Protein Engineering, Design & Selection vol. 22 no. 9 pp. 543–546, 2009
Published online July 3, 2009 doi:10.1093/protein/gzp027

The epitope regions of H1-subtype influenza A, with application to vaccine efficacy

Michael W. Deem[1,2,3] and Keyao Pan[1]

Departments of [1]Bioengineering and [2]Physics and Astronomy, Rice University, 6100 Main Street, MS 142, Houston, TX 77005-1892, USA

[3]To whom correspondence should be addressed.
E-mail: mwdeem@rice.edu

The recent emergence of H1N1 (swine flu) illustrates the ability of the influenza virus to create antigens new to the human immune system, even within a given hemagglutinin and neuraminidase subtype. This new H1N1 strain is sufficiently distinct, for example, from the A/Brisbane/59/2007 (H1N1)-like virus strain of influenza in the 2008/09 Northern hemisphere vaccine that protection is not expected to be substantial. The human immune system responds primarily to the five epitope regions of the hemagglutinin protein. By determining the fraction of amino acids that differ between a vaccine strain and a viral challenge strain in the dominant epitope regions, a measure of antigenic distance that correlates with epidemiological studies of H3 influenza A vaccine efficacy in humans with $R^2 = 0.81$ is derived. This measure of antigenic distance is called p_{epitope}. The relation between vaccine efficacy and p_{epitope} is given by $E = 0.47 - 2.47 \times p_{\text{epitope}}$. We here identify the epitope regions of H1 hemagglutinin, so that vaccine efficacy may be reliably estimated for H1N1 influenza A.

Keywords: antibody/epitopes/influenza A/swine flu

Introduction

The recent outbreak of H1N1 (swine flu) has caused immediate international concern. From its earliest case in mid-March 2009 to mid-May 2009, 8000–9000 infections and 70–80 deaths were recorded in 40–50 countries and regions, and as of mid-May, over 90% of infections and deaths were in Mexico and the USA (World Health Organization, 2009). Historically, three subtypes of influenza A virus have been able to circulate in the human population. The Spanish flu pandemic in 1918–20 was H1N1, which circulated in the world until 1957. H1N1 reappeared in 1977 and persists today (Nakajima *et al*., 1978). The Asian flu pandemic in 1956–58 was H2N2, which spread widely in the human population during the time interval 1957–68 (Palese *et al*., 2006). The Hong Kong flu pandemic in 1968–69 was H3N2, which has circulated in the human population as the dominant subtype until recently (Palese *et al*., 2006). Other subtypes rarely infected humans, although cases of H5N1 and H9N2 have been reported.

The 2009 swine flu virus possesses H1 hemagglutinin (HA) and N1 neuraminidase on the surface of the virion, of which the hemagglutinin is the main target of host antibodies. The human immune system responds primarily to the five epitope regions of the hemagglutinin protein (Bush *et al*., 1999; Macken *et al*., 2001). Host antibodies bind to five epitopes in hemagglutinin and lead to high escape evolution rates of amino acids in the epitopes. An early identification of H1 epitopes was carried out by antibody mapping of the A/PR/8/1934 (H1) hemagglutinin, with an additional study of laboratory mutations (Caton *et al*., 1982). However, these H1 epitopes contain far fewer amino acids than do the epitopes in H3 determined by modern methods (Macken *et al*., 2001) and are incomplete. Alignment of H1 strains in 1918–2009 indicates many mutation positions outside the originally identified epitopes. We here use sequence alignment and information entropy to complete the definition of H1 epitopes.

Vaccination is an effective way to reduce the influenza morbidity and mortality. The efficacies of influenza vaccines vary from year to year, in part due to different antigenic distances between the circulating influenza strains and the vaccine. Antigenic distance between a vaccine strain and a viral strain can be estimated by the number of mutations in the hemagglutinin sequence between the two strains (Smith *et al*., 1999, 2004). Ferret animal model studies are used to further refine the notion of antigenic distance. These methods correlate with epidemiological studies of vaccine efficacy in humans with $R^2 = 0.59$ and 0.43–0.57, respectively (Gupta *et al*., 2006; Zhou *et al*., 2009). By considering only those mutations that occur in the dominant epitope, the p_{epitope} theory provides a prediction of vaccine efficacy that correlates with epidemiological studies of vaccine efficacy in humans with $R^2 = 0.81$.

Vaccine efficacy has a linear correlation with the antigenic distance between the vaccine strain and the circulating virus strain (Gupta *et al*., 2006; Zhou *et al*., 2009). Since p_{epitope} correlates well with influenza vaccine efficacy in humans, it can be used to estimate antigenic distance. For example, when p_{epitope} is larger than 0.19, the vaccine no longer offers protection. This correlation can be used to find optimal strains for vaccine design with minimal antigenic distance from expected circulating strains. Here we calculate the antigenic distance for H1 influenza A and apply the method to evaluate efficacy of a candidate swine flu vaccine.

Methods

Mapping the epitope from H3 hemagglutinin to H1 hemagglutinin

In this paper, the amino acid positions of H1 (A/PR/8/1934) and H3 (A/Aichi/2/1968) hemagglutinin are denoted using H1 and H3 numbering, respectively (Winter *et al*., 1981; Nobusawa *et al*., 1991). Two hemagglutinin sequences of A/PR/8/1934 (H1) and A/Aichi/2/1968 (H3) are aligned using ClustalW. Amino acids in H1 sequence corresponding to

epitope A–E in H3 are defined as mapped epitope A–E in H1. Similarity between H3 epitopes and corresponding mapped H1 epitopes was verified by aligning their three-dimensional structures (PDB code: H3 = 1HGF and H1 = 1RU7). The RMSD values of alpha carbons in five epitopes between H3 and H1 are 2.18, 0.63, 1.19, 2.43 and 1.90 Å, respectively.

The HA alignment of A/California/04/2009 with A/PR/8/1934 and of A/California/04/2009 with A/Aichi/2/1968 shows that neither A/PR/8/1934 nor A/Aichi/2/1968 has a gap-free alignment with A/California/04/2009. Thus, a new numbering scheme is required for the 2009 H1N1 ('swine flu') hemagglutinin. The new numbering starts at the same amino acid position as the H1 numbering (Winter *et al*., 1981; Nobusawa *et al*., 1991). Amino acid 130 in A/California/04/2009 corresponds to a gap in A/PR/8/1934, indicating that amino acid position >130 in A/California/04/2009 has the number equal to 1 plus the corresponding number of the amino acid in A/PR/8/1934. In Table I, we use the A/California/04/2009 numbering scheme. HA alignment of swine flu strains deposited in NCBI and GISAID up until 18 May 2009 shows that no mutation occurred in amino acid 130. By examining the three-dimensional structure of A/PR/8/1934 (PDB code: 1RU7), we find that amino acid 129 (A/PR/8/1934 numbering) is exposed and position 130 is partially buried inside the molecule. Consequently, although amino acid 130 (A/California/04/2009 numbering) in swine flu HA may or may not be significant to antibody binding, we have no evidence that it is in the epitope.

Extension of mapped epitope using entropy method

The definition of information entropy of site k is

$$S(k) = -\sum_{i=1}^{20} \frac{n_i}{N} \ln \frac{n_i}{N}$$

where n_i is the number of times that amino acid i ($i = 1–20$) is found in site k of the aligned full-length strains. N is the number of those full-length sequences, equal to 2294.

Table I. Amino acids in epitopes A, B, C, D and E of H1 (A/California/04/2009 numbering, modified from Caton *et al*., 1982)

Epitope	Amino acids
A	118 120 121 122 126 127 128 129 132 133 134 135 137 139 140 141 142 143 146 147 149 165 252 253
B	124 125 152 153 154 155 156 157 160 162 183 184 185 186 187 189 190 191 193 194 195 196
C	34 35 36 37 38 40 41 43 44 45 269 270 271 272 273 274 276 277 278 283 288 292 295 297 298 302 303 305 306 307 308 309 310
D	89 94 95 96 113 117 163 164 166 167 168 169 170 171 172 173 174 176 179 198 200 202 204 205 206 207 208 209 210 211 212 213 214 215 216 222 223 224 225 226 227 235 237 239 241 243 244 245
E	47 48 50 51 53 54 56 57 58 66 68 69 70 71 72 73 74 75 78 79 80 82 83 84 85 86 102 257 258 259 260 261 263 267

For A/PR/8/1934 numbering, amino acid numbers above 130 would have 1 subtracted from them.

The threshold of the information entropy values to identify epitopes is determined adaptively. Amino acids in epitopes are under immune pressure selection and evolve to avoid recognition by antibodies, yielding large information entropy values for amino acids in epitopes among all H1 strains. Additionally, classical epitopes are defined to be on the surface of the three-dimensional structure of the hemagglutinin. We decreased the entropy threshold until amino acids inside the structure constitute a significant proportion of those amino acids newly incorporated into the epitopes. Here, we select 0.075 as the threshold.

Phylogenetic tree and its root

Three hundred and twenty H1N1 swine flu protein sequences as of 18 May 2009 were downloaded from the NCBI and GISAID databases. ClustalW is applied to align these strains. Two hundred and sixty-six sequences containing residues 27–324 (A/California/04/2009 numbering, modified from Caton *et al*., 1982) were used to create the phylogenic tree with PHYLIP (Felsenstein, 1989). Two hundred and sixty-six subsequences that included residues 27–324 were extracted, and duplicated subsequences were removed. The 23 unique aligned subsequences were then used as the input of the program PHYLIP. The distance matrix from protein sequences (protdist) and the Fitch–Margoliash and least squares methods with evolutionary clock (kitsch) (Felsenstein, 1989) were sequentially applied to generate the phylogenic tree in Fig. 1. The output of protdist is the input of kitsch. In the phylogenetic tree, the strain A/California/07/2009 (FJ966974) is one of the outgroups, which are closest to the root.

Calculation of $p_{epitope}$

The antigenic distance between two strains is calculated from the amino acid sequence in five epitopes of hemagglutinin. For each epitope, the P-value is defined as the proportion of different amino acids between these two strains. The largest of the five P-values is defined as $p_{epitope}$, and the corresponding epitope is defined as the dominant epitope (Gupta *et al*., 2006; Zhou *et al*., 2009).

Results

Antigenic distance

As of 18 May 2009, there are 320 H1N1 swine flu strains in the NCBI and GISAID databases, and we focus on the interval covering residues 27–324 (A/California/04/2009 numbering, modified from Caton *et al*., 1982) in 266 of 320 sequences covering this interval. Among these 266 sequences, the closest strain to the root of a phylogenetic reconstruction (Felsenstein, 1989) is A/California/07/2009 (FJ966974) (Fig. 1). There are 20 mutations in the population of sequences, referenced to this strain, with a maximum Hamming distance of 4. Using the $p_{epitope}$ method, we find the largest $p_{epitope}$ value is 0.059 (dominant epitope = E). A/California/04/2009 (FJ966082) is a candidate for vaccine design. There are 20 mutations in the population of sequences, referenced to A/California/04/2009, with a maximum Hamming distance of 5. The largest $p_{epitope}$ value between this candidate strain and all sequences deposited to date is 0.059 (dominant epitope = E), suggesting a worst-case

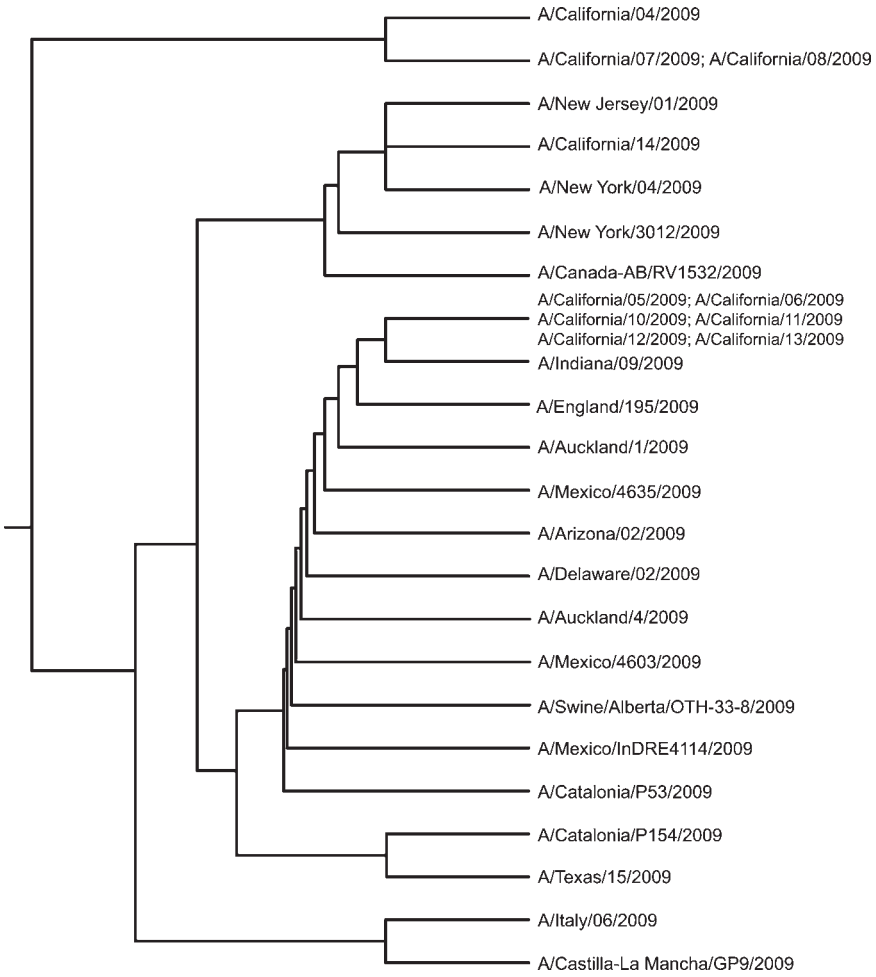

Fig. 1. Phylogenetic tree of swine flu hemagglutinins deposited in NCBI and GISAID until 18 May 2009. For each tip containing over two strains, representative strains are marked.

vaccine efficacy against strains deposited to date of 69.1% of that of a perfect-match, $p_{\text{epitope}} = 0$, vaccine.

Epitope identification

To construct p_{epitope} for H1N1, the identities of the epitope regions in H1 are needed. An early antibody mapping experiment on the PR/8 strain of H1N1 identified 9, 6, 6, 5 and 6 amino acids belonging to the five epitopes Sa, Sb, Ca_1, Ca_2 and Cb, respectively (Caton *et al.*, 1982). These H1 epitopes map by sequence homology to the H3 epitopes B+D, B, C+D, A+D and E, respectively. Interestingly, for those H1 epitopes that map to multiple H3 epitopes, individual antibodies bound to all H1 epitopes mapping to the same set of H3 epitopes, suggesting that the identification of the H1 epitopes may be subject to refinement. The number of amino acids determined by antibody mapping to be in the five epitope residues of H1 is about one-third of the number of amino acids identified in the five epitopes of H3 influenza A (Smith *et al.*, 1999, 2004).

We determined the likely remaining members of the five epitope regions in H1 by information entropy methods. On 5 May 2009, we downloaded all 2735 H1 human influenza A strains from NCBI. We retained the 2294 full-length sequences. We constructed a sequence entropy diagram (Fig. 2) from these data. The refined H1 epitopes were constructed by (i) mapping the known H3 epitopes (Smith *et al.*, 1999, 2004) to the H1 sequence, in rough agreement with the skeleton of sites identified by early antibody mapping experiments (Caton *et al.*, 1982) (sequence

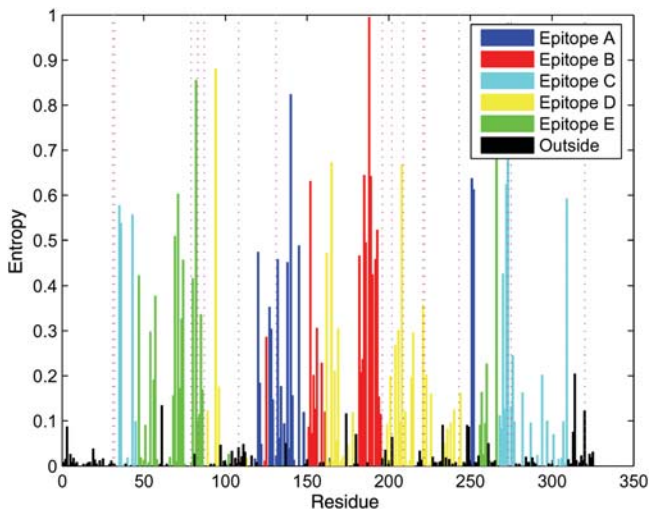

Fig. 2. Sequence entropy for the human strains of H1 (A/PR/8/1934 numbering, as in Caton *et al.*, 1982). Positions belonging to predicted epitopes are color coded by the epitope identity.

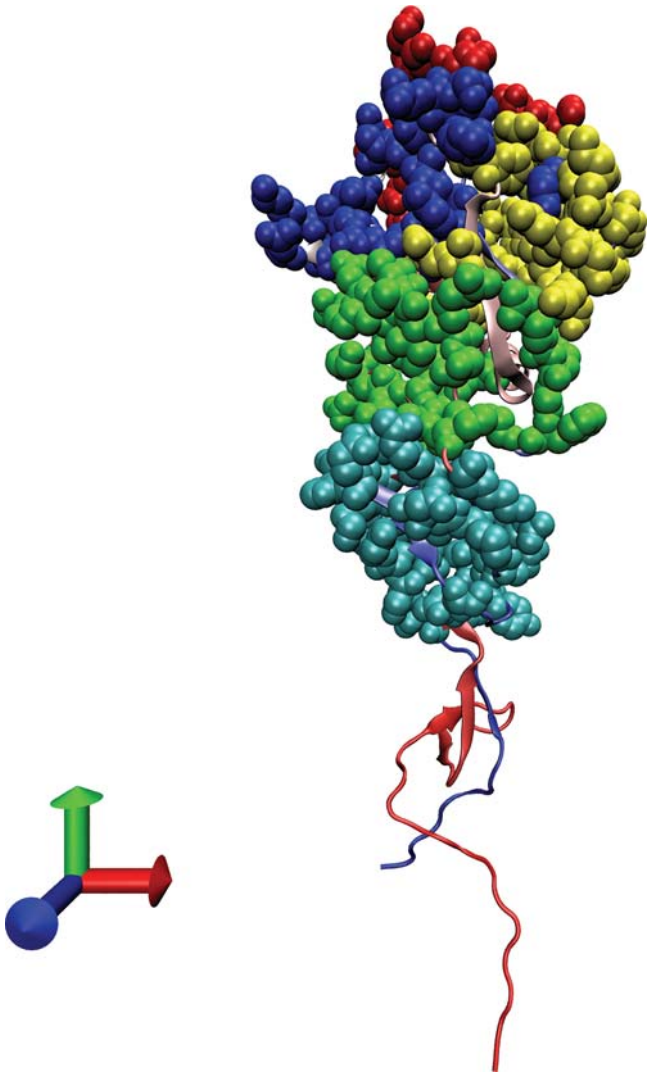

Fig. 3. Color-coded epitopes in the H1 structure (PDB code: 1RU7).

mapping determined by ClustalW alignment of A/PR/8/1934 to A/Aichi/2/1968), and (ii) 31 additional sites identified as being under selective immune pressure, on the surface of the hemagglutinin protein, and with information entropy values >0.075.

Interestingly, eight amino acids were identified with information entropy >0.075 that were outside the conventional definition of epitopes. Three of these were in the tail region of the hemagglutinin (positions 3, 314 and 320, numbering as in Caton *et al*., 1982). The other five were not surface residues (positions 61, 174, 233, 248 and 249) (Fig. 3). Although these residues can affect the geometry at the surface, and so can be under selective pressure, they are not available for presentation to antibodies and so cannot be within a classically defined antibody epitope.

Discussion

With the epitopes in H1 identified, p_{epitope} can be constructed for H1 influenza A. The parameter p_{epitope} provides a quantitative definition of antigenic distance. With this measure of antigenic distance, vaccine strains as 'close' as possible to potential circulating strains can be identified. This capability should be useful in the design of the H1N1 component of the annual influenza vaccine. This capability should also be useful for special situation H1N1 vaccines, such as a vaccine for the recently emerged swine flu.

The p_{epitope} measure of antigenic distance can be used to estimate vaccine efficacy. Vaccine efficacy is $(u-v)/u$, where u is the probability (or rate) that unvaccinated people are infected and v the probability (or rate) that vaccinated people are infected. By analogy with the H3N2 study (Gupta *et al*., 2006; Zhou *et al*., 2009), we expect vaccine efficacy will be well predicted by the equation $E = 0.47 - 2.47 \times p_{\text{epitope}}$. This equation predicts, for example, that a single mutation in dominant epitope B would lead to a vaccine efficacy that is 76.1% of that of a perfect match between vaccine and virus. This equation also predicts that vaccine efficacy is no longer positive for $p_{\text{epitope}} > 0.19$.

Funding

Funding to pay the Open Access publication charges for this article was provided by the FunBio program of DARPA.

References

Bush,R.M., Fitch,W.M., Bender,C.A. and Cox,N.J. (1999) *Mol. Biol. Evol.*, **16**, 1457–1465.
Caton,A.J., Brownlee,G.G., Yewdell,J.W. and Gerhard,W. (1982) *Cell*, **31**, 417–427.
Felsenstein,J. (1989) *Cladistics*, **5**, 164–166.
Gupta,V., Earl,D.J. and Deem,M.W. (2006) *Vaccine*, **24**, 3881–3888.
Macken,C., Lu,H., Goodman,J. and Boykin,L. (2001) The value of a database in surveillance and vaccine selection. In Osterhaus,A.D.M.E., Cox,N. and Hampson,A.W. (eds), *Options for the Control of Influenza IV*. Elsevier Science, Amsterdam.
Nakajima,K., Desselberger,U. and Palese,P. (1978) *Nature*, **274**, 334–339.
Nobusawa,E., Aoyama,T., Kato,H., Suzuki,Y., Tateno,Y. and Nakajima,K. (1991) *Virology*, **182**, 475–485.
Palese,P., Tumpey,T.M. and Garcia-Sastre,A. (2006) *Immunity*, **24**, 121–124.
Smith,D.J., Forrest,S., Ackley,D.H. and Perelson,A.S. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 14001–14006.
Smith,D.J., Lapedes,A.S., de Jong,J.C., Bestebroer,T.M., Rimmelzwaan,G.F., Osterhaus,A.D.M.E. and Fouchier,R.A.M. (2004) *Science*, **305**, 371–376.
Winter,G., Fields,S. and Brownlee,G.G. (1981) *Nature*, **292**, 72–75.
World Health Organization (2009), http://www.who.int/csr/don/2009_05_17/en/index.html
Zhou,H., Pophale,R. and Deem,M.W. (2009) Computer-assisted vaccine design. In Wang,Q. and Tao,Y.J. (eds), *Influenza: Molecular Virology*. Horizon Scientific Press, Norfolk, UK.

Received May 28, 2009; revised May 28, 2009; accepted June 3, 2009

Edited by Harold Sheraga