J Mol Evol (2011) 72:90–103
DOI 10.1007/s00239-010-9405-4

# Selective Pressure to Increase Charge in Immunodominant Epitopes of the H3 Hemagglutinin Influenza Protein

**Keyao Pan · Jinxue Long · Haoxin Sun · Gregory J. Tobin · Peter L. Nara · Michael W. Deem**

Received: 22 January 2010 / Accepted: 25 October 2010 / Published online: 18 November 2010


**Abstract** The evolutionary speed and the consequent immune escape of H3N2 influenza A virus make it an interesting evolutionary system. Charged amino acid residues are often significant contributors to the free energy of binding for protein–protein interactions, including antibody–antigen binding and ligand–receptor binding. We used Markov chain theory and maximum likelihood estimation to model the evolution of the number of charged amino acids on the dominant epitope in the hemagglutinin protein of circulating H3N2 virus strains. The number of charged amino acids increased in the dominant epitope B of the H3N2 virus since introduction in humans in 1968. When epitope A became dominant in 1989, the number of charged amino acids increased in epitope A and decreased in epitope B. Interestingly, the number of charged residues in the dominant epitope of the dominant circulating strain is never fewer than that in the vaccine strain. We propose these results indicate selective pressure for charged amino acids that increase the affinity of the virus epitope for water and decrease the affinity for host antibodies. The standard PAM model of generic protein evolution is unable to capture these trends. The reduced alphabet Markov model (RAMM) model we introduce captures the increased selective pressure for charged amino acids in the dominant epitope of hemagglutinin of H3N2 influenza ($R^2 > 0.98$ between 1968 and 1988). The RAMM model calibrated to historical H3N2 influenza virus evolution in humans fit well to the H3N2/Wyoming virus evolution data from Guinea pig animal model studies.

**Keywords** Influenza · Virus evolution · Pepitope

**Electronic supplementary material** The online version of this article (doi:10.1007/s00239-010-9405-4) contains supplementary material, which is available to authorized users.

K. Pan · M. W. Deem (✉)
Department of Bioengineering, Rice University,
6100 Main Street, Houston, TX 77005, USA
e-mail: mwdeem@rice.edu

M. W. Deem
Department of Physics & Astronomy, Rice University,
6100 Main Street, Houston, TX 77005, USA

J. Long · G. J. Tobin · P. L. Nara
Biological Mimetics, Inc., 124 Byte Drive, Frederick,
MD 21702, USA

H. Sun
Department of Biomedical Engineering,
The Johns Hopkins University, 3400 N Charles Street,
Baltimore, MD 21218-2694, USA

G. J. Tobin · P. L. Nara
Department of Biomedical Sciences, Rm. 2052,
College of Veterinary Medicine, Iowa State University,
1600 S 16th Street, Ames, IA 50011, USA

## Introduction

Influenza A virus causes annual global epidemics resulting in severe morbidity and mortality. The dominant circulating virus today is the H3N2 virus, which emerged in 1968 and is defined by two kinds of surface glycoproteins: H3 hemagglutinin and N2 neuraminidase. It is currently believed that hemagglutinin is relevant to virus attachment and entry into the cell, while neuraminidase facilitates virus release (Frank 2002). Hemagglutinin also plays a central role in the process of immuno escape, in which the antibodies mainly attack five epitopes, denoted as epitopes

A–E, on the surface of the hemagglutinin protein (Muñoz and Deem 2005; Macken et al. 2001). Because of antigenic changes through time, influenza vaccines are redesigned each year to provide improved protection against evolved circulating strains. The efficacy of the annual vaccine is variable due to the escape mutation of the influenza virus (Smith et al. 1999), especially mutation at the five epitopes on the hemagglutinin (Gupta et al. 2006).

By analyzing the results of over 50 epidemiological studies of H3N2 influenza during the period 1968–2004 (Gupta et al. 2006; Sun and Deem 2009; Zhou et al. 2009; Pan and Deem 2009) showed that the escape mutation of influenza A virus can be measured by $p_{\mathrm{epitope}}$, the proportion of mutated amino acids in the dominant epitope of hemagglutinin, where the dominant epitope is defined as the epitope with the largest such proportion among the five epitopes. Compared with $p_{\mathrm{sequence}}$, the proportion of mutated amino acids in the whole sequence of hemagglutinin, and the ferret antisera assays, $p_{\mathrm{epitope}}$ between vaccine strains and dominant circulating strains in the same flu season correlated better with the vaccine efficacies in the northern hemisphere (Gupta et al. 2006). Therefore, $p_{\mathrm{epitope}}$ is an appropriate measurement for the antigenic distances between vaccine strains and dominant circulating strains. With the definition of the dominant epitope, the escape mutation at the dominant epitope induces the largest antigenic distance between vaccine strains and dominant strains, and endows the dominant epitope with the immunodominance.

In Gupta et al.'s (2006) model, which correlates well with vaccine efficacy in humans, every mutated amino acid is assigned the same weight. However, free energy calculations suggest that different amino acid substitutions have different contributions to the escape from the immune pressure. In general, the calculated differences in binding free energy $\Delta\Delta G = \Delta G_{\mathrm{mutated}} - \Delta G_{\mathrm{wildtype}}$ are different for different mutations, where $\Delta G_{\mathrm{mutated}}$ and $\Delta G_{\mathrm{wildtype}}$ denote binding free energy between two proteins one of which has and does not have a point mutation, respectively. For the experimentally measured difference in binding free energy $\Delta\Delta G$ between human growth hormone (hGH) and its first bound receptor (hGHbp), individual alanine substitutions of hydrophobic amino acids on the epitope of hGHbp induced the largest increase in $\Delta\Delta G$, followed by charged amino acids (Clackson and Wells 1995). Nevertheless, we show that charged amino acids correlate more strongly with viral evolution (Tables 3, 4). Nakajima et al. (2007) found that the majority of escape mutations of H3 hemagglutinin of the strain A/Kamata/14/91 were the mutations that introduce charged amino acids, and that the frequency of selected mutations to charged residues was significantly higher than that expected by random chance. A related study by Smith et al. (2004) found a similar over-representation of mutations to charged amino acids in the evolution of influenza.

We here consider the effect of different physical properties on the escape from immune pressure. We focus on the charged amino acids in the epitopes. These amino acids are strongly hydrophilic, and they reduce the tendency of antibodies to bind to hemagglutinin. Charged amino acids play a critical role in protein–protein interaction by creating salt bridges and salt bridge networks. Charged amino acids introduce specificity in binding (Sinha et al. 2002). Amino acid substitutions involving charged residues in the vicinity of receptor-binding region of HA affect the binding affinity between HA and its receptor (Kaverin et al. 2000). The evolution of charged amino acids, therefore, may provide useful information on viral escape from antibody pressure. A discussion of charge evolution in proteins has been given by Leunissen et al. (1990), who observed a large variance in the evolutionary trends among different protein families. Here we use stochastic methods to model the evolution of charged amino acids on the epitopes of H3 hemagglutinin strains collected from humans since 1968.

The metaanalysis of 50 epidemiological human vaccine efficacy studies shows that the single dominant epitope is the critical region that determines the epidemiological vaccine efficacy (Gupta et al. 2006). There are five non-overlapping epitopes on the surface of H3 HA molecule, namely epitope A–E, to which different sets of antibodies bind. In each epitope, the $p$ value is defined as the fraction of mutated amino acids (Gupta et al. 2006). The dominant epitope is defined as the epitope with the greatest $p$ value. The greatest $p$ value is $p_{\mathrm{epitope}}$. Epidemiological data on the vaccine efficacies in 18 previous flu seasons when H3N2 subtype was dominant were collected from approximately 50 studies (Gupta et al. 2006). The identities of the vaccine strains and dominant circulating strains were also obtained to calculate $p_{\mathrm{epitope}}$. H3N2 vaccine efficacy correlates with $p_{\mathrm{epitope}}$ with $R^2 = 0.81$. This strong correlation shows that $p_{\mathrm{epitope}}$ defined by the single dominant epitope is a quantitative definition of antigenic distance. Importantly, the $p_{\mathrm{epitope}}$ calculated from the dominant epitope correlated better with vaccine efficacy than did antigenic distance including all HA amino acids (Gupta et al. 2006).

The results of (Sun et al. 2006) show that subdominant epitopes are not the critical regions for vaccine efficacy. In an effort to improve the definition of antigenic distance, four modifications of the definition of antigenic distance were tested for their ability to improve the correlation with vaccine efficacy: (1) incorporating $p$ values from subdominant epitopes, (2) distinguishing conservative and nonconservative amino acids mutations, (3) mutations in amino acids adjacent to the epitopes, and (4) the calculation of mutations in neuraminidase (Sun et al. 2006). These four modifications of the definition of antigenic distance, including use of subdominant epitope $p$ values, all failed to substantially improve the correlation with vaccine efficacy data in the years 1971–2004. These results motivate our focus on the dominant epitiope in the present analysis.

The reduced alphabet Markov model (RAMM) described in this paper is an amino acid substitution model. It is built from the H3 hemagglutinin strains circulating in 1972–1987 when epitope B was the dominant epitope. This time span is shortly after the emergence of H3N2 virus in 1968, and this newly emerged virus subtype needed some time to adapt to host immune system, because emergence of new subtypes such as H2N2 in 1957 and H3N2 in 1968 went with Asian flu and Hong Kong flu outbreaks, indicating that subtypes like H3N2 were more virulent in the beginning and less adaptive to human. Further, the phylogenetic tree of H3N2 also shows that H3N2 evolved faster at the very beginning than in the later stage (Smith et al. 2004). So the pattern of evolution illustrates the escape mutation of the virus before substantial adaptation to host immune system was developed. Because the sequence database has been derived from patient samples and categorized by antigenic strain and date of collection, the human data do not necessarily reflect fixed variants, but, rather, snapshots along an evolutionary continuum.

Mutations of amino acids in different positions in the epitope are viewed as independent and identical Markov chains, whose parameters are the transition matrix $\mathbf{P}$ or the instantaneous rate matrix $\mathbf{Q}$. Markov models of protein evolution include the point accepted mutation (PAM) model (Dayhoff et al. 1978) and the block substitution matrix (BLOSUM) model (Henikoff and Henikoff 1992). These models are derived by counting mutations in aligned amino acid sequences, and this approach provides the transition matrices $\mathbf{P}(t)$ of a Markov chain in a period of evolutionary time $t$. Adachi and Hasegawa (1996) introduced the maximum likelihood method to estimate the elements of the transition matrix, and maximum likelihood was also employed to estimate the evolutionary time $t$ when fixing the transition matrix $\mathbf{P}(t)$ (Müller and Vingron 2000). The instantaneous rate matrix $\mathbf{Q}$ was calculated from the Laplace transform of $\mathbf{P}(t)$ (Müller and Vingron 2000; Müller et al. 2002). For a review of applications to 2000, see (Thorne 2000). Some recent studies estimated the effect of possible multiple mutations at the same position within evolutionary time $t$ (Veerassamy et al. 2003; Kosiol and Goldman 2005). The instantaneous rate matrix has been estimated from observed frequencies of 20 amino acids (Goldman and Whelan 2002).

The transition matrices $\mathbf{P}$ and the instantaneous rate matrices $\mathbf{Q}$ of most previous models are $20 \times 20$ matrices trained by analyzing databases with numerous alleles in many taxa. In the present case, the training data are limited, which can cause overfitting and introduce large errors to the fit model. One way to circumvent this difficulty is to decrease the number of parameters: 20 amino acids may be classified into several groups with similar biophysical properties. Here we grouped 20 amino acids as 5 charged amino acids and 15 uncharged amino acids.

To test whether the mathematical models can be applied to data outside of the existing human data, we investigated their relevance to animal models of influenza infection. Guinea pigs have been shown to be infected with unadapted H3N2 strains. In most cases, the infection is limited to the upper respiratory tract, causes little apparent morbidity, and can be spread to cage mates via aerosol (Lowen et al. 2006). As part of our analysis of the $p_{\text{epitope}}$ calculations, we developed additional evolutionary data derived through analysis of progeny viruses in animal model systems. Interestingly, the pattern of variations in the Guinea pig infection correlated well with the predictions of the model and with the human evolutionary data.

## Materials and Methods

We defined the charged amino acids as {Asp, Glu, Arg, Lys, His}. Historical sequence data indicated that the number of charged residues increased continually on the dominant epitope of the hemagglutinin of the circulating H3N2 strains in Fig. 1, except for the data in the year 1996–1997 which was presumably a continuation of the decreasing trend in the charge of epitope B since 1993. The vaccine strains in Fig. 2 also had increasing numbers of charged amino acids in a period of time without epitope shift. The charge in the dominant epitope sometimes stopped increasing, which means the charge saturated. Moreover, since the observed dominant epitope in a given year was either epitope A or epitope B, we defined the subdominant epitope as the epitope in the set {epitope A, epitope B} other than the dominant one. Figures 1 and 2 also show that the number of charged residues decreased in the subdominant epitope in both the circulating strains and the vaccine strains. For vaccine strains and circulating strains collected in 19 years, vaccine efficacies in each year had a strong correlation ($R^2 = 0.81$) with the $p_{\text{epitope}}$ in that year (Gupta et al. 2006), but $R^2$ decreased quickly if we use the weighted sum of $p_{\text{epitope}}$ and $p_{\text{epitope}}$ in previous years (Pan and Deem, unpublished data). Therefore we treated the immune escape of the virus as a Markov process (Minin and Suchard 2008). Additionally, since there is limited information on the correlation between different positions on a given epitope, we assumed that the mutations at different positions on the epitope followed identical and independent Markov chains.

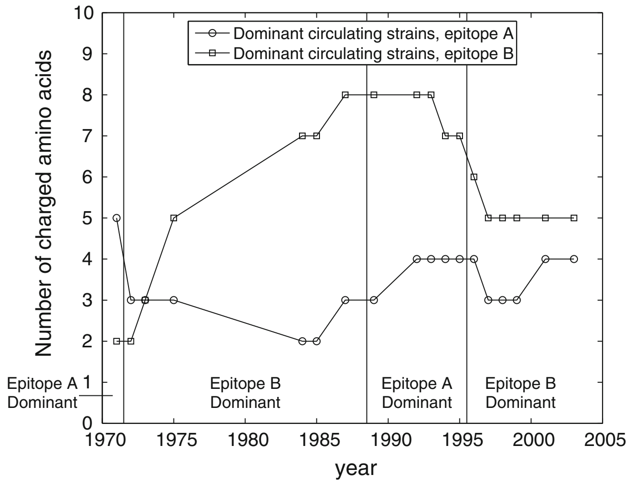


**Fig. 1** Number of charged amino acids for each year on the epitope A and epitope B of the dominant circulating strains from 1971 to 2003

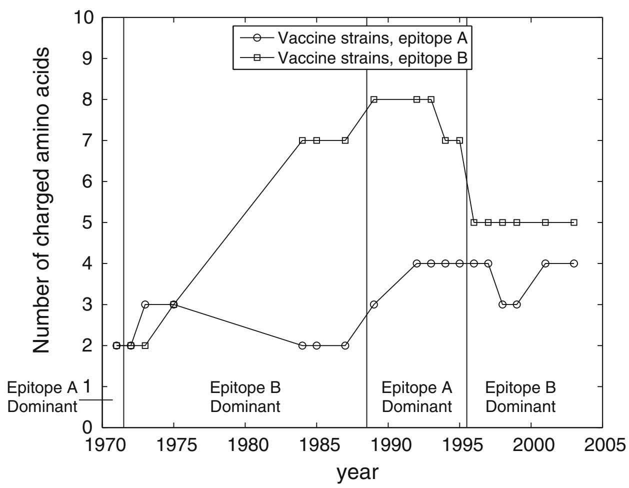


**Fig. 2** Number of charged amino acids for each year on the epitope A and epitope B of the vaccine strains from 1971 to 2003. There were four consecutive intervals where epitope A or epitope B was dominant

The epitope A and epitope B of the circulating strains and the strains deposited in the GenBank database were collected in 18 years between 1968 and 2003 and listed in Table 1, where the numbers of charged amino acids are presented as number for the dominant circulating strain and the average number for the circulating strains in GenBank database in the same year (Gupta et al. 2006). These data were also plotted (Figs. 1, 3). The dominant epitope was epitope A in the 1971 strain, epitope B in the {1972, 1973, 1975, 1984, 1985, 1987} strains, epitope A in the {1989, 1992, 1993, 1994, 1995} strains, and epitope B in the {1996, 1997, 1998, 1999, 2001, 2003} strains. All the strains except the 1971 strain fell into three multi-year time intervals defined by unchanged dominant epitopes. The numbers of charged amino acids on the dominant epitopes of each strain were counted and utilized. The total number of charged amino acid on a certain epitope, $N_c$, was modeled by a binomial distribution with probability

$$\Pr\{N_c = n\} = \binom{N}{n} P_c^n(t \mid \epsilon, \delta_1)[1 - P_c(t \mid \epsilon, \delta_1)]^{N-n} \quad (1)$$

with the mean number of amino acids in the epitope equal to $NP_c(t \mid \epsilon, \delta_1)$, where $N$ was the total number of amino acids on the epitope.

## Discrete-Time Markov Chain

We first applied a discrete-time Markov chain using 1 year as the time unit because the available data are the strains deposited in databases with 1-year time resolution. The probability distribution was defined as $\boldsymbol{\pi}(t) = (P_c(t), P_u(t)) = (P_c(t), 1 - P_c(t))$ with the initial value $(P_{c0}, 1 - P_{c0})$, where $P_c$ and $P_u$ were the probabilities that a charged amino acid and an uncharged amino acid existed in a given position on the dominant epitope. Following this definition, the transition matrix is a $2 \times 2$ matrix, taking parameters that describes the evolutionary process. We chose the mutation probability for each position, $\epsilon$, and the bias for mutation to charged amino acids from charged or uncharged amino acids, $\delta_1$ and $\delta_2$, respectively, to characterize the virus evolution model discussed in this paper. Thus, the transition matrix is

$$\mathbf{P} = (1 - \epsilon)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \epsilon \begin{pmatrix} \alpha_{cc} & \alpha_{cu} \\ \alpha_{uc} & \alpha_{uu} \end{pmatrix} \quad (2)$$

with $\alpha_{cc} = \frac{5}{20} + \delta_1, \alpha_{cu} = \frac{15}{20} - \delta_1, \alpha_{uc} = \frac{5}{20} + \delta_2, \alpha_{uu} = \frac{15}{20} - \delta_2$, where $\epsilon$ is the amino acid substitution rate (the new amino acid could be identical to the original amino acid); and $\alpha_{cc}, \alpha_{cu}, \alpha_{uc}$, and $\alpha_{uu}$ are the conditional probabilities that a charged residue mutated to a charged residue, a charged residue mutated to an uncharged residue, an uncharged residue mutated to a charged residue, and an uncharged residue mutated to an uncharged residue, respectively, if the substitution occurred. The parameters $\delta_1$ and $\delta_2$ were the bias for the probability that a charged amino acid mutated to a charged amino acid, an uncharged amino acid mutated to a charged amino acid, respectively. Under the further assumption that $\delta_1 = \delta_2 = \delta$, the probability distribution in year $t$ is

$$\begin{aligned}(P_c(t \mid \epsilon, \delta), P_u(t \mid \epsilon, \delta)) &= (P_{c0}, 1 - P_{c0})\mathbf{P}^t \\ &= (P_{c0}, 1 - P_{c0})\mathbf{Q}^{-1}\mathbf{D}^t\mathbf{Q} \\ &= \Big(\frac{1}{4}(1 + 4\delta - (1 - \epsilon)^t \\ &\quad (1 + 4\delta - 4P_{c0})), 1 - P_c(t \mid \epsilon, \delta)\Big) \end{aligned} \quad (3)$$

where the matrix $\mathbf{Q}$ was invertible and the matrix $\mathbf{D}$ was diagonal.

**Table 1** Epitopes A and B of dominant circulating strains

| Year | Dominant circulating strain | Dominant epitope | Number of charged amino acids in epitope A | Sequence of epitope A of the dominant circulating strain | Number of charged amino acids in epitope B | Sequence of epitope B of the dominant circulating strain |
|---|---|---|---|---|---|---|
| 1971–1972 | HongKong/1/68 (AF201874) | A | 5/2.22 | TGTVTQDGNAKGPGKRRNM | 2/2.78 | TGTKSGSTVNSTNQETSLVQA |
| 1972–1973 | England/42/72 (AF201875) | B | 3/2.95 | NGTVTQNGNAKGPDSSRNM | 2/3.00 | TGYKSECTVNSTNQVTSLVQA |
| 1973–1974 | PortChalmers/1/73 (AF092062) | B | 3/3.00 | NGTVTQNGNAKGPDSGRNM | 3/3.00 | TGYKSGSAVNSTDQETNLVQA |
| 1975–1976 | Victoria/3/75 (ISDNVIC75) | B | 3/3.18 | NGNVTQNGSAKGPDNGRNM | 5/4.18 | TGYKLGSTVNSTDKETDLVQA |
| 1984–1985 | Mississippi/1/85 (AF008893) | B | 2/2.00 | NGNVTQSGYAKGSVNSRNM | 7/6.88 | TGYKSESKANSTDKETNLVRA |
| 1985–1986 | Mississippi/1/85 (AF008893) | B | 2/2.13 | NGNVTQSGYAKGSVNSRNM | 7/7.03 | TGYKSESKANSTDKETNLVRA |
| 1987–1988 | Shanghai/11/87 (AF008886) | B | 3/3.20 | NDNVTQSGYAKGSVNSRNM | 8/8.20 | TGHESEYKANSTDRETNLVRA |
| 1989–1990 | England/427/88 (AF204238) | A | 3/3.42 | NDNVAQSGCAKGSVNSRNM | 8/8.20 | TGHESEYKANSTDRETNLVRA |
| 1992–1993 | Beijing/32/92 (AF008812) | A | 4/4.64 | NDNVAQDGYAKGSVNSRNM | 8/7.79 | TGHKSEYKANSTDRDTSLVRA |
| 1993–1994 | Beijing/32/92 (AF008812) | A | 4/4.77 | NDNVAQDGYAKGSVNSRNM | 8/7.02 | TGHKSEYKANSTDRDTSLVRA |
| 1994–1995 | Johannesburg/33/94 (AF008774) | A | 4/4.71 | NNNVAQDKYAKGSVNSRNM | 7/6.94 | TGHKLEYKANSTDSDTSLVRA |
| 1995–1996 | Johannesburg/33/94 (AF008774) | A | 4/4.35 | NNNVAQDKYAKGSVNSRNM | 7/6.36 | TGHKLEYKANSTDSDTSLVRA |
| 1996–1997 | Wuhan/359/95 (AF008722) | B | 4/4.01 | NGNVAQDTYAKGSVKSRNM | 6/6.01 | TGHKLEYKANSTDSDTSIVQA |
| 1997–1998 | Sydney/5/97 (AJ311466) | B | 3/4.07 | NSNVAQNTYAKSSIKSRNM | 5/5.53 | TGHQLKYKANSTDSDTSIAQA |
| 1998–1999 | Sydney/5/97 (AJ311466) | B | 3/4.03 | NSNVAQNTYAKSSIKSRNM | 5/5.18 | TGHQLKYKANSTDSDTSIAQA |
| 1999–2000 | Sydney/5/97 (AJ311466) | B | 3/3.95 | NSNVAQNTYAKSSIKSRNM | 5/5.01 | TGHQLKYKANSTDSDTSIAQA |
| 2001–2002 | Panama/2007/99 (ISDNCDA001) | B | 4/4.70 | NSNVAQNTSAKRSNKSRNM | 5/5.02 | TGHQLKYKANSTDSDISLAQA |
| 2003–2004 | Fujian/411/2002 (ISDN38157) | B | 4/4.53 | NSNVTQNTSAKRSNKSRNM | 5/4.98 | TGTHLKYKANGTDSDISLAQA |

Two numbers of charged amino acids in each epitope in each year are presented; the first one is calculated from the dominant circulating strain, and the second one is the simple arithmetic average of all strains collected in that year and deposited in GenBank

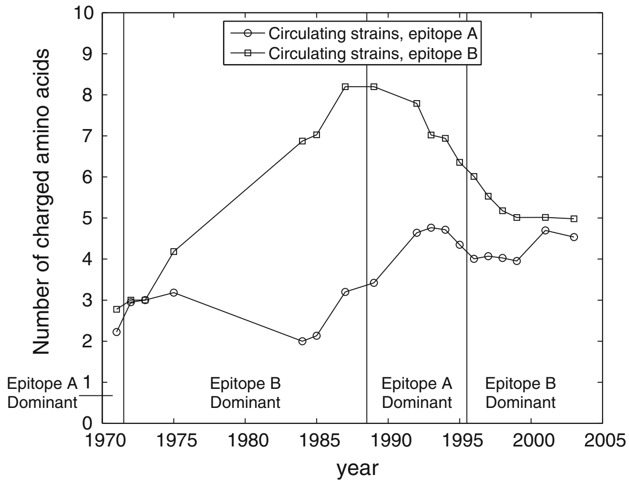


**Fig. 3** Average number of charged amino acids for each year on the epitope A and epitope B of the strains deposited in the GenBank database from 1971 to 2003. The curves in this figure are smoother than those in Fig. 1 due to the averaging over database strains. The difference in the numbers of charged amino acids between this figure and Fig. 1 is smaller than one for most years

## Continuous-Time Markov Chain

Besides the discrete-time Markov chain, the continuous-time Markov chain was also commonly employed in several bio-related fields including the estimation of residue mutation rates (Tseng and Liang 2006). In the continuous-time Markov chain, the transition rate matrix $\mathbf{Q}$ is first defined describing the evolution of probability distribution in a infinitesimal time interval, with $\mathbf{Q1} = 0$, where $\mathbf{1}$ is a column vector with all the elements equal to unity. The transition matrix is

$$\mathbf{P}(t) = \begin{pmatrix} P_{cc}(t) & P_{cu}(t) \\ P_{uc}(t) & P_{uu}(t) \end{pmatrix} = \exp(\mathbf{Q}t) \tag{4}$$

with the probability distribution at year $t$ was again defined as $\boldsymbol{\pi}(t) = (P_c(t), P_u(t)) = \boldsymbol{\pi}(0)\mathbf{P}(t)$ with the initial probability distribution $\boldsymbol{\pi}(0) = (P_{c0}, 1 - P_{c0})$. And if the Markov chain contains finite states, the Kolmogorov backward equation (KBE)

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{QP}(t) \quad \text{(KBE)} \tag{5}$$

holds.

For the evolution process of charged amino acid, the $\mathbf{Q}$ matrix was defined as

$$\mathbf{Q} = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix} \quad \lambda_1, \lambda_2 \geq 0. \tag{6}$$

The solution for the transition matrix $\exp(\mathbf{Q}t)$ was

$$P_{cc}(t) = \frac{\lambda_2}{\lambda_1 + \lambda_2} + \frac{\lambda_1}{\lambda_1 + \lambda_2} e^{-(\lambda_1+\lambda_2)t}$$
$$P_{cu}(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2} - \frac{\lambda_1}{\lambda_1 + \lambda_2} e^{-(\lambda_1+\lambda_2)t}$$
$$P_{uc}(t) = \frac{\lambda_2}{\lambda_1 + \lambda_2} - \frac{\lambda_2}{\lambda_1 + \lambda_2} e^{-(\lambda_1+\lambda_2)t}$$
$$P_{uu}(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2} + \frac{\lambda_2}{\lambda_1 + \lambda_2} e^{-(\lambda_1+\lambda_2)t}.$$

The term $P_c$ is given by

$$P_c(t \mid \lambda_1, \lambda_2, P_{c0}) = \frac{\lambda_2}{\lambda_1 + \lambda_2} + \left(P_{c0} - \frac{\lambda_2}{\lambda_1 + \lambda_2}\right) e^{-(\lambda_1+\lambda_2)t}. \tag{7}$$

## Maximum Likelihood Estimation

The maximum likelihood estimation method optimizes the parameters in a given parametric form by maximizing the log-likelihood function with the definition $l(\boldsymbol{\theta} \mid \mathbf{x}) = \ln P(\mathbf{x} \mid \boldsymbol{\theta})$. The observed data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ are usually independent, therefore the maximum likelihood function is $l(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{k=1}^{m} \ln P(\mathbf{x}_k \mid \boldsymbol{\theta})$. In our model, the data $\mathbf{x}_k = (n, t)$ in different years were assumed to be independent, where $n$ was the number of charged amino acids, $t$ was the time in years, and the parameter set was $\boldsymbol{\theta} = (\epsilon, \delta, P_{c0})$. From the relationship between the total number of charged amino acids and the probability distribution at each position (1), the maximum likelihood equation was written as

$$\begin{aligned} l(\epsilon, \delta, P_{c0} \mid \mathbf{x}) &= \sum_{k=1}^{m} \ln P(\mathbf{x}_k \mid \boldsymbol{\theta}) \\ &= \sum_{k=1}^{m} \ln \binom{N}{n_k} + \sum_{k=1}^{m} n \ln P_c(t_k \mid \epsilon, \delta, P_{c0}) \\ &\quad + \sum_{k=1}^{m} (N - n) \ln[1 - P_c(t_k \mid \epsilon, \delta, P_{c0})] \end{aligned} \tag{8}$$

with the expression of $P_c$ in (3).

The maximum likelihood function was maximized by solving the following equations

$$\begin{cases} \frac{\partial}{\partial \epsilon} l(\epsilon, \delta, P_{c0} \mid \mathbf{x}) = 0 \\ \frac{\partial}{\partial \delta} l(\epsilon, \delta, P_{c0} \mid \mathbf{x}) = 0 \\ \frac{\partial}{\partial P_0} l(\epsilon, \delta, P_{c0} \mid \mathbf{x}) = 0 \end{cases} \tag{9}$$

The numerical solution was identical to that of the least square fit to the first two significant figures. We note that solutions of (9) were not ill-conditioned while the nonlinear fitting algorithm for (3) led to ill-conditioned Jacobians.

### Guinea Pig Animal Model

The Guinea pig model of human influenza infection was adopted to model the process of infection and reinfection (Lowen et al. 2006). A sample of A/Wyoming/2003 (H3N2) was obtained from the Centers for Disease Control and Prevention (CDC). Sequence analysis of multiple plaque isolates from the CDC virus demonstrated that the sample contained a mixture of viruses with several hemagglutinin (HA) sequences, all in close agreement with the A/Wyoming/03/2003 sequence reposited in GenBank (accession number AAT08000). An isolate (WyB4) representing the dominant HA sequence contained within the CDC virus mixture was purified and propagated.

In the first infection experiment, four immunologically naïve Guinea pigs were inoculated with the CDC virus mixture to model virus evolution in the absence of robust immunity. Three days following inoculation, progeny virus were purified from nasal washes for sequence analysis of the HA genes. After a recovery period of 28 days, two of the animals were reinfected with the CDC virus mixture to model virus escape in the presence of increased immune pressure. Nasal washes were collected after 3 days, progeny virus isolated, and HA gene sequences analyzed. In a second infection experiment, six naïve Guinea pigs were inoculated with the purified WyB4 isolate. Again, virus progeny were purified from nasal wash samples for HA sequence analysis. In a third infection experiment, Guinea pigs were inoculated with three immunizations of recombinant Wyoming HA protein having the same sequence as the WyB4 isolate. After the animals had mounted robust immune responses, they were challenged with either WyB4 or the CDC virus mixture. Progeny virus were isolated from nasal washes for HA gene sequence analyses.

## Results

### Charge Increases in the Dominant Epitope

The parameters obtained by the fitting algorithm for the first time interval (epitope B was dominant, from 1972 to 1987) were $\boldsymbol{\theta}^{\mathrm{T}} = (\epsilon, \delta, P_{\mathrm{c0}}) = (0.207, 0.113, 0.0937)$ with the square of the correlation coefficient between observations and the model $R^2 = 0.98$. The mathematical details are given in Supplementary Material. Here the transition matrix was

$$\mathbf{P} = \begin{pmatrix} 0.868 & 0.132 \\ 0.075 & 0.925 \end{pmatrix}. \tag{10}$$

We plotted the observed data and the data predicted by the model with the parameters fixed by maximum likelihood estimation in Fig. 4. We fit $\epsilon$ and $\delta$ to the data in the range 1972–1987 and used these values for all years. We fit the value of $P_{\mathrm{c0}}$ in the range 1972–1987 and found it equal to the data point at 1972. For other intervals we used the observed initial values.

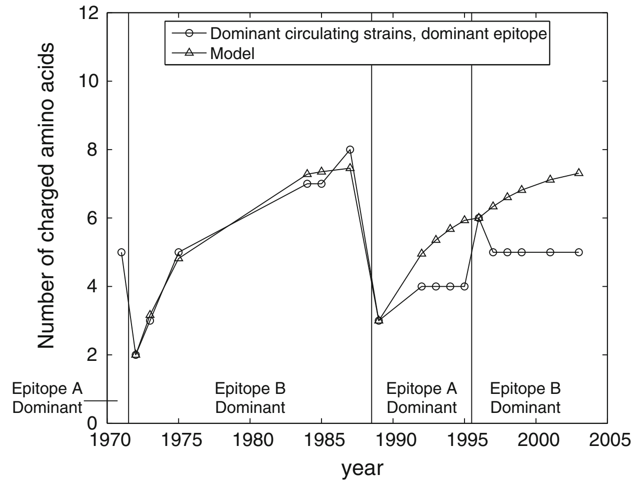


**Fig. 4** Number of charged amino acids for each year on the dominant epitopes of the circulating strains. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 are plotted. Since the estimated $P_{\mathrm{c0}}$ was close to the observed number, we used the observed numbers of charged amino acids in the first year of the interval to calculate $P_{\mathrm{c0}}$

The model parameters here yielded the average number of point mutations in epitope $\mathrm{B} = 19N\epsilon/20 = 4.07$, and the observed numbers of point mutation in epitope B during 1 year were: 1971–1972, 4; 1972–1973, 6; 1984–1985, 0. Those are all the available 1-year time spans in the training data, and this model falls inside the standard error of the actual annual mutation rates on the dominant epitopes. That $\delta = 0.11$ showed that the conditional probability that a charged residue mutated to a charged residue was 0.36 while the probability that an uncharged residue mutated to an uncharged residue was 0.64. The observed number of charged amino acids on epitope B in 1972 was 2, and corresponding $P_{\mathrm{c0}} = 2/21 \approx 0.095$, which agrees well with the fit value $P_{\mathrm{c0}} = 0.094$.

We also obtained all the strains deposited in the GenBank database that were collected in the same year as a circulating strain existed. The numbers of charged amino acids $N_{\mathrm{c}} = NP_{\mathrm{c}}(t \mid \epsilon, \delta, P_{\mathrm{c0}})$ were counted for each year as the mean value of the numbers of charged amino acid on the dominant epitope of the strains collected in that year. The standard deviations of the numbers of charged amino acid on the dominant epitope were also calculated for each year. Again, there is no assurance that any of these mutations can be considered end-point or fixed variants as the evolutionary process continued from year to year. By using these $P_{\mathrm{c}}(t \mid \epsilon, \delta, P_{\mathrm{c0}})$ to fit the parameters $\boldsymbol{\theta} = (\epsilon, \delta, P_{\mathrm{c0}})$, $\epsilon$

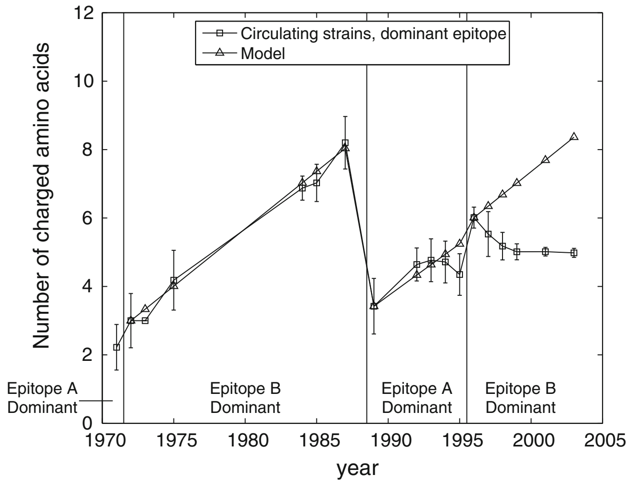


**Fig. 5** Number of charged amino acids for each year on the dominant epitopes of the average of all the database strains in the same year. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 were plotted. We used the observed numbers of charged amino acids in the first year of the interval to calculate $P_{c0}$. Error bars for the circulating strains are one standard deviation calculated from all NCBI strains collected in that year

approached to zero when the number of iterations of the fitting algorithm increased to infinity, while the product $\epsilon\delta$ approached to 0.016 compared with the product $\epsilon\delta = 0.023$ in the case with only circulating strains. Both the calculated and observed $P_{c0}$ equal to 0.14. The model fit well with all the datapoints in this time span with $R^2 = 0.99$. We note that the Markov model is not able to reproduce convex and rising data, and this is the reason for $\epsilon \to 0$ for these data. In fact the data in Fig. 5 are convex probably because the simple arithmetic averaging is only a rough approximation, and we expect concave and rising data, which the Markov model can represent, from a properly weighted population average. The maximum likelihood method estimated the same parameter values to within two significant figures as the least square fitting method.

The continuous-time Markov chain possessed the similar parameter set $\boldsymbol{\theta} = (\lambda_1, \lambda_2, P_{c0})$ to the discrete-time Markov chain. For the circulating strains in 1972–1987, the nonlinear least-squares fitting yielded the parameters $\lambda_1 = 0.15, \lambda_2 = 0.084, P_{c0} = 0.094$ with $R^2 = 0.98$. For the database strains, there were $\lambda_1 = -0.031, \lambda_2 = 0.011, P_{c0} = 0.14$ with $R^2 = 0.99$. As before, simple arithmetic averaging produces convex and rising data, which the Markov model is not able to represent, and which is not expected from a population average.

The RAMM model fit well the data in circulating strains and those in the GenBank database in the first time interval (1972–1987). This model had a larger discrepancy in the prediction versus both groups of data in the second (1989–1995) and third time interval (1996–2003).

## Evolution of Amino Acids in Epitopes of Hemagglutinin Is Non-Universal: Comparison with PAM Matrix

To compare with the RAMM model, we also employed the conventional PAM matrix (Dayhoff et al. 1978) to predict the evolution of charged amino acids. We first calculated the PAM1 matrix $\mathbf{M} = \{M_{ij}\}$ describing the evolution of one amino acid position during a time unit in which the probability that a point mutation occurred in this position was fixed by Dayhoff et al. (1978) to 0.01 (the meaning of the suffix 1). Evolution during $n$ time units was depicted by PAM$n$ matrix with PAM$n$ = PAM1$^n$ = $\mathbf{M}^n = \{M'_{ij}\}$. The probability for a point mutation occur was then

$$P = \sum_{i=1}^{20} (1 - M'_{ii})\pi_i. \tag{11}$$

We let this probability equal to the annual mutation rate calculated by the RAMM model and verified by the historical data, $19\epsilon/20 = 0.196$, and so PAM22 should be selected to calculate the mutations during 1 year. The more frequently used PAM20 matrix, however, underpredicts the annual mutation rate and generates larger errors. While the PAM22 matrix reproduces the observed number of total mutations in the dominant epitope, as we shall see, it underpredicts the number of charged mutations per year. This means there is an additional selective force for charged amino acids in the epitopes of hemagglutinin relative to protein evolution in general. To calculate the $2 \times 2$ transition matrix with the definition

$$\mathbf{P} = \begin{pmatrix} P_{cc} & P_{cu} \\ P_{uc} & P_{uu} \end{pmatrix}$$

there were

$$P_{cu} = P\{\text{charged} \to \text{uncharged} \mid \text{charged}\} = \frac{\sum_{i=\text{charged}} \left( \sum_{j=\text{uncharged}} M'_{ij} \right) \pi_i}{\sum_{i=\text{charged}} \pi_i} \tag{12}$$

$$P_{uc} = \frac{\sum_{i=\text{uncharged}} \left( \sum_{j=\text{charged}} M'_{ij} \right) \pi_i}{\sum_{i=\text{uncharged}} \pi_i} \tag{13}$$

$$P_{cc} = 1 - P_{cu} \tag{14}$$

$$P_{uu} = 1 - P_{uc}. \tag{15}$$

yielding the solution of the transition matrix $\mathbf{P}$ based on PAM22 matrix

$$\mathbf{P} = \begin{pmatrix} 0.883 & 0.117 \\ 0.040 & 0.960 \end{pmatrix} \tag{16}$$

which is slightly different from RAMM matrix (10). The evolution predicted by PAM22 matrix for the circulating

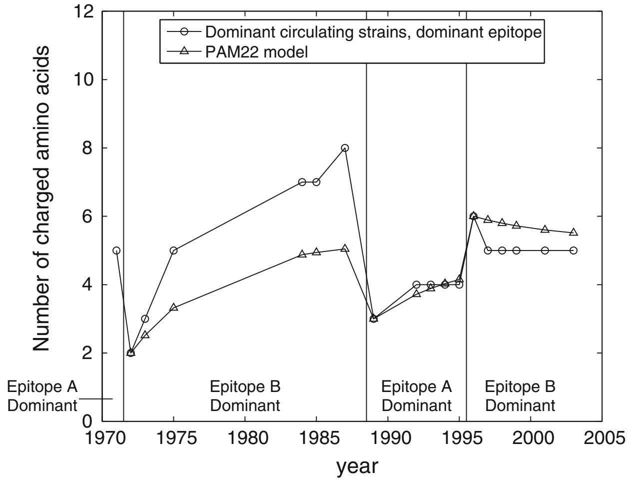


**Fig. 6** Number of charged amino acids for each year on the dominant epitopes of the circulating strains. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 with PAM22 matrix were plotted. $P_{c0}$ was fixed to the observed numbers of charged amino acids in the first year of the interval. The PAM22 matrix cannot reproduce the data

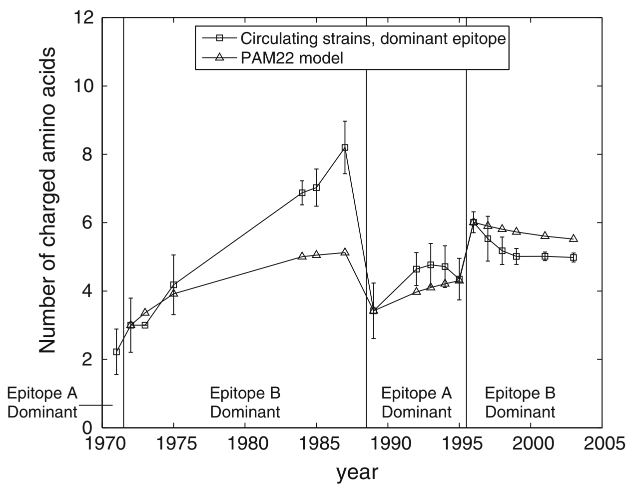


**Fig. 7** Number of charged amino acids for each year on the dominant epitopes of the average of all the database strains in the same year. Both the observed data from 1971 to 2003 and the predicted data from 1972 to 2003 with PAM22 matrix were plotted. $P_{c0}$ was fixed to the observed numbers of charged amino acids in the first year of the interval. The PAM22 matrix cannot reproduce the data. Error bars for the circulating strains are one standard deviation calculated from all NCBI strains collected in that year

strains and equally weighted database strains is plotted in Figs. 6 and 7, respectively.

## Guinea Pig Animal Model Verifies the Increase in Charge

To buttress our analyses of the evolution of charged amino acids in historical virus sequences, we examined the frequency of charge changes in the HA sequences derived from the progeny virus using the Guinea pig animal infection model. The animals were divided into two groups, naïve and immune. The immune animals were either previously infected or immunized with HA protein. The animals were given either an inoculum of a dominate strain with the frequency 64.3%, as well as three closely related variants with the frequencies 14.3, 14.3, and 7.1% respectively, totaling 35.7%, denoted CDC, or an inoculum of the dominant plaque-purified strain of the CDC stock mentioned above, denoted WyB4. The HA genes from replicated virus were sequenced from nasal washing samples at three days following intra-nasal inoculation. As observed in Table 2, naïve animals given the CDC inoculum resulted in an increase in the frequency of variants in the nasal wash virus. The WT:mutant ratio of 64.3%:35.7% in the CDC inoculum was reduced to 48.3%:51.7% in the progeny virus, indicating that the dominant viral strain in the starting material was reduced 16.0% while new HA mutants increased 16.0%. Moreover, most of the progeny variants characterized in the nasal wash samples were not previously detected in the CDC stock virus suggesting that the progeny variants arose during infection in the Guinea pig. The Guinea pig animal model also yielded new strains with multiple mutations as well as new strains which fundamentally changed the biochemical character of the mutated amino acid. Although many of the new variants may came from neutral mutation, the scenario that some of the new variants increased the replicative fitness cannot be precluded. Thus, it seems likely that the egg-adapted variants in the CDC stock had reduced replicative fitness relative to the WyB4 strain. Unexpectedly, reinfection of the same Guinea pigs with the CDC stock resulted in an increase of 11.3% in the frequency of the wildtype HA sequence and a concomitant equal decrease in the frequency of progeny strains having mutated HA genes.

**Table 2** Sequence analysis of progeny virus isolated from nasal washes of infected Guinea pigs

| Virus | Immune status | Number of sequences | Wild type | Mutants |
|---|---|---|---|---|
| CDC | Naïve | 58 | 28 (48.3%) | 30 (51.7%) |
| CDC | Reinfected | 82 | 62 (75.6%) | 20 (24.4%) |
| WyB4 | Naïve | 62 | 43 (69.4%) | 19 (30.6%) |
| WyB4 | HA-immunized | 210 | 159 (75.7%) | 51 (24.3%) |

CDC virus designates mixture of Wyoming HA sequence variants contained in initial virus stock. WyB4 virus designates purified stock from predominant isolate of CDC virus. Immune status refers to whether the animals were naïve, i.e., neither previously infected nor immunized with purified HA protein, or immune. Number of sequences refers to the number of HA genes examined in progeny virus isolated from nasal washes

In the second set of experiments in which naïve Guinea pigs were infected with the WyB4 plaque-purified clone of the dominant strain in the CDC stock, progeny virus had a 69.4%:30.6% WT:mutant HA gene ratio relatively close to that seen with the CDC stock in naïve animals. In the third experiments in which immunized animals were infected with the WyB4 virus, the ratio of WT:mutant HA genes increased to 75.7%:24.3%, similar to the result observed in infections with the CDC stock. Taken together, infection of naïve animals resulted in a frequency of WT HA genes of approximately 65–70% while infection of immune animals produced an increase in the WT frequency to $\sim$76% regardless of whether the CDC stock or WyB4 virus was used. By comparing the percentages of mutants listed in Table 2, three pairs of numbers are statistically significantly different ($p < 0.02$), which are CDC naïve vs. CDC reinfected, CDC naïve vs. WyB4 naïve, and CDC naïve vs. WyB4 HA-immunized, respectively. The null hypothesis cannot be rejected for the other three pairs when testing the statistically significant difference ($p > 0.3$). That is, only the low WT ratio in CDC naïve is statistically significant. Presumably the immunized animals eliminated the virus more quickly, and so fewer mutations were formed.

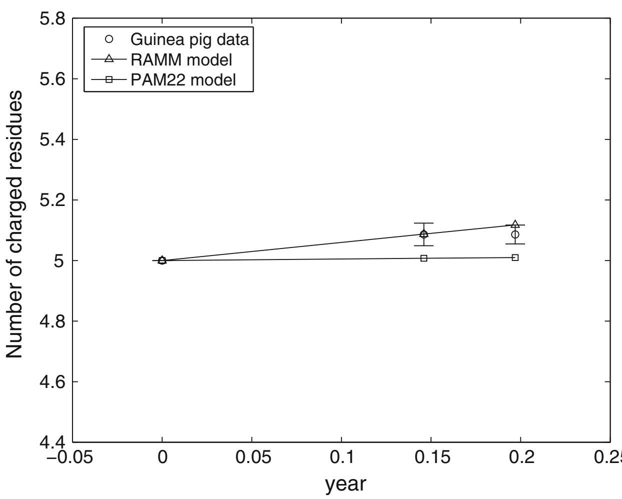


**Fig. 8** Comparison of charged residue changes between theoretical models and sequence data derived from Guinea pigs inoculated with the CDC A/Wyoming/2003 virus mixture. The RAMM and the PAM theoretical models were considered, as well as three data points from the Guinea pig infections: First point: Wyoming inoculum; Second point: progeny strains from infection of naïve Guinea pigs; and Third point: progeny strains from infection of previously infected Guinea pigs. The time of naïve and reinfection strains was estimated from the average number of amino acid mutations, counting both wild type and mutated strains, in the whole HA1 sequence. With the assumption derived from historical data that the annual mutation rate was 5.2 amino acids/HA1 sequence/year, we divided those average numbers of mutations by 5.2 to obtain the times. Error bars are one standard error

Figure 8 compares the number of charged residues predicted by both the PAM22 model and the reduced alphabet Markov model (RAMM), with those observed in the progeny of Guinea pigs infected with the A/Wyoming/2003 mixture obtained from the CDC. Because epitope B was the dominant epitope in the Wyoming virus circulating in 2003, we focused our analysis on the residues contained within this epitope. Five charged residues were present in the 21 amino acid B epitope in the initial Wyoming inoculum. The mean number of charged residues increased to 5.09 in the progeny from naïve animals and 5.09 in the progeny of reinfected animals.

The evolutionary times between the strains from the original inocula, the progeny of naïve Guinea pigs, and the progeny of reinfected animals are estimated by the mean number of mutations in the whole sequence of strains collected from Guinea pigs, divided by the approximate annual historical number of mutations in the whole HA1 sequences (mean historical mutations during the years 1971–2003 = 5.2 amino acids/HA1 sequence/year). Note that regardless of whether charged amino acids are involved, the numbers of mutations in the whole sequence instead of epitope B are used to calibrate the evolutionary times. The strains from the naïve Guinea pig experiments were characterized, and the total number of point mutations in these strains was then divided by the number of strains to calculate the population average number of point mutations in the progeny strains. This average number was divided by 5.2, the mean number of historical mutations per year in the whole HA1 sequences, to obtain the evolutionary time $\Delta t$ for the naïve Guinea pig experiment. The $\Delta t$ for the reinfection experiment was calculated following the same method. The number of charged amino acids in epitope B, the dominant epitope for the Wyoming strain, increased in a trajectory that closely followed the path predicted by the RAMM model. In contrast the generic protein evolution PAM22 model predicted no such increase in charged residues. The evolutionary times and the numbers of charged amino acids in epitope B are calculated respectively by different approaches, therefore the evolutionary time is independent of the increase of charged amino acids. Thus, the agreement between RAMM model and the experiment is nontrivial.

Figure 9 shows a similar plot of charged residues observed from the experimental infection of naïve and immune Guinea pigs with the WyB4 isolate compared to predicted values. The methods used in the analysis shown in Fig. 8 were applied to the data derived from infection with the homogenous virus. The WyB4 appeared to accumulate charged residues in the epitope B somewhat more rapidly than the Wyoming mixture obtained from the CDC. These data suggest either that the WyB4 strain has a replicative advantage in Guinea pigs and therefore evolves

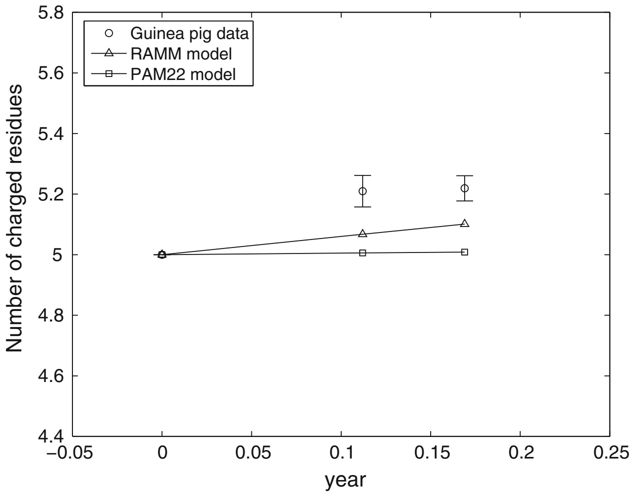


**Fig. 9** Comparison of charged residue changes between theoretical models and sequence data derived from progeny virus isolated from Guinea pigs inoculated with the homogeneous WyB4 virus isolate. The number of predicted and observed charged residues were analyzed with the method used for the data in Fig. 8. Error bars are one standard error

faster, or that accumulation of charged residues, especially in immune animals, represents a selective advantage.

Comparing Figs. 8 and 9, the WyB4 isolate appeared to evolve more in naïve animals than did the mixed CDC stock. A greater number of mutations per mutant strain were observed in naïve animals receiving the WyB4 isolate virus than the naïve animals receiving the CDC swarm. Among the 30 collected mutants of CDC naïve, the numbers of mutants with 1–4 point mutations were 22, 3, 4, and 1, respectively, with the average number of mutations per mutant equaled 1.47 (standard error: 0.16), and 15 of 44 (34.1%) mutations occurred in epitopes A–E. Among the 19 collected mutants of WyB4 naïve, the number of mutants with 1–4 point mutations were 6, 10, 2, and 1, respectively, with the average number of mutations per mutant equaled 1.89 (standard error: 0.19), and 21 of 36 (58.3%) mutations occurred in epitopes A–E. The difference of the average number of mutations per mutant of CDC naïve and WyB4 naïve is hence significant ($p = 0.039$). Again, the mechanisms behind these data are not clear, but they suggest the appearance of some form of heterotypic immunity in the CDC stock virus. That is, the immune system of the Guinea pig either put more pressure on the WyB4 strain to evolve or took longer to clear the WyB4 and so allowed it to evolve longer relative to the CDC stock virus. An estimation of evolutionary time from the total number of mutations in the HA protein suggests the explanation is more likely the former.

## Partitioning the Amino Acids by Charge is Optimal

The RAMM model discussed in this paper deals with the numbers of amino acids belonging to a certain category in different years. Besides the charged–uncharged categorization of the 20 amino acids, we considered six different categories, which were polar (Arg, Asn, Asp, Cys, Glu, Gln, His, Lys, Ser, Thr, Tyr), hydrophobic (Ala, Gly, Ile, Leu, Met, Phe, Pro, Trp, Val), basic (Arg, Lys, His), acidic (Asp, Glu), amides (Asn), aliphatic (Gly, Ala, Val, Leu, Ile), to generate different data sets for the RAMM model. Four combinations among these seven categories were also introduced with the amino acids in each category being counted once, which were charged + basic, charged + acidic, charged + polar, hydrophobic + aliphatic, respectively. To further generalize the $R^2$ value given by the RAMM model, different combinations of epitopes were focused on. Here the sequence of H3 hemagglutinin (HA1) between position 44 and 312 was split into six subsets: epitope A–E, and the amino acids outside any epitope, defined as O. Table 3 shows that charged amino acids are among the categories best fit by the RAMM model. Table 4 lists the $R^2$ values of the RAMM model for seven categories of amino acids that are counted in six combinations of epitopes containing epitope B, the dominant epitope in the year of 1972–1987. Compared with Table 3, Table 4 indicates that focusing on the dominant epitope leads to the approximately optimal fitting.

**Table 3** $R^2$ values for amino acid categories and epitope B, the dominant epitope, using dominant circulating strains in 1972–1987 from which the model is trained

| Categories | $R^2$ |
|---|---|
| Charged | 0.9810 |
| Polar | 0.9620 |
| Hydrophobic | 0.9586 |
| Basic | 0.9872 |
| Acidic | 0.8743 |
| Amides | 0.6217 |
| Aliphatic | 0.7698 |
| Combinations | |
| Charged + Basic | 0.9932 |
| Charged + Acidic | 0.9567 |
| Charged + Polar | 0.9568 |
| Hydrophobic + Aliphatic | 0.8528 |

Seven categories and four combinations of categories are presented here. These four combinations involving charged and hydrophobic amino acids are chosen as the supplement to the seven categories, because charged and hydrophobic amino acids, especially charged ones, are critical in protein–protein interaction and evolution (Clackson and Wells 1995; Nakajima et al. 2007; Sinha et al. 2002; Kaverin et al. 2000; Leunissen et al. 1990)

**Table 4** $R^2$ values for amino acid categories and combinations of epitopes involving epitope B, the dominant epitope, using dominant circulating strains in 1972–1987 from which the model is trained

| Epitopes | AB | BC | BD | BE | BO | ABCDEO |
|---|---|---|---|---|---|---|
| Charged | 0.8928 | 0.9782 | 0.9431 | 0.9913 | 0.9810 | 0.8935 |
| Polar | 0.8187 | 0.9620 | 0.7780 | 0.8528 | 0.9766 | 0.7656 |
| Hydrophobic | 0.9586 | 0.9771 | 0.7928 | 0.9580 | 0.9884 | 0.9766 |
| Basic | 0.9872 | 0.9872 | 0.9913 | 0.9456 | 0.9872 | 0.9917 |
| Acidic | 0.4730 | 0.9210 | 0.6337 | 0.7245 | 0.6838 | 0.3916 |
| Amides | 0.9771 | 0.7922 | 0.9834 | 0.2171 | 0.8443 | 0.9326 |
| Aliphatic | 0.4976 | 0.9009 | 0.7648 | 0.3787 | 0.7895 | 0.6129 |

Residues outside an epitope are denoted by O. The dominant circulating strains in the time span of 1972–1987 have epitope B as the dominant epitope. The H3N2 virus emerged in 1968, therefore less adaptation to host immune system was developed compared with other time span. The previous discussion indicates that epitope B had the immunodominance in this period of time, and this table shows the limited contribution of amino acids outside the dominant epitope to the pattern of evolution

## Discussion

### Data Fitting and Verifying the RAMM Model

For the historical data, our analysis was based on both the dominant circulating strain and the average of all circulating strains deposited in NCBI database. The frequency of database strains is not currently available. In addition, because clear sequential and temporal relationships between the sequences are unavailable, the fixation of mutations is unknown. Although some trends in the mutations have emerged (such as increased numbers of both charged residues in defined epitopes and total numbers of glycosylation sites throughout the HA ectodomain), the majority of the sequences do not appear to be fixed.

For the Guinea pig animal model, the frequency of both wildtype and variants were collected and tabulated. These data allowed us to calculate a population average of all the strains with no loss of frequency information when counting the average number of mutations and the increase of charged amino acids. For the Guinea pig experiment, we have sequences for all the strains characterized in all animals, and can thus calculate the true population average. However, at this time we do not have firm data on fixation rates of variants in the Guinea pig infection. Some of the observed variants could have been fixed over the 4–6 rounds of replication that occurred in the three-day infection, but a serial transmission experiment performed in immune animals would be required to provide quantitative data on mutations have become immunologically fixed. Thus, the quality of the Guinea pig data and the human database with respect to fixation of mutations is comparably similar.

Note the RAMM model calibrated to years of human epidemiological data fit well the Guinea pig experimental data. Figures 8 and 9 compare the observed evolution of the HA envelope in the CDC Wyoming mixed stock and the purified WyB4 isolate with the evolution predicted by the two models. As shown above, the model fit the in vivo data well for the CDC stock but underestimated the accumulation of charged residues observed in progeny from the WyB4 infections. CDC stock was an ensemble of a dominant and three closely related variants while WyB4, the plaque-purified isolate and the dominant member of the CDC stock was the virus that appeared to have greater replicative capacity in Guinea pigs as compared to the variants in the CDC stock. The CDC stock infection of naïve animals more closely represents a cross-section of a transmitting quasi-species related to all the circulating strains existing in 1 year, while the WyB4 infection of naïve animals may model the dominant circulating strain.

Table 4 provides further support to the hypothesis of a single dominant epitope. These data show the $R^2$ values of the correlations between observations and the model using numbers of charged amino acids in a variety of combinations of epitopes as well as a general null model using the whole sequence of H3 HA. Different amino acid categories are also incorporated in Table 4 to show the $R^2$ values of the correlation between observations and the model. Thus, as in previous studies (Sun et al. 2006), Table 4 confirms that calculations using the single dominant epitope lead to the approximately optimal fitting between the observations and the model. None of the inclusions of subdominant epitopes yields $R^2$ values much greater than that of focusing on the single dominant epitope.

The results obtained from the discrete-time and continuous-time Markov chains were essentially identical. By comparing expressions (3) and (7), the parameter set of the discrete-time Markov chain $(\epsilon, \delta, P_{c0})$ and that of the continuous-time Markov chain $(\lambda_1, \lambda_2, P_{c0})$ are related by

$$\begin{cases} \epsilon = 1 - e^{-(\lambda_1+\lambda_2)} \\ \delta = \frac{\lambda_2}{\lambda_1+\lambda_2} - \frac{1}{4} \end{cases} \tag{17}$$

The discrete-time Markov chain was appropriate for the case in this paper because the time unit for the data was year. Currently there is limited time resolution available, hence using discrete-time model with 1-year time resolution did not lose any information. The virus evolves continuously in different geographic regions, however, and spreads among these regions by the migration of humans and birds. Furthermore, the global virus strains in a given time point constitutes an ensemble where the population of different strains varies. Available data are the dominant circulating strains in each year, as well as the database strains. Therefore, the average numbers of charged amino acids should be the weighted sum of all the database strains in each year, rather than their simple arithmetic means. It will be of significant help to know the frequency of each database strain so that a population average can be calculated.

### Model Reversibility

Most previous models assumed the reversibility of the Markov chain, whose transition matrices $\mathbf{P} = \{P_{ij}\}$ and equilibrium probability distributions $\boldsymbol{\pi} = \{\pi_i\}$ satisfied the detailed balance equation

$$\pi_i P_{ij} = \pi_j P_{ji}. \quad (18)$$

Such models were mainly designed to describe the evolution of bacteria, archaea, and eukaryota living in a relative constant environment. The evolution of these species is probably dominated by a stable and loose natural selection that is tolerant to mildly deleterious polymorphism (McDonald 2006). Viruses, on the other hand, perform their biological function in other organisms, and usually their surface protein antigens are targeted by the immune systems in the antigen sequence space that makes the virus evolution under strong, directed, and changing selection. This arrow of time due to directed selection invalidates the reversibility assumption, because $\boldsymbol{\pi}$ is changing, and this higher degree of freedom must be accounted for when formulating the model. In the data shown here, the detailed balance condition (18) was not established, because the immune pressure and selection are constantly changing. In other words, when starting from a non-steady state initial condition, the system will first converge toward the steady state before fluctuating about that steady state until the immune pressure changes again. We studied the non-equilibrium dynamics of the evolution: the evolution of the probability distribution was dominated by a changing proportion of charged amino acids.

Compared with Figs. 4 and 5, PAM22 fit better with both the circulating strains and the database in later years than in early years partly because the equilibrium probability distribution of PAM22 model (16) solved by (18) was $P_c = 0.25, P_u = 0.75$, hence the expected numbers of charged amino acid in epitope A and epitope B were 5.31 and 4.81, close to the numbers in later years depicted in Figs. 1 and 3. On the other hand, PAM22 drifted away from both the circulating strains and the database strains in the early stage after the emergence of H3N2 virus, when the immune escape was underway most strongly. This phenomena showed that H3N2 virus did not follow the general law of protein evolution in a constant environment in the early years, while its evolution returned to steady state about 20 years after its emergence in 1968. The higher fixation rate of charged residues determined by the RAMM model versus the general PAM model is due to immune pressure.

### Fluctuation and Spatial Distribution of Charge

The dominant epitopes (Gupta et al. 2006) in history were epitope A and epitope B, so in a certain year one of them was dominant while the other was subdominant. As a common trend, the number of charged amino acids on the dominant epitope increased and that on the subdominant epitope decreased. The explanation is likely that an increase in number of charged amino acids reduces the free energy of binding of antibodies to the epitope, by increasing the affinity of the epitope for water. Although the number of charged amino acid on the subdominant epitope decreased almost monotonically, it never decreased to the initial level.

Nakajima et al. (2007) pointed out that the epitopes of A/Aichi/2/1968 and A/Kamata/14/1991 share similar shapes, hence the spatial distribution of charged amino acids on the epitopes could be estimated from the available structure of Aichi strains in 1968 (PDB entry: 1KEN). We plotted such figures for epitope A and B for each circulating strain. The number and position of hydrophobic amino acids were relatively conservative: hydrophobic amino acids existed at position (130, 138, 168) on epitope A and position (163, 194, 196, 198) on epitope B in each circulating strain. Positions (130, 138, 168) located in three corners of epitope A while positions (163, 194, 196, 198) constituted a continuous region on the center of epitope B. The hydrophobic center on epitope B was gradually surrounded by charged amino acids during the virus evolution in 1972–1987, so in the epitope shift year of 1989, epitope B contained a hydrophobic center surrounded by charged amino acid, a structure found in other protein-protein interaction cases (Clackson and Wells 1995). This arrangement of amino acids was preserved in epitope B in later years, which may be the reason that the number of charged amino acids stayed in a high level in epitope B. Epitope A, however, did not evolve into the stable structure with a hydrophobic center surrounded by charged residues; thus the number of charged amino acids fell to a lower level when epitope A was subdominant than when epitope B was subdominant. We also repeated the same analysis on the recently emerged H5N1 strains, using as epitopes the residues aligned to H3N2 epitopes. We did not observe a significant change in the number of charged amino acids, perhaps since the H5N1 virus has not evolved substantially in human.

## Conclusion

Mutation of the H3 hemagglutinin on the surface of the the influenza virus decreases the ability of the immune system to recognize the flu and decreases the efficacy of the annual vaccine. We show that influenza tends to increase the number of charged amino acids in the regions of hemagglutinin that the immune system recognizes, probably because this reduces ability of antibodies to bind hemagglutinin. An interesting corollary of this selection is that the number of charges in the dominant epitope of the dominant circulating virus strain is never fewer than that in the vaccine strain, chosen early in the season. We developed a model of the evolution of charge in hemagglutinin by partitioning 20 amino acids into two categories: charged and uncharged, calibrated this model on virus evolution data in humans, and demonstrated the model on Guinea pig animal model studies.

Protein evolution models such as PAM model and BLOSUM model typically apply to the evolution of bacteria, archaea, and eukaryota. For influenza virus, the harsh and changing environment due to immune pressure on the virus, makes its evolution a non-equilibrium dynamics, especially in the time period after its initial emergence in humans. The RAMM model supports the hypothesis that the rate of charge evolution is greater in regions of hemagglutinin recognized by the immune system than in proteins in general. Such temporal and spatial heterogeneity requires a method such as we have presented here for modeling the virus evolution.

**Acknowledgments** The original virus stock used in the Guinea pig infections was purchased from the Centers for Disease Control and Prevention. This work was partially supported by DARPA.



## References

Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459–468

Clackson T, Wells JA (1995) A hot spot of binding energy in a hormone-receptor interface. Science 267:383–386

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, vol 5, pp 345–352

Frank SA (2002) Immunology and evolution of infectious disease. Princeton University Press, Princeton, NJ

Goldman N, Whelan S (2002) A novel use of equilibrium frequencies in models of sequence evolution. Mol Biol Evol 19:1821–1831

Gupta V, Earl DJ, Deem MW (2006) Quantifying influenza vaccine efficacy and antigenic distance. Vaccine 24:3881–3888

Henikoff S, Henikoff JG (1992) Amino-acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Kaverin NV, Matrosovich MN, Gambaryan AS, Rudneva IA, Shilov AA, Varich NL, Makarova NV, Kropotkina EA, Sinitsin BV (2000) Intergenic HA–NA interactions in influenza A virus: postreassortment substitutions of charged amino acid in the hemagglutinin of different subtypes. Virus Res 66:123–129

Kosiol C, Goldman N (2005) Different versions of the Dayhoff rate matrix. Mol Biol Evol 22:193–199

Leunissen JA, van den Hooven HW, de Jong WW (1990) Extreme differences in charge changes during protein evolution. J Mol Evol 31:33–39

Lowen AC, Mubareka S, Tumpey TM, García-Sastre A, Palese P (2006) The Guinea pig as a transmission model for human influenza viruses. Proc Natl Acad Sci USA 103:9988–9992

Macken C, Lu H, Goodman J, Boykin L (2001) The value of a database in surveillance and vaccine selection. In: Osterhaus ADME, Cox N, Hampson AW (eds) Options for the control of influenza IV. Elsevier, Amsterdam, accession number ISDN38157. http://www.flu.lanl.gov/

McDonald JH (2006) Apparent trends of amino acid gain and loss in protein evolution due to nearly neutral variation. Mol Biol Evol 23:240–244

Minin VN, Suchard MA (2008) Counting labeled transitions in continuous-time markov models of evolution. J Math Biol 56:391–412

Muñoz ET, Deem MW (2005) Epitope analysis for influenza vaccine design. Vaccine 23:1144–1148

Müller T, Vingron M (2000) Modeling amino acid replacement. J Comput Biol 7:761–776

Müller T, Spang R, Vingron M (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol 19:8–13

Nakajima S, Nakajima K, Nobusawa E, Zhao J, Tanaka S, Fukuzawa K (2007) Comparison of epitope structures of H3HAs through protein modeling of influenza a virus hemagglutinin: mechanism for selection of antigenic variants in the presence of a monoclonal antibody. Microbiol Immunol 51:1179–1187

Pan K, Deem MW (2009) Comment on Ndifon et al., "On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness". Vaccine 27:5033–5034

Sinha N, Mohan S, Lipschultz CA, Smith-Gill SJ (2002) Differences in electrostatic properties at antibody-antigen binding sites: Implications for specificity and cross-reactivity. Biophys J 83:2946–2968

Smith DJ, Forrest S, Ackley DH, Perelson AS (1999) Variable efficacy of repeated annual influenza vaccination. Proc Natl Acad Sci USA 96:14001–14006

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, Fouchier RAM (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305:371–376

Sun J, Deem MW (2009) Statistical mechanics of the immune response to vaccines. In: Zaman MH (ed) Statistical mechanics of cellular systems and processes. Cambridge University Press, pp 177–213

Sun J, Earl DJ, Deem MW (2006) Localization and glassy dynamics in the immune system. Mod Phys Lett B 20:63–95

Thorne JL (2000) Models of protein sequence evolution and their applications. Curr Opin Genet Dev 10:602–605

Tseng YY, Liang J (2006) Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. Mol Biol Evol 23:421–436

Veerassamy S, Smith A, Tillier ERM (2003) A transition probability model for amino acid substitutions from blocks. J Comput Biol 10:997–1010

Zhou H, Pophale R, Deem MW (2009) Computer-assisted vaccine design. In: Wang Q, Tao YJ (eds) Influenza: molecular virology. Horizon Scientific Press, Norwich, UK