JOURNAL OF THE ROYAL SOCIETY Interface

Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza

Keyao Pan and Michael W. Deem

J. R. Soc. Interface 2011 **8**, doi: 10.1098/rsif.2011.0105 first published online 4 May 2011

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2011) **8**, 1644–1653
doi:10.1098/rsif.2011.0105
Published online 4 May 2011

Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza

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

[1]*Department of Bioengineering, and* [2]*Department of Physics and Astronomy, Rice University, 6100 Main Street, Houston, TX 77005, USA*

Many viruses evolve rapidly. For example, haemagglutinin (HA) of the H3N2 influenza A virus evolves to escape antibody binding. This evolution of the H3N2 virus means that people who have previously been exposed to an influenza strain may be infected by a newly emerged virus. In this paper, we use Shannon entropy and relative entropy to measure the diversity and selection pressure by an antibody in each amino acid site of H3 HA between the 1992–1993 season and the 2009–2010 season. Shannon entropy and relative entropy are two independent state variables that we use to characterize H3N2 evolution. The entropy method estimates future H3N2 evolution and migration using currently available H3 HA sequences. First, we show that the rate of evolution increases with the virus diversity in the current season. The Shannon entropy of the sequence in the current season predicts relative entropy between sequences in the current season and those in the next season. Second, a global migration pattern of H3N2 is assembled by comparing the relative entropy flows of sequences sampled in China, Japan, the USA and Europe. We verify this entropy method by describing two aspects of historical H3N2 evolution. First, we identify 54 amino acid sites in HA that have evolved in the past to evade the immune system. Second, the entropy method shows that epitopes A and B on the top of HA evolve most vigorously to escape antibody binding. Our work provides a novel entropy-based method to predict and quantify future H3N2 evolution and to describe the evolutionary history of H3N2.

Keywords: virus; influenza; evolution; prediction; selection; entropy

1. INTRODUCTION

A common strategy by which viruses evade pressure from the immune system is to evolve and change their antigenic profile. Viruses with a low evolutionary rate that infect only humans, such as the small pox virus [1], can be effectively controlled by vaccinating the human population. By contrast, viruses with a high evolutionary rate, such as HIV, hepatitis B and influenza A, resist being eliminated by the immune system by generating a plethora of mutated virus particles and causing chronic or repeated infection. In this study, we take subtype H3N2 influenza A virus as a model evolving virus. Influenza A virus circulates in the human population every year, typically causing three to five million severe illnesses and 250 000–500 000 fatalities all over the world [2]. Haemagglutinin (HA) and neuraminidase (NA) are two kinds of virus surface glycoproteins encoded by the influenza genome. The subtype of influenza is jointly determined by the type of HA ranging from H1 to H16, and that of NA ranging from N1 to N9. On the surface of the virus membrane, HA exists as a cylindrical trimer containing three HA monomers, and each monomer comprises two domains, HA1 and HA2. HA is also a key factor in virus evolution, because it is the major target of antibodies, and HA escape mutation changes the antigenic character of the virus presented to the immune system. The H3N2 virus causes the largest fraction of influenza illness. H3 HA is under selection by the immune response mainly on the five epitope regions in the HA1 domain [3], labelled epitopes A to E, as shown in figure 1. The immune pressure and the escape mutation drive the evolution of the H3N2 virus. The underlying mutation rate of the HA gene is 1.6×10^{-5}/amino acid position/day (the average mutation rate of influenza A virus is equivalent to 1.6×10^{-5}/residue/day, or 5.8×10^{-3}/residue/year [4]), measured using the method modified from that in an earlier study on the HA mutation rate [5]. Note that the mutation rate does not necessarily equal the evolution rate, or the fixation rate. The mutation rate equals the evolutionary rate only if the evolution is neutral. The non-neutrality of the HA evolution is shown in §3. Evolution of the HA viral protein causes occasional mismatch between the virus and the vaccine and

*Author for correspondence (mwdeem@rice.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2011.0105 or via http://rsif.royalsocietypublishing.org.

Received 23 February 2011
Accepted 11 April 2011

decreases vaccine effectiveness [6,7]. As more amino acid substitutions are introduced into influenza sequences, the antigenic characteristics of influenza strains drift away [8], and influenza epidemic severity of subtype H1N1 [9] and subtype H3N2 [10] increases.

The H3N2 virus has a distinguished evolutionary history, largely affected by the immune pressure. The H3N2 virus emerged in the human population in 1968 and has been circulating in the population since 1968. The phylogenetic tree of H3 HA since 1968 has a linear topology in which most sequences are close to the single trunk of the tree, and the lengths of the branches are short [11–13]. Historical HA sequences fall into a series of clusters, each of which has similar genetic or antigenic features and circulates for 2–8 years before being replaced by the next cluster [12,14]. The evolution of different amino acid positions of HA shows remarkable heterogeneity: a subset of positions undergo frequent change, while some positions are conserved [15]. This heterogeneity is quantified by the Shannon entropy at each position of the amino acid sequence of HA [16]. Shannon entropy has been used to locate protein regions with high diversity, such as the antigen-binding sites of T-cell receptors [17]. Shannon entropy has been used to identify antibody-binding sites, or epitopes, which are under immune pressure and so are rapidly evolving [16]. The heterogeneity of amino acid substitution suggests that point mutations randomly occurring in distinct positions have different contributions to virus fitness.

The selection pressure on the H3N2 virus to evolve is reflected in the difference between the H3 HA sequences in two consecutive seasons. We consider Northern Hemisphere strains. When the epidemic initiates in a new season, we assume that each position of an HA sequence inherits the amino acid from a sequence of the previous season or has a different amino acid owing to random mutation and selection. This assumption comes from the fact that the H3N2 virus circulating in each influenza season migrates from a certain geographical region in which the virus is preserved between two influenza seasons [13,18]. In the absence of selection, the histogram of the 20 amino acid usage in one position in the current season is similar to that in the same position in the previous season except for changes owing to the small random mutation rate. The difference between the two histograms beyond that expected owing to mutation quantifies selection.

Synthesizing these factors, we introduce an entropy method to describe the evolution of influenza. The entropy method extracts an evolutionary pattern from aligned sequences. Shannon entropy quantifies the amount of sequence information in each position of aligned sequences [19,20]. The sequence information reflects the variation, which is equivalently diversity, in each position, and so Shannon entropy has been used to measure the diversity in each position [21–23]. Shannon entropy has also been used to measure the structural conservation in the protein-folding dynamics [24,25]; see Valdar [26] for a detailed review of the applications of Shannon entropy. On the other hand, relative entropy measures gain of sequence information at each position and requires a background amino acid frequency distribution [27]. Relative entropy was also used as a sequence conservation measure to detect functional protein sites ([28]; see the supplemental data for the derivation of eqn 7 in Halabi *et al.* [29]). Further, a dimension reduction technique using relative entropy has identified sectors in proteins [29,30]. As an extension of these previous works, we apply Shannon entropy and relative entropy to jointly measure two quantities in each position: sequence information in one season and gain of sequence information from one season to the next season. Simultaneous analysis of Shannon entropy and relative entropy sheds light on the evolutionary pattern of the H3N2 virus evolution when data from multiple seasons are available. In the HA1 domain, positions in the epitope regions have increased Shannon entropy, and this feature was applied to locate the epitopes of H1 HA [16]. We here use Shannon entropy to quantify the virus diversity in each amino acid position in each season. The entropy relative to the previous season [31] is also used to analyse the evolution of the HA1 domain in one single season and to quantify the selection pressure on the virus in each amino acid position in each season. The selection and the virus diversity are two significant state variables determining the dynamics of evolution.

The article is organized as follows. Section 2 presents the data used in this work and details of the entropy model. Section 3 uses the Shannon entropy of the sequence in one season to predict the evolution quantified by relative entropy from this season to the next season. Results are also presented for the flow of virus migration between the four geographical regions of China, Japan, the USA and Europe. In §3, we demonstrate the entropy method in two applications, the results of which agree with prior knowledge on H3N2 evolution. Finally, we discuss our results and present our conclusions.

2. MATERIAL AND METHODS

2.1. Sequence data

The HA sequences considered in this work are the amino acid sequences of the HA of the human influenza A H3N2 virus. We only use Northern Hemisphere sequences because 90 per cent of the world population lives in Northern Hemisphere. The influenza season in Northern Hemisphere is defined as the time interval from October in one year to September in the next year. We downloaded 5309 Northern Hemisphere sequences labelled with month of collection from the NCBI Influenza Virus Resource on 16 January 2011. Sequences too short to cover residues 1–329 of HA were removed, and the remaining sequences were trimmed to only keep residues 1–329 in the HA1 domain. Any sequence with an undefined amino acid in residues 1–329 was removed. We consider 18 seasons from 1992–1993 to 2009–2010 during which most available sequences were collected. In total, we preserved and aligned 4292 sequences in these 18 seasons containing amino acids 1–329.

2.2. Histograms of 20 amino acids

The first step is to quantify the alignment of the amino acid sequences. The aligned historical H3 HA sequences

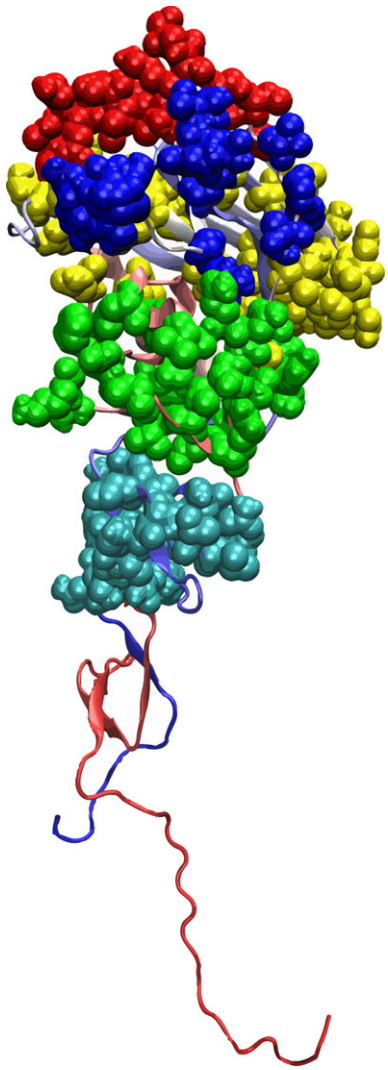

Figure 1. The tertiary structure of the HA1 domain of H3 HA (PDB code: 1HGF). The surface of HA1 facing outward is the exposed surface when the HA trimer is formed. The other two HA1 domains (not shown) in the HA trimer are located at the back of the structure displayed here. The solid balls represent five epitopes. Colour code: blue, epitope A; red, epitope B; cyan, epitope C; yellow, epitope D; green, epitope E.

form a matrix $\mathbf{A}$ with 4292 rows and 329 columns. The element $\mathbf{A}_{l,j}$ denotes the identity of the amino acid in sequence l and position j. The 4292 sequences were clustered into 18 groups by the seasons of sampling from the 1992–1993 season to the 2009–2010 season. Note that most of the sequences before the 1992–1993 season were not labelled with month of collection and were excluded from this study. We denote by $i = 0,1,\ldots,17$ the seasons between 1992–1993 and 2009–2010. For position j in season i, the relative frequency of each amino acid k, $f(k,i,j)$, $k = 1,\ldots,20$, was calculated from the vector $\mathbf{A}_{l(i),j}$, where the index array $l(i)$ holds the indices of sequences sampled in season i.

2.3. Shannon entropy as diversity

The Shannon entropy is one useful quantification of diversity in a single position. A large Shannon entropy has the physical meaning that the amino acid in the given position is prone to be substituted. This physical meaning was also applied in Deem & Pan [16]. The diversity at a single position takes the format of Shannon entropy because of the sensitivity of Shannon entropy to diversity.

This physical meaning of the Shannon entropy does not necessarily involve the joint frequency distributions for two and more positions, and we do not consider the joint frequency in this paper. Rather, we define the diversity only in the level of a single amino acid position. Consequently, the defined diversity is additive for a number of positions. The idea of adding diversity in each position of the sequence comes from classic works such as Schneider *et al.* [19], which added the Shannon entropy in each position to measure the total diversity in an aligned binding site.

Therefore, the diversity of the virus in each position in each season is represented by the Shannon entropy that quantifies the amount of information in the histogram or distribution under study. For the sequences sampled in all the seasons, positions with a high evolutionary rate have a higher Shannon entropy compared with the conserved positions [16]. The sequences in each season are assumed to be collected concurrently. The Shannon entropy is a quantification of the diversity of amino acids in one position, and so the diversity in position j in season i is calculated from the histogram $\mathbf{f}(i,j) = [f(1,i,j),\ldots,f(20,i,j)]^{\mathrm{T}}$ by Shannon entropy

$$D_{i,j} = -\sum_{k=1}^{20} f(k,i,j)\log f(k,i,j) \tag{2.1}$$

in which $k = 1,\ldots,20$ is the identity of the amino acid in position j in season i.

2.4. Relative entropy as selection pressure

Selection in each position j in season i is reflected by the significant difference between the 20-bin histogram in the current season $\mathbf{f}(i,j)$ and that in the previous season $\mathbf{f}(i-1,j)$. In the absence of selection, random mutation and genetic drift are the dominant forces generating $\mathbf{f}(i,j)$ from $\mathbf{f}(i-1,j)$. In each position, random mutation creates a slightly modified histogram, from which amino acids are randomly chosen to appear in season i by the effect of genetic drift.

The source of random mutation is the spontaneous error of the RNA polymerase replicating the influenza virus RNA. The random mutation rate in different regions of HA is thought to be homogeneous, regardless of whether the regions are in antigenic sites or not [32]. Therefore, random mutation is modelled as a Poisson process $\mathbf{M}(\mu,g(k))$ equally affecting all the positions. Here μ is the mutation rate of the influenza A virus that equals to 5.8×10^{-3}/residue/season [4], and $g(k)$, $k = 1,\ldots,20$, is the relative frequency of each amino acid in the whole alignment $\mathbf{A}$. The probability that the original amino acid t mutates to amino acid u is

$$\mathbf{M}_{u,t}(\mu,g) = \frac{\mu g(u)}{1-g(t)}. \tag{2.2}$$

So, after mutating for one season, the histogram in position j in season $i-1$ is obtained by

$$\hat{\mathbf{f}}(i,j) = \mathbf{M}(\mu, g)\mathbf{f}(i-1,j). \tag{2.3}$$

This histogram serves as the background distribution for season i from which the sequences in season i are built.

If selection is absent, the effect of genetic drift is to create sequences in the current season by randomly choosing amino acids in each position from a background distribution $\hat{\mathbf{f}}(i,j)$. We denote by N_i the number of sequences in season i. The probability that N_i amino acids in position j have the histogram $\mathbf{f}(i,j)$ is [29]

$$\Pr\{\mathbf{f}(i,j)\} \approx \exp(-N_i S_{i,j}), \tag{2.4}$$

where

$$S_{i,j} = \sum_{k=1}^{20} f(k,i,j)\log\frac{f(k,i,j)}{\hat{f}(k,i,j)} \tag{2.5}$$

is the relative entropy between the observed histogram, $f(k,i,j)$, and the background histogram, $\hat{f}(k,i,j)$ [31].

The null hypothesis that selection is absent in the evolution is rejected if the relative entropy $S_{i,j}$ is great enough such that the probability in equation 2.4 is less than 0.05, that is, the relative entropy $S_{i,j}$ is greater than $-\log(0.05)/N_i$ in season i. Note that the majority of residues were stable in most of the seasons, and in this case the relative entropy is $S_{i,j} = \log(1/(1-\mu)) \approx \mu$. To avoid classifying these stable residues erroneously as positions under selection, a proper threshold of relative entropy needs to be larger than the mutation rate μ. Additionally, a fraction λ of the circulating HA1 sequences were not deposited in the database because of the sampling bias of the HA1 sequences. In an extreme case, in a stable position j with the real histogram of 20 amino acids $[1-\lambda,\lambda,\ldots,0]^{\mathrm{T}}$ in all the seasons, the histograms of the sequences sampled in two consecutive seasons $i-1$ and i are $\mathbf{f}(i-1,j) = [1,0,\ldots,0]^{\mathrm{T}}$ and $\mathbf{f}(i,j) = [1-\lambda,\lambda,\ldots,0]^{\mathrm{T}}$, respectively, and so the relative entropy introduced by the sampling bias is

$$S^{\mathrm{bias}} \approx (1-\lambda)\log\frac{1-\lambda}{1-\mu} + \lambda\log\frac{\lambda}{\mu/19} \tag{2.6}$$

in spite of the absence of selection in position j in season i . The relative entropy S^{bias} equals to 0.1 if a sampling bias $\lambda = 2.5\%$ exists in the HA1 database sequences. We fix the threshold of the relative entropy in season i to

$$\begin{aligned} S_i^{\mathrm{thres}} &= \max\left\{-\frac{\log(0.05)}{N_i}, (1-\lambda)\log\frac{1-\lambda}{1-\mu} + \lambda\log\frac{\lambda}{\mu/19}\right\} \\ &\approx \max\left\{\frac{3}{N_i}, 0.1\right\}. \end{aligned} \tag{2.7}$$

The numbers of collected HA1 sequences N_i were fewer than 30 only in the 1995–1996 season ($i=3$) with $N_3 = 25$. The thresholds $S_3^{\mathrm{thres}} = 0.12 > 0.1$ in the 1995–1996 season because of the small numbers of HA1 sequences. In all the other 17 seasons, the numbers of sequences N_i were greater than 30, and so the thresholds $S_i^{\mathrm{thres}} = 0.1$.

3. RESULTS

In this section, we show the positive correlation between the Shannon entropy in season i and the relative entropy from season i to season $i+1$. This correlation means that the larger the virus diversity in one season, the higher the virus evolutionary rate from this season to the next season. We draw the H3N2 migration pattern by comparing the relative entropy. The migration pattern reveals a novel migration path from the USA to Europe and shows that the virus evolutionary rate is higher in the epicentre, China, than in the migration paths. We also demonstrate the entropy method in two applications. First, we compute the average Shannon entropy and relative entropy in each position over the past 17 seasons to identify positions under selection pressure. Second, we compare Shannon entropy and relative entropy in epitope regions to find the contribution of each epitope to H3N2 evolution. Results of these two applications agree with previous studies and additionally show the heterogeneity of the selection pressure over different amino acid positions of HA, with increased pressure in the epitopes, as well as the dominance of epitopes A and B.

3.1. Correlation between Shannon entropy and relative entropy

Relative entropy $S_{i+1,j}$ in amino acid position j from season i to season $i+1$ linearly increases with the Shannon entropy $D_{i,j}$ in position j in season i. For the sequences sampled from the 1992–1993 season ($i=0$) to the 2008–2009 season ($i=16$), $329\times17 = 5593$ ordered pairs ($D_{i,j}$, $S_{i+1,j}$) are calculated. All except two pairs fall into eight bins in which the values of $D_{i,j}$ belong to eight intervals [0,0.1), [0.1,0.2), …, [0.7,0.8), respectively. The first bin with $D_{i,j}$ in [0,0.1) is discarded because it contains numerous conserved amino acid positions. The values of $S_{i+1,j}$ are averaged, respectively, in each of the seven remaining bins. As described in figure 2, the average relative entropy in each bin shows a positive correlation with midpoints of the $D_{i,j}$ interval, $R^2 = 0.70$. An amino acid position j with high Shannon entropy $D_{i,j}$ in season i is expected to present high relative entropy $S_{i+1,j}$ from season i to $i+1$. The evolution in position j from season i to $i+1$ quantified by relative entropy is therefore predicted using the mean and standard error of $S_{i+1,j}$ in the bin chosen by $D_{i,j}$ in season i.

A positive correlation is also observed between the mean values of $D_{i,j}$ and $S_{i+1,j}$ in a variety of positions j in each season i. In each season between 1992–1993 ($i=0$) and 2008–2009 ($i=16$), we average the Shannon entropy $D_{i,j}$ and relative entropy $S_{i+1,j}$ over the positions j with $D_{i,j} > 0.1$. The data point with $i=1$, (0.22,1.38), has a large standard error of the relative entropy $S_{i+1,j}$ and is excluded in the analysis below. The remaining average Shannon entropy $\langle D\rangle_i$ ($i=0,2,\ldots,16$) correlates with the average relative entropy $\langle S\rangle_{i+1}$ ($i=0,2,\ldots,16$) with $R^2 = 0.50$, as

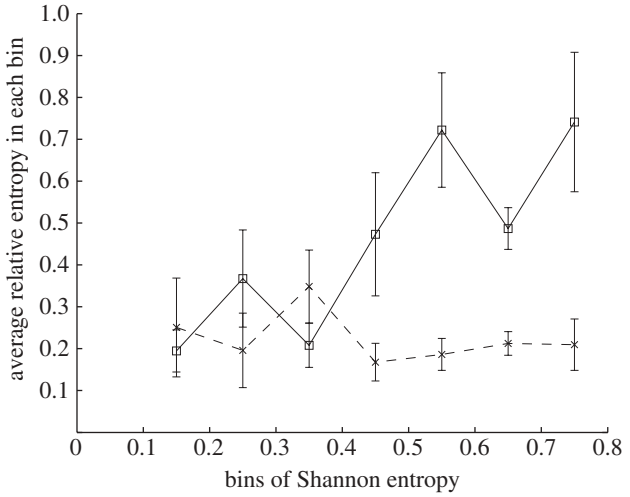

Figure 2. Mean and standard error of relative entropy $S_{i+1,j}$ in each bin of Shannon entropy. Shannon entropy and relative entropy in each of the 329 positions and in each of the 17 seasons between 1992–1993 ($i = 0$) and 2008–2009 ($i = 16$) fall into one of the eight bins. The first bin with Shannon entropy less than 0.1 is discarded. Bins with larger Shannon entropy $D_{i,j}$ also have larger relative entropy $S_{i+1,j}$. Shannon entropy $D_{i,j}$ and relative entropy $S_{i+1,j}$ in iteration $i = 51$–100 of the neutral evolution model (crosses) are used to calculate mean and standard error of relative entropy in each bin of Shannon entropy distribution in the same way. No increasing trend is found. Error bar is one standard error. Open squares represent H3 data.

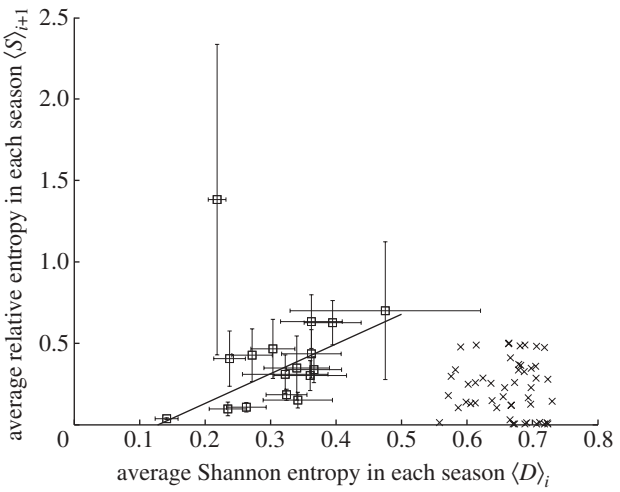

Figure 3. Average Shannon entropy $\langle D\rangle_i$ versus average relative entropy $\langle S\rangle_{i+1}$ for each season between 1992–1993 ($i = 0$) and 2008–2009 ($i = 16$). For each season i, a set of amino acid positions j with Shannon entropy $D_{i,j}$ greater than 0.1 are chosen. For all the j in this set of positions, $\langle D\rangle_i$ is the average of the Shannon entropy $D_{i,j}$ values and $\langle S\rangle_{i+1}$ is the average of relative entropy $S_{i+1,j}$ values. Horizontal and vertical error bars are the standard errors of Shannon entropy and relative entropy, respectively. The solid line, $\langle S\rangle_{i+1} = 1.82\langle D\rangle_i - 0.23$, is a least squares fit of $\langle D\rangle_i$ to $\langle S\rangle_{i+1}$ ($i = 0,2,\ldots,16$). A strong correlation with $R^2 = 0.50$ exists between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ excluding the point (0.22,1.38) with $N_i = 1$, which has a large standard error of the relative entropy $S_{i+1,j}$. Using the same method, $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ are calculated from a neutral evolution model, $i = 51$–100, and plotted. No visible correlation exists between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ from the neutral evolution model (crosses). Open squares represent H3 data.

shown in figure 3. A least-squares fit gives $\langle S\rangle_{i+1} = 1.82\langle D\rangle_i - 0.23$. Thus, the expected average relative entropy from the current season i to the next season $i+1$ can be calculated from the average Shannon entropy $\langle D\rangle_i$ in the current season i.

The above relationships between Shannon entropy and relative entropy cannot be generated by a neutral evolution model. To demonstrate this result, we create an ensemble of 1000 identical sequences with 50 amino acid positions. Each iteration of the model simulates H3N2 evolution during one season. In each iteration, the number of mutated amino acids N_{mut} in each sequence follows a Poisson distribution with mean $\mu = 2.0$, which is the annual substitution rate in history. The N_{mut} mutated positions are then randomly assigned in the corresponding sequence. We randomly select $p_{\mathrm{cut}} = 10\%$ of the sequences to build the sequence ensemble in the next iteration. The Shannon entropy $D_{i,j}$ and relative entropy $S_{i+1,j}$ generated in iteration $i = 51$–100 are processed using the same method as for H3 sequences in history. First, as shown in figure 2, no increasing trend appears in the means of $S_{i+1,j}$ in the seven bins from [0.1,0.2) to [0.7,0.8). Second, no correlation ($R^2 = 0.003$) is observed between $\langle D\rangle_i$ and $\langle S\rangle_{i+1}$ (figure 3). When we change the parameters μ between 1.0 and 10 and p_{cut} between 1 and 100 per cent in the algorithm, the simulation still does not yield the visible increasing trend in figure 2 or the correlation observed in figure 3. As a result, we conclude that neutral evolution alone is not able to generate the pattern between Shannon entropy and relative entropy of H3 sequences. It was previously shown that the fixation rate of H3N2 evolution cannot be explained only by neutral evolution [15]. In this study, the monotonically increasing linear relationship between relative entropy and Shannon entropy in figures 2 and 3 suggests that selection pressure substantially contributes to H3N2 evolution.

3.2. Annual virus migration

The entropy method is also used to analyse the global migration pattern of the virus. Most of the Northern Hemisphere H3 sequences were collected in East–Southeast Asia, the USA and Europe. East–Southeast Asia is suggested to be the reservoir of the annual H3N2 epidemic [13]. To increase the geographical resolution in East–Southeast Asia, we use two regions, China and Japan, as the representative of East–Southeast Asia, because each of these regions has a population over 50 million and has a consistent time series of H3 sequence data from the 2001–2002 to the 2007–2008 season.

We select the sequences in four geographical regions, i.e. China, Japan, the USA and Europe in seven seasons from 2001–2002 to 2007–2008. In all the six pairs of consecutive seasons, we calculate for each region four average relative entropy values of the whole sequence. These four average relative entropy values are calculated using the sequences in each of the four regions in the previous season as the reference. The results are shown in table 1. Sequences collected in China in the previous season yield the minimum relative entropy to the sequences in the current season collected in China

Table 1. The relative entropy between HA sequences in the different regions in the current influenza season and sequences in these regions in the previous season. The minimum relative entropy in each column is marked in bold. The p values of the Wilcoxon signed-rank test between the minimum relative entropy and other relative entropy values in the same column are in the parentheses. HA sequences were collected from four geographical regions: China, Japan, the USA and Europe. Seven seasons from 2001–2002 to 2007–2008 are used here. The relative entropy values listed in this table are averaged for all the sites and all the six pairs of consecutive seasons. These results imply that the H3N2 viruses in China, Japan, and the USA migrate from China, while the H3N2 virus in Europe migrates from USA.

region of the previous season	region of the current season	China	Japan	USA	Europe
China		**0.057**	**0.040**	**0.044**	0.064 (0.0017)
Japan		0.114 (2.1×10^{-5})	0.094 (0.0049)	0.076 (0.0012)	0.059 (0.032)
USA		0.105 (2.1×10^{-10})	0.087 (3.3×10^{-8})	0.070 (6.4×10^{-5})	**0.056**
Europe		0.135 (3.8×10^{-9})	0.115 (3.4×10^{-6})	0.094 (4.8×10^{-7})	0.074 (0.15)

($p < 2.1\times10^{-5}$, Wilcoxon signed-rank test), Japan ($p < 0.0049$, Wilcoxon signed-rank test) and the USA ($p < 0.0012$, Wilcoxon signed-rank test). Sequences in the USA in the previous season have the minimum relative entropy to the sequences in Europe in the current season ($p < 0.15$, Wilcoxon signed-rank test). Relative entropy data in table 1 imply the virus migration from China, as the geographical reservoir, to Japan and the USA and suggest a migration from the USA to Europe. The result in table 1 also implies that the H3N2 virus circulating in China seeds the virus in China, Japan and the USA in the next season, and the virus in the USA probably seeds the virus in Europe in the next season.

Comparison of the relative entropy data in table 1 in the H3N2 reservoir and migration paths also reveals the H3N2 virus migration pattern. Using the Wilcoxon signed-rank test, the relative entropy data of H3 HA in China in two consecutive seasons are significantly greater than those in the three migration paths: from China to Japan ($p = 0.035$), from China to the USA ($p = 0.0030$) and from the USA to Europe ($p = 0.0017$). The relative entropy data, therefore, confirm China as the H3N2 reservoir and imply that novel H3N2 viruses are emerging in China, not during the migration process.

3.3. Positions under selection

The values of diversity as the Shannon entropy $D_{i,j}$ and selection as relative entropy $S_{i,j}$ are available for the sequences collected from the 1993–1994 season to the 2009–2010 season. First, we apply the mean field approximation to remove the variation of selection and diversity over the time, and only consider the variation of Shannon entropy and relative entropy in different positions and regions over the past 17 seasons. A profile of the pattern of Shannon entropy and relative entropy in position j comprises the average selection, the number of seasons under selection and the average diversity. The average selection $\bar{S}_j$ is expressed by the mean of relative entropy in each position over the 17 seasons

$$\bar{S}_j = \frac{1}{17}\sum_{i=1}^{17} S_{i,j} \tag{3.1}$$

and is displayed in figure 4*a*. The number of seasons under selection in each position j is calculated by

$$N_j = \sum_{i=1}^{17} H(S_{i,j} - S_i^{\mathrm{thres}}), \tag{3.2}$$

where H is the Heaviside step function. The numbers are shown in figure 4*b*. The average diversity $\bar{D}_j$ in each position is calculated by averaging the Shannon entropy over the 17 seasons from 1993–1994 to 2009–2010

$$\bar{D}_j = \frac{1}{17}\sum_{i=1}^{17} D_{i,j} \tag{3.3}$$

and is displayed in figure 4*c*.

Figure 4*d* presents the distribution of the selection. Around 76 per cent of the amino acid positions 1–329 of HA have an average selection close to zero and fall into the leftmost bin. The numbers of seasons when selection $S_{i,j}$ in these positions were greater than the threshold level S_i^{thres} are shown in figure 4*e*. The average diversities $\bar{D}_{i,j}$ in all the positions are shown in figure 4*f*. If position j is under selection with $S_{i,j} > S_i^{\mathrm{thres}}$ in greater than two of the 17 seasons between 1993–1994 and 2009–2010, or $N_j > 2$, this position j is classified as a position under selection in the evolutionary history of the H3N2 virus. The 54 positions with $S_{i,j} > S_i^{\mathrm{thres}}$ in greater than two seasons are listed in table 2.

Patterns of selection and diversity similar to those observed in historical sequences in figure 4 are generated by a Monte Carlo simulation model, as displayed in the electronic supplementary material, figure S1. The basis of the Monte Carlo simulation is that the antibody binds to one of two epitope regions on the surface of the HA1 domain [6], and the dominant epitope bound by the antibody is under immune pressure and undergoes a higher substitution rate [11]. The detailed description and discussion of the Monte Carlo model is in the electronic supplementary material, appendix S1.

3.4. Comparison of different regions

A human antibody binds to five epitopes in the H3 HA [3]. The five epitopes are located in different parts of the HA1 domain of the cylinder-like structure of the H3 HA. Epitopes A and B are on the top of the HA1 structure and are exposed in the HA trimer. Epitope D is on

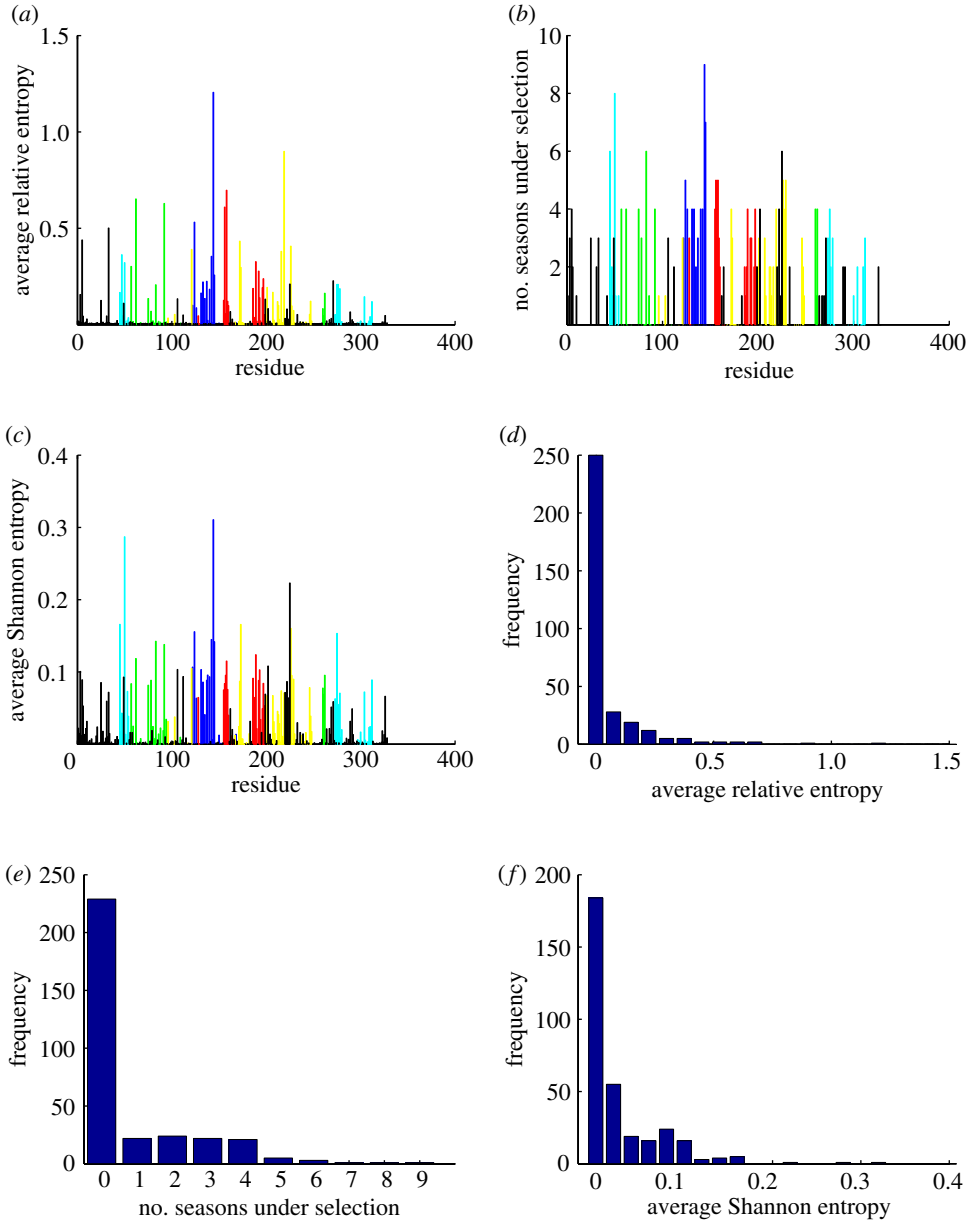

Figure 4. (*a*) Average selection in each position quantified by relative entropy during the past 17 seasons from 1993–1994 to 2009–2010, calculated by $\bar{S}_j = \sum_{i=1}^{17} S_{i,j}/17$. The colours represent positions in epitopes A to E and positions outside the epitopes, as in figure 1. (*b*) Number of seasons for each position when the relative entropy was greater than the threshold S_i^{thres}, i.e. the position was under selection. (*c*) Average diversity in each position quantified by Shannon entropy in the seasons from 1993–1994 to 2009–2010, calculated by $\bar{D}_j = \sum_{i=1}^{17} D_{i,j}/17$. (*d*) Distribution of the average selection in each position displayed in (*a*). (*e*) Distribution of the numbers of seasons under selection displayed in (*b*). (*f*) Distribution of the average diversity in each position shown in (*c*).

the top of the HA1 structure and is partly buried inside the HA trimer. Epitopes C and E are at the central area of the exposed surface of the HA1 domain as shown in figure 1. Using the entropy method, we will show that epitopes A and B are under the highest average selection over all the seasons. These results can be interpreted as the antibody binds mostly to the top exposed part of the structure of the HA trimer defined by epitopes A and B, and so the selection in these two epitopes is with higher intensity.

We divide the HA1 domain of the H3N2 HA into six regions, namely epitopes A to E, and positions not in any of the epitopes. These regions show significantly distinct patterns of evolution. In each season from 1993–1994 to 2009–2010, we averaged selection and diversity in each epitope and the positions not in any of the epitopes. The fraction of positions j under selection defined by $S_{i,j} > S_i^{\mathrm{thres}}$ was also calculated. The averages for 17 seasons are listed in table 3. It is evident that the values in table 3 vary across the epitopes. The selection and diversity in epitopes A and B are greater than those in epitopes C, D and E for each of selection, fraction of positions under selection and diversity. The fraction of positions under selection is significantly greater than those in epitopes C, D and E ($p < 0.038$, using Wilcoxon signed-rank test). The values in epitopes C, D and E are significantly greater than those not in any of the epitopes ($p < 0.0019$ for selection,

Table 2. Amino acid positions j under selection. To be included, the positions must be under selection, $S_{i,j} > S_i^{\text{thres}}$, in greater than two seasons.

region	amino acid positions
epitope A	122 124 126 131 133 137 140 142 144 145
epitope B	128 155 156 157 158 159 189 192 193 197
epitope C	45 50 273 275 278 312
epitope D	121 172 173 201 207 219 226 227 229 246
epitope E	57 62 75 78 83 92 260 262
out of epitopes	3 5 25 33 49 106 202 222 225 271

Table 3. Annual selection, fraction of positions under selection, and diversity in epitopes A to E, positions not in any of the epitopes, and the whole HA1 sequence.

region	selection	fraction of positions under selection	diversity
epitope A	0.187	0.152	0.077
epitope B	0.157	0.134	0.062
epitope C	0.077	0.087	0.048
epitope D	0.100	0.072	0.037
epitope E	0.111	0.094	0.049
out of epitopes	0.021	0.019	0.013
the whole sequence	0.060	0.051	0.028

$p < 0.0011$ for fraction of positions under selection and $p < 6.0 \times 10^{-4}$ for diversity, using Wilcoxon signed-rank test). Consequently, epitopes A and B display the highest level of selection and diversity.

4. DISCUSSION

The Shannon entropy and the relative entropy are introduced here to quantify the diversity and the selection pressure of the evolving H3N2 virus. The foundation of entropy calculation is the assumption that the virus sequences used in entropy calculation are from a random unbiased sampling of the virus circulating in the human population. However, the sampling density of the H3N2 virus varies in different geographical regions, hence creating a sampling bias.

We have addressed this issue at the continent level by analysing data from different regions separately. We now additionally study the effect of sampling bias within one country. For example, among the H3 HA sequences labelled with month of collection in the NCBI Influenza Virus Resource Database, the New York state sequences account for about one-third of the USA sequences. However, New York state has only about 6.5 per cent of the USA population. We chose eight seasons from 2001–2002 to 2008–2009 with abundant USA sequences collected in and out of New York state during this period of time available in the database. Using the procedure in §2, we calculated the histogram of 20 amino acids in each amino acid position in each season for the New York state sequences, and that for the non-New York state sequences. Each of the $329 \times 8 = 2632$ pairs of histograms in and out of New York state was compared using the χ^2-test for homogeneity. The p-values of 2581 pairs are greater than 0.05. That is, 98.1 per cent of the pairs are not significantly different. The high sampling density in New York state does not affect the histograms of 20 amino acids in each position, implying that the sampling bias of the H3N2 virus is uncorrelated to the amino acid usage patterns.

By applying Shannon entropy and relative entropy to the aligned HA sequences labelled with month of collection, we obtain the evolution and migration pattern of the H3N2 virus in §3. First, Shannon entropy and relative entropy quantify the diversity of and the selection pressure over the virus, relatively. Relative entropy from the current season i to the next season $i+1$ linearly increased with the Shannon entropy in the current season i (figures 2 and 3). Second, relative entropy quantifies the similarity of two groups of virus and implies the migration path of the H3N2 virus (table 1). In the following text, we compare our methods and results with the literature.

The relative entropy reveals the H3N2 migration pattern. Previous studies applied phylogenetic methods in an attempt to locate the epicentre of the H3N2 epidemic in each season. Rambaut *et al.* [18] studied the dynamics of influenza sequence diversity in the temperate regions in both hemispheres to imply that the H3N2 virus originates in the tropics and migrates to the temperate regions in both hemispheres. Russell *et al.* [13] obtained the antigenic and genetic evolution rate in each region and distances of the H3N2 strains to the trunk of the phylogenic tree. This information indicated East and Southeast Asia as the epicentre, from which the H3N2 virus spreads to North America, Europe and Oceania in each season [13]. Recently, Bedford *et al.* [33] suggested the centre of the H3N2 migration network being China, Southeast Asia and the USA by estimating the migration rate between different regions in the world. Here the relative entropy, the gain of sequence information, is used as a novel measure of the sequence similarity. The H3N2 migration path is the directed graph in which each path has the minimum relative entropy, or the maximum sequence similarity. These studies reach a consensus that South China is located in the epicentre of influenza epidemics. Here, we additionally identify a novel migration path from the USA to Europe and show that virus evolutionary rate is higher in the epicentre than in the migration paths.

Previous studies have identified positions that have led to the immune escape of influenza, by resolving historical mutations. A 1997 study examining 254 H3 nucleotide sequences from 1984 to 1996 identified 14 positions that are under selection, using the dN/dS ratio method [34]. In a 1999 study involving 357 nucleotide sequences from 1983 to 1997, 18 positions were identified to be under selection, using the dN/dS ratio method [35]. A 2003 paper used the alignment of 525 nucleotide sequences from 1968 to 2000 to calculate the codon diversity and the amino acid diversity [36]. This paper reported 25 positions with the largest codon diversity and amino acid diversity. A 2007 paper located 63 positions of positive selection by alignment of 2248 sequences from 1968 to 2005 and considering substitutions at the amino acid level [15].

Our study identified 54 positions under selection at the amino acid level, using 4292 aligned sequences from the 1992–1993 season to the 2009–2010 season. Considering this historical body of work, it is apparent that the number of amino acid positions identified to be under selection has increased with the number and the time span of sequences used in the study and as the discriminating power of the data has increased. In addition, different criteria to identify the positions under selection have been introduced in the previous studies [15,34–36] and in the present study. Shih *et al.* [15] identified positions to be under selection when an amino acid substitution occurred during successive years. We here classify a position j as under selection if its relative entropy $S_{i,j}$ is greater than the threshold S_i^{thres} in greater than two seasons. These two methods identify many identical positions as well as some distinct positions.

The heterogeneity of the methods also contributes to identification of different sets of positions under selection. We note that these methods fall into two categories. The first category operates at the codon level. The $\mathrm{d}N/\mathrm{d}S$ ratio method [34,35] calculates the non-synonymous and synonymous mutation rate of the codon. The Plotkin & Dushoff [36] method comparing the codon diversity and the amino acid diversity is a variation of the $\mathrm{d}N/\mathrm{d}S$ ratio method. The second category operates at the amino acid level. Shih *et al.* [15] identified the positions with amino acid switch occurring in history to be under selection. Our entropy method recognizes the positions with relative entropy higher than the threshold, S_i^{thres}, in greater than two seasons. A large $\mathrm{d}N/\mathrm{d}S$ of a codon does not necessarily mean an amino acid switch in the same position because amino acid substitution could be unfixed. Methods at the amino acid level, such as the amino acid switch [15] and our entropy method, can identify positions with low $\mathrm{d}N/\mathrm{d}S$ to be under selection because these methods do not consider nucleotide substitutions. Positive selection does not necessarily lead to a fixed amino acid switch, and in this case the entropy method can still detect positive selection. Unlike the amino acid switch method [15], the entropy method applied in this study is able to detect unfixed amino acid substitutions arising from selection. Our entropy method releases the requirement of fixed amino acid substitution in Shih *et al.* [15] but adds one requirement: the positions under selection need to present large relative entropy in greater than two seasons. Consequently, these methods identify slightly different sets of positions to be under selection.

5. CONCLUSION

We use Shannon entropy and relative entropy as two state variables of H3N2 evolution. The entropy method is able to predict H3N2 evolution and migration in the next season. First, we show that the rate of evolution increases with the virus diversity in the current season. The Shannon entropy data in one season strongly correlate with the relative entropy data from that season to the next season. If higher Shannon entropy of the virus is observed in one season, a higher virus evolutionary rate is expected from this season to the next season. Second, the relative entropy values between virus sequences from China, Japan, the USA and Europe indicate the H3N2 virus migration from China to Japan and the USA, and identify a novel migration path from the USA to Europe. The relative entropy values in and out of China, the epicentre, show that the evolutionary rate is higher in China than in the migration paths. Moreover, the entropy method was demonstrated on two applications. First, the selection pressure of H3 HA is mainly in 54 amino acid positions. Second, the top exposed part in the three-dimensional structure of the HA trimer covered by epitopes A and B is under the highest level of selection. These results substantiate current thinking on H3N2 evolution, and show that the selection pressure is focused on a subset of amino acid positions in the epitopes, with epitopes A and B on the top of HA being dominant and making the largest contribution to H3N2 evolution. These predictions and applications show that the entropy method is not only predictive but also descriptive.

K.P.'s research was supported by a training fellowship from the Keck Center Nanobiology Training Programme of the Gulf Coast Consortia (NIH grant no. R90 DK071504). This project was also partially supported by DARPA grant HR 0011-09-1-0055.

REFERENCES

1 Li, Y., Carroll, D. S., Gardner, S. N., Walsh, M. C., Vitalis, E. A. & Damon, I. K. 2007 On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proc. Natl Acad. Sci. USA* **104**, 15 787–15 792. (doi:10.1073/pnas.0609268104)

2 World Health Organization Media Centre influenza fact sheet 211. 2009 See http://www.who.int/mediacentre/factsheets/fs211/en/index.html.

3 Wiley, D. C., Wilson, I. A. & Skehel, J. J. 1981 Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature* **289**, 373–378. (doi:10.1038/289373a0)

4 Nobusawa, E. & Sato, K. 2006 Comparison of the mutation rates of human influenza A and B viruses. *J. Virol.* **80**, 3675–3678. (doi:10.1128/JVI.80.7.3675-3678.2006)

5 Parvin, J. D., Moscona, A., Pan, W. T., Leider, J. M. & Palese, P. 1986 Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. *J. Virol.* **59**, 377–383.

6 Gupta, V., Earl, D. J. & Deem, M. W. 2006 Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine* **24**, 3881–3888. (doi:10.1016/j.vaccine.2006.01.010)

7 Deem, M. W. & Lee, H. Y. 2003 Sequence space localization in the immune system response to vaccination and disease. *Phys. Rev. Lett.* **91**, 068 101. (doi:10.1103/PhysRevLett.91.068101)

8 Liao, Y. C., Lee, M. S., Ko, C. Y. & Hsiung, C. A. 2008 Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics* **24**, 505–512. (doi:10.1093/bioinformatics/btm638)

9 Wu, A. P., Peng, Y. S., Du, X. J., Shu, Y. L. & Jiang, T. J. 2010 Correlation of influenza virus excess mortality with

antigenic variation: application to rapid estimation of influenza mortality burden. *PLoS Comput. Biol.* **6**, e1000882. (doi:10.1371/journal.pcbi.1000882)
10 Wolf, Y. I., Nikolskaya, A., Cherry, J. L., Viboud, C., Koonin, E. & Lipman, D. J. 2010 Projection of seasonal influenza severity from sequence and serological data. *PLoS Curr.* **2**, RRN1200. (doi:10.1371/currents.RRN1200)
11 Ferguson, N. M., Galvani, A. P. & Bush, R. M. 2003 Ecological and immunological determinants of influenza evolution. *Nature* **422**, 428–433. (doi:10.1038/nature01509)
12 Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E. & Fouchier, R. A. M. 2004 Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371–376. (doi:10.1126/science.1097211)
13 Russell, C. A. *et al.* 2008 The global circulation of seasonal influenza A (H3N2) viruses. *Science* **320**, 340–346. (doi:10.1126/science.1154137)
14 Plotkin, J. B., Dushoff, J. & Levin, S. A. 2002 Haemagglutinin sequence clusters and the antigenic evolution of influenza A virus. *Proc. Natl Acad. Sci. USA* **99**, 6263–6268. (doi:10.1073/pnas.082110799)
15 Shih, A. C., Hsiao, T. C., Ho, M. S. & Li, W. H. 2007 Simultaneous amino acid substitutions at antigenic sites drive influenza A haemagglutinin evolution. *Proc. Natl Acad. Sci. USA* **104**, 6283–6288. (doi:10.1073/pnas.0701396104)
16 Deem, M. W. & Pan, K. 2009 The epitope regions of H1–subtype influenza A, with application to vaccine efficacy. *Protein Eng. Des. Sel.* **22**, 543–546. (doi:10.1093/protein/gzp027)
17 Stewart, J. J., Lee, C. Y., Ibrahim, S., Watts, P., Shlomchik, M., Weigert, M. & Litwin, S. 1997 A shannon entropy analysis of immunoglobulin and T cell receptor. *Mol. Immunol.* **34**, 1067–1082. (doi:10.1016/S0161-5890(97)00130-2)
18 Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K. & Holmes, E. C. 2008 The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615–619. (doi:10.1038/nature06945)
19 Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. 1986 Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431. (doi:10.1016/0022-2836(86)90165-8)
20 Schneider, T. D. & Stephens, R. M. 1990 Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100. (doi:10.1093/nar/18.20.6097)
21 Sander, C. & Schneider, R. 1991 Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68. (doi:10.1002/prot.340090107)
22 Shenkin, P. S., Erman, B. & Mastrandrea, L. D. 1991 Information-theoretical entropy as a measure of sequence variability. *Proteins* **11**, 297–313. (doi:10.1002/prot.340110408)
23 Gerstein, M. & Altman, R. B. 1995 Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175. (doi:10.1006/jmbi.1995.0423)
24 Mirny, L. A. & Shakhnovich, E. I. 1999 Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196. (doi:10.1006/jmbi.1999.2911)
25 Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R. & Baker, D. 2000 Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303–312. (doi:10.1006/jmbi.1999.3663)
26 Valdar, W. S. J. 2002 Scoring residue conservation. *Proteins* **48**, 227–241. (doi:10.1002/prot.10146)
27 Williamson, R. M. 1995 Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. *J. Theoret. Biol.* **174**, 179–188. (doi:10.1006/jtbi.1995.0090)
28 Wang, K. & Samudrala, R. 2006 Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinform.* **7**, 385. (doi:10.1186/1471-2105-7-385)
29 Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. 2009 Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786. (doi:10.1016/j.cell.2009.07.038)
30 Lockless, S. W. & Ranganathan, R. 1999 Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299. (doi:10.1126/science.286.5438.295)
31 Kullback, S. & Leibler, R. A. 1951 On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86. (doi:10.1214/aoms/1177729694)
32 Ina, Y. & Gojobori, T. 1994 Statistical analysis of nucleotide sequences of the haemagglutinin gene of human influenza A viruses. *Proc. Natl Acad. Sci. USA* **91**, 8388–8392. (doi:10.1073/pnas.91.18.8388)
33 Bedford, T., Cobey, S., Beerli, P. & Pascual, M. 2010 Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS Pathog.* **6**, e1000918. (doi:10.1371/journal.ppat.1000918)
34 Fitch, W. M., Bush, R. M., Bender, C. A. & Cox, N. J. 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl Acad. Sci. USA* **94**, 7712–7718. (doi:10.1073/pnas.94.15.7712)
35 Bush, R. M., Fitch, W. M., Bender, C. A. & Cox, N. J. 1999 Positive selection on the H3 haemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* **16**, 1457–1465.
36 Plotkin, J. B. & Dushoff, J. 2003 Codon bias and frequency-dependent selection on the haemagglutinin epitopes of influenza A virus. *Proc. Natl Acad. Sci. USA* **100**, 7152–7157. (doi:10.1073/pnas.1132114100)