In contrast, influenza pandemics occur due to antigenic shift, resulting from reassorting gene segments within the virus to which the host immune system has had no previous exposure. Typically, these strains that cause pandemics will undergo antigenic drift in years after, causing the less severe epidemics or seasonal flu. It should be interesting to trace mutations that may occur in H1N1 following the 2009 Swine Flu pandemic season.

A virus undergoes a sub-process of antigenic drift known as positive selection in order to select for advantageous amino acids to avoid the host immune response. The process of positive selection has had a great deal of influence on the host species. The host immune system is able to build up subtype and strain specific antibodies to the surface glycoproteins of the HA [6]. The strongest antibody response occurs when the individual is infected with their first variant [6]. However, the RNA polymerase complex makes errors during replication of viral HA genes—one estimate is 10\(^{5.5}\) mutations/site/replication[57]. Due to the fact that RNA viruses do not have a proofreading mechanism, mutations will alter the shape and the electric charge of viral surface antigens of hemagglutinin. When a virus selects for a different amino acid at a certain codon position, this is known as a non-synonymous (dN) mutation, whereas a mutation selecting for the same amino acid is known as a synonymous (dS) mutation. A codon position that has a high dN:dS ratio is thought to be under positive selection. Much variation occurs in antibody binding sites [60], in comparison to nonepitopic sites, in order to create diversity to escape the host antibodies without significantly altering stereochemical or functional properties of the segment[3, 37, 42, 59].

However, an obstacle to studying selection in influenza is sampling bias. In laboratory sequencing, efforts are made to find sequences that are antigenetically novel, thus containing a high dN:dS ratio. This sampling bias causes error in evolutionary analyses as well as the representation of types of influenza viruses circulating.

Viral strains used to make vaccines for the seasonal influenza epidemics are typically made well in advance of the flu season. Much research and development goes into tracking persistent lineages and strains that may be responsible for the epidemics, and typically, phylogenetic trees are used to monitor the progressive evolution and antigenic drift of the virus. Influenza lineages typically have a finite lifespan of a couple years and new lineages typically arise from small populations of strains that are descendants of the ancestors of viruses causing the most recent epidemic, which makes prediction difficult [7]. HA nucleotide sequences provide plenty of evolutionary details; however, it is difficult to analyze the link between genetic change and antigenetic
variation [6]. Although much progress has been made [31], antigenic mutations and mutation patterns are not 100% predictable, which makes it hard to select proper vaccine strains.

It is suggested that four amino acid replacements occurring in at least two of the five antigenic sites are needed in the HA1 of the H3N2 strain to create a new antigenic variant [10]. Due to the great deal of variation in this pattern from year to year, there is a need to look at other correlations which may point to a link between genetic change and antigenic variation. It was previously shown that the strain containing the most positive selection sites within a given influenza season would serve as progenitors for future influenza lineages [7]. The presence of these sites near antibody binding sites and receptor binding pockets were key factors in sustaining lineages [7]. It is suggested that screening new strains for additional replacements at these particular sites may provide an early warning of potentially successful viral strains, and thus improve upon current influenza surveillance and vaccine strain selection methods [6].

The hemagglutinin segment consists of two subunits, HA1 and HA2, which are 329 and 220 residues long, respectively. The major binding regions are located on the globular head of HA1 subunit. Much of the variation and selection occurs within this larger subunit that binds to the host receptors. There are five antigenic regions within H3N2 (Fig. 3) and five receptor binding sites within H1N1, both of which are highly susceptible to positive selection driven by the host’s immune response. It is important to study the structural differences rather than just relying on sequence analysis because some amino acids structurally interacting with the host’s antibody or sialic acid may have a larger influence on the virus’ activities. Positive selections or mutations on these sites are more likely due to survival pressure. There are six conserved residues in H3N2 that comprise the surface of the pocket: Tyr98, Trp153, Glu190, Leu194, His183, and Thr155. Residues 134-138 form the right side of the pocket and residues 224-228 form the left side of the pocket [58].

In this work, we carefully selected viral strains from each given year to analyze, assessed mutation changes during the past decade, and evaluated structural changes caused by these mutations using a variety of computational methods, to collectively study adaptive antigenic mutations of the human influenza H1N1 and H3N2 hemagglutinin. Specifically, we detected overall signal of selection in H1N1 and H3N2, examined positive selection sites in each lineage, estimated selection pressure on each site, and studied the structural changes of significant mutations. We expect to better understand the evolution of the influenza A virus, which will lead to the prediction of proper vaccine strains.

II. MATERIALS AND METHODS

A. Obtaining Sequences

The hemagglutinin sequences used for the positive selection analysis were obtained from the FluGenome database [33], in order to assess sequence lineage annotation (according to FluGenome) and host human specificity. The sequences were obtained from the past ten flu seasons and separated out by season. Using the accession numbers of the obtained sequences, coding nucleotide sequences of at least 1,500 nucleotides in length were obtained from the NCBI Influenza Database. Sequences were aligned using MUSCLE [12]. The signal peptide was predicted using the SignalP 3.0 Server [2] and cut using MEGA [51]. The HA2 subunit was also cleaved prior to analysis, yielding a 987 nucleotide alignment. Ambiguous sequences were removed from the alignment as well as those that produced a gap, unless the number of sequences that contributed to the gap was larger than 5 sequences. Representative samples included in the season to season analysis were manually selected from all parts of the world (if sequences were available) to avoid regional sampling bias as well as to provide for sequence diversity.

B. Detecting Overall Signature of Selection

Analysis was conducted using the modules contained in the standalone Hy-Phy package [43].

The codon data was first analyzed using the Single Breakpoint Module [24] for detection of recombination. The Single Breakpoint Recombination test revealed that no significant breakpoints existed within the data using the cAIC criterion; therefore, the input trees could be used for further analysis.

Maximum parsimony phylogenetic trees constructed with TNT were used for the analysis [18]. Two lineages (1A and 1B) were inferred from the H1N1 HA phylogenetic tree and one lineage, 3A, from H3N2. (Trees available upon request.) Within TNT, the heuristic New Technology search method was implemented. It is known that subtle tree alterations in tree structure can affect nucleotide substitution assignment [7]; therefore, a consensus tree was made using the best scoring trees to avoid structural bias.

Evidence of positive selection was conducted using the module PARRIS[46]. This module has four phases to detect positive selection: nucleotide model maximum likelihood fit, null model M1 (no selection) fit, alternative model M2 (selection) fit, and finally, LRT to test M1 against M2 with two (2) degrees of freedom.

C. Detecting the Positive Selected Codons

Overall, 544 H1N1 and 752 H3N2 sequences were used in the decade analysis. Three methods were used: random effects likelihood (REL), single likelihood ancestor counting (SLAC), and fixed-effects likelihood (FEL).

The H1N1 tree was rooted with the sequence A/South Australia/25/2000 and the H3N2 tree was rooted with the A/Canterbury/50/2000 strain, each being a representative of the earliest year of the data.

For identifying positive selections, we considered the following criteria: an LRT of greater than 10 for FEL and a p-value smaller than 0.2 for both REL and SLAC. To focus more on the most significant selection sites, here we present only those satisfying two out of the three criteria had to be met: an LRT of greater than 15 for FEL, a p-value smaller than 0.1 for REL, and a p-value of smaller than 0.1 for SLAC. For
the all methods, the parameter for the substitution model was the HKY85/ MG94xHKY85 model. The dN/dS bias parameter options was set to neutral (dN/dS=1).

D. Protein Structure Analysis

In order to carry out structural analysis on H1N1 and H3N2 HA1 proteins, we used 8 representative strains from H1N1 and 10 strains from H3N2 families in the last decade and built a homology model for each protein from the above strains. The structure of the 2009 pandemic H1N1 HA protein with PDB accession 3LZG[62] was used as a template to build homologous models for other strains from H1N1 and protein with PDB accession 1EO8[14] was used as a template for all 10 strains of H3N2 (Accession numbers for the viral strains available upon request). The protein homology modeling was carried out using the MOE program [1]. The query sequence was searched against the PDB in MOE using the FASTA methodology[41]. The resulting hit list was evaluated by E-value and by Z-score, which is an estimate of the statistical significance of an alignment score. The alignment between the query sequence strains and the template proteins was evaluated in MOE using the Needleman and Wunsch algorithm[38]. The sequence alignment between the query strains and the template proteins was carried out with ClustalW [26] with high sequence identity, ranging from 87% to 100%. After sequence alignment in MOE, the homologous model for each strain was built using the Homology Model approach in MOE.

All H1N1 model strains were structurally aligned to 3LZG and H3N2 strain proteins to 1EO8 using the DaliLite program [20]. All H1N1 models and H3N2 models were imported to MOE and the receptor binding sites for the HA1 domains of H3N2 isolates and the corresponding sites for H1N1 isolates were identified (Fig. 2 and Fig. 3, respectively).

III. RESULTS AND DISCUSSION

A. Positive Selection

Strong positive selection was detected in the HA of both H3N2 and H1N1 viruses. Specifically, the PARRIS analysis revealed a p-value of 3.7x10^{-5} for the M2a model (positive selection) for H3N2, which is a strong indicator of positive selection. The PARRIS analysis of H1N1 revealed a p-value of 1.0x10^{-7}, which is an even stronger indicator of positive selection.

The H1N1 analysis of positive selection sites yielded eight positively selected sites. For the 1B lineage, there were seven: 98, 144, 190, 192, 193, 260, and 313. For 1A, there was one: 206. Most positions have been previously identified with an antibody binding region or receptor binding domain (Table 1). Under the more strict criteria, positions 144, 190, 192, and 206 were identified as the most significant positive selection sites.

It is important to keep in mind that previous findings indicate egg-adapted mutations occurred in positions 144 and 190 due to passage in chicken eggs prior to sequencing [16, 45, 63]. Past knowledge has described position 144 as being involved with antigenetic drift and receptor binding in H1N1 [28, 47, 64]. However, much of the functional significance of the mutations in this position remains unsolved. Previous studies also indicate that residue190 in 1934 human (H1N1) as well as residues 190 and 193 in 1930 swine (H1N1) were found to directly interact with bound human-like α(2,6)-receptors [17]. It is believed that D190/D225 mutation is necessary for human (H1N1) receptor-binding [44, 49, 53]; however, our results indicate V190, G190, and N190 all exist in the human population, thus refuting previous hypotheses of specificity. Some studies show that D190N of A(H1N1) HA was shown to result in a lower binding affinity for human-like α(2,6) receptors, and a higher binding affinity for avian-like α(2,3) receptors[16]. These results indicate the necessity for the monitoring and updating of previous findings. It also indicates the need to find functional reasoning behind these mutations.

Position 192 is previously noted as a critical amino acid for contact in the Sb receptor binding domain, along with positions 190 and 193 [49]. Others have correlated R192M and A193T with being part of H1N1 amandadine-resistant strains[64]. Codon 192 of the 1B lineage has mainly selected for the R192K mutation (Table 3). However, in 2008, another selection event occurred in Asia and Europe to cause a K192R reversion mutation. In the same year in North America, there was additional positive selection with a K192N mutation. Furthermore, R192S and R192M have taken place late in the decade from 2006-2009. This goes to show how rapidly the influenza virus can select for advantageous amino acids in a short time period.

Recent studies concur that position 206 is under selection pressure, but more interestingly, this position has not been under selection with previous H1N1 pandemic outbreaks[40]. It can be suggested that the S206T mutation may affect the infectivity and transmissibility of 2009 H1N1 in humans due to its presence in the Ca binding domain.

It is necessary to take into account the binding sites under strong selection pressure for both lineages of human H1N1 currently in circulation, since antigenic variation occurs from year to year. It may be best to design vaccines and drugs that target residues currently under selection pressure to keep up with the mutating hemagglutinin structure.

Table 1 Positive selection sites of H1N1 detected by different methods (REL, SLAC, and FEL) from strains between 2000 and 2010.

<table>
<thead>
<tr>
<th>FEL</th>
<th>REL</th>
<th>SLAC</th>
<th>Ab-binding Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>144*</td>
<td>144</td>
<td>144</td>
<td>Ca**</td>
</tr>
<tr>
<td>190*</td>
<td>190</td>
<td>190</td>
<td>Sb**</td>
</tr>
<tr>
<td>192*</td>
<td>192</td>
<td></td>
<td>Sb</td>
</tr>
<tr>
<td>193</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>206*</td>
<td>206</td>
<td>206</td>
<td>Ca**</td>
</tr>
<tr>
<td>260</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>313</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* sites satisfying at least two of the stringent criteria
**: one amino acid away from the Ab-binding site
Table 2 Summary table of codon substitutions for H1N1 derived from phylogenetic codon substitution maps using HyPhy. Substitutions are given by year and region.

<table>
<thead>
<tr>
<th>Influenza Season</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>144</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu → Gly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys → Arg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>190</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp → Val</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp → Gly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp → Asn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>192</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg → Arg *Reversion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys → Asn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg → Ser</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg → Met</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>206</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: Reversion indicates a reversal of a previous substitution.*
The H3N2 analysis of positive selection suggested 21 selected sites, of which eight meet the stringent criteria: 53, 138, 142, 144, 157, 173, 225, and 226. In contrast to the H1N1 analysis, a majority of these codons lie within known antigenic regions (Table 3). Codon 225 and 226 both make up the left side of the receptor binding pocket and codon 138 is within the right side of the pocket, in addition to their respective antigenic sites. Previous studies have indicated 18 positive selected sites[8] in an analysis of 357 viruses from 1984 to 1996, and a study in 1997 reasoned that 6 hypervariable positions [13] exist in human H3N2. Residues 144, 157, and 173 were identified as potential immunodominant positions in previous work[29].

Table 3 Positive selection sites of H3N2 detected by different method (REL, SLAC, and FEL) from strains between 2000 and 2010

<table>
<thead>
<tr>
<th>REL</th>
<th>SLAC</th>
<th>FEL</th>
<th>Antigenic Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53*</td>
<td>53</td>
<td>53</td>
<td>C</td>
</tr>
<tr>
<td>92</td>
<td></td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>112</td>
<td></td>
<td></td>
</tr>
<tr>
<td>138*</td>
<td>138</td>
<td>138</td>
<td>A</td>
</tr>
<tr>
<td>142*</td>
<td>142</td>
<td>142</td>
<td>A</td>
</tr>
<tr>
<td>144*</td>
<td>144</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>145</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>156</td>
<td>156</td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>157*</td>
<td>157</td>
<td>157</td>
<td>B</td>
</tr>
<tr>
<td>158</td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>173*</td>
<td>173</td>
<td>173</td>
<td>D</td>
</tr>
<tr>
<td>189</td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>220</td>
<td>220</td>
<td>220</td>
<td>D</td>
</tr>
<tr>
<td>223</td>
<td></td>
<td></td>
<td>D</td>
</tr>
<tr>
<td>225*</td>
<td>225</td>
<td>225</td>
<td>D</td>
</tr>
<tr>
<td>226*</td>
<td>226</td>
<td></td>
<td>D</td>
</tr>
<tr>
<td>229</td>
<td>229</td>
<td></td>
<td>D</td>
</tr>
<tr>
<td>275</td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>291</td>
<td>291</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* sites satisfying at least two of the stringent criteria

Codon 53 has previously been selected under pressure due to a mouse monoclonal antibody[27], but functional relevance in regard to the human host is poorly understood.

Codons 138 and 226 have previously been identified as hypervariable codons[13]. These codons have been observed to change during growth in eggs and have evolved at a much higher rate than any other residues[23][13]. Residue 138 has primarily selected for the A138S mutation with one exception of A138T in North America during 2005. However, the A138T is a mutation previously identified with passage in eggs[55]. The A138S may be regarded as favorable without any need to select for a different amino acid. This selection site may seem odd due to its role in receptor-binding specificity[8]. Further research needs to be done to understand the functional advantage of this mutation.

Codon 142 has previously been identified as being positively selected as well[8]. This position has previously been labeled as having a high Shannon entropy and low information gain (IG) value, which would indicate a low probability to cause an antigenic variant[21].

Codon 144 has been reported to be under positive selection and much functional relevance of mutations in this position has been documented[4]. There has been evidence of positive selection at this position during egg passage prior to isolation[9]. In contrast to position 142, changes in 144 and site A have been associated with creating antigenic variants responsible for yearly epidemics[59]. The loop comprising antigenic site A makes few contacts with other parts of the structure, so selection changes in 144 may not have a significant impact on the overall structure of hemagglutinin[4]. This site also has a well documented history with its interactions with antibodies as the post-infection sera contains antibodies that mainly recognize residue 144[25, 56], which may be the driving force of selection pressure at this position. Reports indicate that G144 is most likely not essential for the increased affinity of human H3N2 synthetic sialyglycopolymers with a 6'-sialyl(N-acetyllactosamine) receptor[34], and it has been noted that antibodies tend to show less reactivity to G144 when compared to D144[54]. N144D was highly selected for through the past decade, with a V144G only occurring once, suggesting the D144 is more favorable for the H3N2 virus. This N144D mutation is thought to have ended a lineage, and in the 2002/2003 season, strains with 144N gave rise to the A/Fujian/411/2002 antigen variant strains[64]. Previous glycosylation sites are also reported at residue 144[4, 52, 64], and our findings indicate that at the beginning of the decade, N144 was the primary amino acid present. There have been substitutions of N144I early on in the decade, indicating an unmasking of a glycosylation site. Evidence also exists of reversion to I144N at which point the virus gains back a glycosylation site. The most recent selection favors the N144K mutation, an unmasking of the glycosylation site at this position[64] (available as supplementary material).

Very little functional significance is associated with residue 157. Previous reports suggest that P157 creates a lack of interaction between antibodies and hemagglutinin. However, the only amino acids present at this position in the past decade were L157 and S157. Selection towards the mutation L157S was taking place in the second half of the decade.

Studies have been mixed on the issue of position 173 contributing to antigenic variants in the past. Bulimo et al. believe that past mutations involving the mutation converting the basic lysine to an acidic glutamic acid have mutated the HA structure to prevent current antibodies from binding and thus have created antigenic variants[5]. On the other hand, Huang et. al. found position 173 to have a low antigenic discriminating score based on a model that considers significant antigenic variants to have a high IG score and high entropy score[21]. Position 173 came under considerable antigenic pressure in the second half of the decade with
mutations K173N/E/R/T/Q and Q173K/N. The amount of variation may implicate position 173 as a frequent antibody binding site; therefore, selection pressure causes a significant amount of mutation.

Codon 225 is involved with receptor binding in HA, suggesting that positive selection in this site may be directly involved with host pressure. From past studies, it appears that mutations in the receptor binding pocket of residues 225 and 227 have not been too much of a determining factor in creating antigenic variants, as changes in sites A, B, and residue 226 have been more significant in determining variants [4, 61]. The selection that was seen the most within residue 225 was the G225D mutation. This mutation, G225D, is thought to increase the virus’s ability to grow in eggs [19, 22], which makes strains having this mutation good vaccine selections. However, this same substitution, combined with the W222R substitution, is known to reduce human H3N2 HA receptor-binding activity and viral replication efficiency [36, 48].

Another substitution of importance is the D225N mutation. It is suggested that D225N, S193F (same residue correlated with H1N1 amantadine resistance strains), or a combination of both may have a role in maintaining the S31N mutation in M2[15]. It is worthwhile to point out that mutations (such as D225N) on position 225 have been observed more frequently in the past three years (Available as supplementary material). This appears relevant as the S31N mutation is a key mutation to amantadine drug resistance. However, the N225E mutation that took place in North America during 2009 and the reversion N225D mutations taking place in North America during 2007 and 2008 may decrease the ability of the virus to maintain the S31N mutation, and thus render the virus susceptible to amantadine. The use of other drugs targeting NA may be the reason for the N225E and N225D mutations, as pressure on N225 to maintain the amantadine resistant S31N mutation lessens. However, these mutations may be due to continuing viral evolution or competitive disadvantages of resistant strains versus sensitive strains [64].

In addition to residue 225, codon 226 is another documented sialic acid receptor binding site for the hemagglutinin segment [61]. Mutations at this residue are also thought to affect receptor binding only, thus having no impact on the fusion event [39]. Changes in this codon, in 2002 are believed to have contributed to the creation of an antigenic variant [61]. According to our analysis, the mutation that took place to create this variant was a V226I mutation. Although the primary selection was for V226I, reversion has taken place at least once in 2005. The V226I (along with the recombinant V226A mutation) selection is thought to improve replication within embryonated chicken egg [32]. However, some studies indicate that a combination of V226I, T155, and T156 results in the inability of growth in eggs [22, 35]. Further analysis shows that a V226L selection has at least taken place twice in Asia during 2002 and 2008. Previous studies have indicated that leucine at position 226 is a preferential amino acid for human α(2,6)–linked sialic acid receptors[11]. Furthermore, studies have shown that the L226Q mutation may reduce affinity of the HA for sialosides with α(2,6) linkage relative to α(2,3) linkages[58], although this mutation was not seen to have taken place during the last decade according to our data.

Codons substitution mapping helps to depict origins of selection sites in terms of year and location (Tables 2 and 4). For instance, mapping of residue 225 suggests that early in this decade the substitution of G225D gave rise to the persisting H3N2 clade of the decade in Southeast Asia. The mutation of D225N in 2005 changed the conformation at the glycosylation site, thus allowing the virus to maintain circulation. Further substitutions of N225 with D/G225 as observed in 2007, 2008, and 2009 allowed for unmasking of the glycosylation site. Glycosylation is a known method, using carbohydrate side chains, to mask antibody binding residues from the host immune system. Reversion or unmasking of a glycosylation sites also aids in escaping the host immune system.

One possible reason that H3N2 possesses more selection sites than H1N1 is the larger population circulating each flu season except for 2009 (Figure 1). This may allow more interaction with host immune systems and thus present more pressure on the virus.

![Number of HA Isolates each Year](image)

**Figure 1 The number of H1N1 and H3N2 isolates during the past decade**

### B. Structural Analysis

The structures for the representative H1N1 and H3N2 strains in the past decade are shown in Figures 2 and 3. Amino acid variations at a given position is listed as A/S/T193, indicating residues Ala, Ser and Thr were observed at position 193 in the past 10 years. Residue 144 in the site Ca is a residue frequently mutated during 2006-2008. In the past three years, however, more positive selections were observed at and around site Sb, featuring residues 190, 192, and 193 which were frequently mutated during virus evolution. Our H1N1 results tended to agree with the previous hypothesis [3], suggesting that lineages in which the most positive selection occurred gave rise to persisting lineages during antigenic drift. (Phylogenetic trees available upon request.) It is pivotal to monitor both the 1A and 1B lineages of H1N1 rather than focusing on one lineage of the serotype. It is interesting to note that the positive selection mutation of S206T already is occurring within the 1A lineage of the H1N1 Swine Flu from year 2009 to 2010. This may be due to the host immune
response or in response to the current vaccines and drugs being administered. In preparing for the 2010 to 2011 influenza season, the recommended vaccine is the strain A/California/07/2009. This strain contains a positively selected codon at 206, which makes it a good candidate for the seasonal vaccine strain for the 1A lineage according to our positive selection and structural analyses. However, potential still exists for the 1B lineage to arise, and for that reason, we suggest strain A/Wisconsin/13/2009, due to its two positive selection sites, 190 and 193 at site Sb. It contained the most positive selection sites (sites dictated by our analysis) in 2009 for the 1B lineage.

In combination with the detailed year over year table (available as supplementary material), it appears that very few positive selections were observed in the first three years in the past decade in H3N2. Positive selections were observed during 2003-2006 in residues 138, 142, and 144, which are located near site A. The latest selection sites during 2008-2010 were mainly observed in residue 157, located in site B, and residues 225 and 226 at the left side of the receptor binding pocket. The investigation of the interactions between H3N2 HA domain and the bound antibody light (L) chain and heavy (H) chain shows that residues 141-144 (Site A, red loop) interact with residues Ser56, Gly57, Val58 and Pro59 of the antibody L chain and that residues Glu62, Asp63 and His75 in Site E (brown loops) form H-bonds with Arg98, Arg94, and Trp100 of the H chain. The survival pressure to escape the antibody L chain binding may explain the frequently observed positive selections at positions 142 and 144. The structural proximity of residues 50, 53, 58, 96, 273, and 275 to the antibody H chain provides a reasonable explanation for positive selections observed in those positions between 2000-2007. On the contrary, the distances between the antibody H and L chains and the positive selection sites on H1N1 HA1 domain are very far away. This may provide reason for a smaller number of observable positive selections in H1N1. Residues 157 and 225 are not in proximity to the antibody H and L chains; however, both residues are on the same surface plane and are located in the readily accessible loop regions. The frequently observed positive selections at 157 and 225 indicate that the loop regions near residues 157 and 225 could be the target sites for new vaccine development. There were six 2009 strains with selection mutations in either 157 or 225; however, both selections were not seen within a single strain in our dataset. All strains containing a mutation at 157 or 225 also contained the selected site at 173. Accession numbers of strains were: ADG21028, ACV72368, ACT67773, ACT67801, ACV72367, and ACT67817. In addition, only two contained the selected site at 226: A/Washington/19/2009 (ACV72367), and A/Georgia/07/2009 (ACV72368).

The selected vaccine strain for H3N2, A/Perth/16/2009, only contained selected sites at 173 and 226. If previous hypothesis are correct, then the strains with the most positive selected sites may be the best candidates for vaccine strains: A/Washington/19/2009 or A/Georgia/07/2009. The Washington strain contained the positive selected sites at 173, 225, and 226 while the Georgia strain contained the selected sites 157, 173, and 226. Our positive selection analyses and structural information suggest residues 157, 173, 225, and 226 warrant further investigations for vaccine development against H3N2 strains.
IV. CONCLUSION

In our analysis, we found eight positively selected codons in H3N2 and four positively selected codons in H1N1 that met stringent criteria for positive selection identifications. We created substitution maps and summary tables in order to identify the mutations within certain regions and time periods. We also linked the mutations structurally in order to identify regions where the best drug docking interactions may occur. We identified codon 206 as a potential docking site for the 1A lineage (swine influenza) and concurred that A/California/07/2009 was a reasonable choice for a vaccine strain due to the presence of the S206T mutation. However, the 1B lineage of H1N1 that has been lingering during the past few decades has displayed strong selection at codons 144, 190, and 192, which is why we suggest A/Wisconsin/13/2009 as an additional vaccine strain to consider. We also suggest A/Washington/19/2009 or A/Georgia/07/2009 for vaccine consideration for H3N2, due to the presence of current selected sites at 225 and 157, respectively. In addition, these strains contain a significant number of selection sites, making them reasonable candidates for a vaccine strain.

For future analysis, we plan to look deeper into the structural and functional relevance of the positively selected residues. We also find it reasonable to compare strains with significant positive selection sites to the WHO selected vaccine strains. As we mentioned in the introduction, it will be interesting to trace mutation changes of H1N1 following the 2009 Swine Flu season. Other work includes improving upon computational methods to predict mutation variation and patterns involved for better vaccine strains prediction.

ACKNOWLEDGMENT

This publication was made possible by NIH grant numbers R01 LM009985-01A1. The authors also acknowledge the UCRCA, the University of Nebraska at Omaha (UNO), for continuous funding support to this research program. We are grateful to Dr. Ruben Donis of the Centers for Disease Control and Prevention (CDC), U.S. Department of Health and Human Services for constructive discussion. We are also give gratitude to Mary Christman, Department of Biology, UNO, for proofreading.

REFERENCES


