Catching Old Influenza Virus with A New Markov Model

HamChing Lam
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota 55455, USA
hamching@cs.umn.edu

Daniel Boley
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota 55455, USA
boley@cs.umn.edu

ABSTRACT

We have developed a novel Markov model which models the genetic distance between viruses based on the Hemagglutinin (HA) gene, a major surface antigen of the avian influenza virus. Using this model we estimate the probability of finding highly similar virus sequences separated by long time gaps. Our biological assumption is based on neutral evolutionary theory, which has been applied previously to study this virus [Gojobori, Moriyama, and Kimura. PNAS Vol 87. 1990]. Our working hypothesis is that after a long enough time gap and with the high mutation rate usually found in RNA viruses, many site mutations should accumulate, leading to distinct modern variants. We obtained 3439 HA protein sequences isolated through years 1918 to 2006 from around the globe, aligned them to a consensus sequence using the NCBI alignment tool, and used a Hamming distance metric on the aligned sequences. We tested our hypothesis by combining a standard Poisson process with a Markov model. The Poisson process models the occurrences of mutations in a given time interval, and the Markov model estimates the probabilities of changes to the genetic distances due to mutations. By coalescing all sequences at a given genetic distance to a single state, we obtain a tractable Markov chain with a number of states equal to the length of the base peptide sequence. The model predicts that the probability of finding highly similar virus after several decades is extremely small. The existence of recent viruses which are very similar to older viruses suggests that potentially there exists some reservoir which preserves viruses over long periods.

Keywords

Influenza virus, Poisson process, Markov Model

1. INTRODUCTION

For the past century researchers have been studying influenza viruses (IV). Belonging to the viral family *Orthomyxoviridae*, influenza viruses have eight unique RNA segments [20] that encode 10 different gene products (PB1 polymerase, PB2 polymerase, PA polymerases, Hemagglutinin (HA), Nucleoprotein (NP), Neuraminidase (NA), Matrix M1 and M2 proteins, and Nonstructural NS1 and NS2 proteins). The target of our study is the Hemagglutinin HA gene product. We have developed a novel Markov model which models the genetic distance between viruses based on the Hemagglutinin (HA) gene, a major surface antigen of the avian influenza virus. Our working hypothesis is that after a long enough time gap, many site mutations should accumulate in the virus due to a lack of a proofreading function [6], leading to distinct modern variants. We based our biological assumption on neutral theory of evolution [7, 12, 17, 8] and that each amino acid site is under a neutral mutation pressure. Previous studies have shown that subtypes of influenza virus are subjected to higher silent substitution rate [7, 24], which is consistent with the neutral theory of molecular evolution. Although their studies were conducted using nucleotide sequences, we believe that the same general concept and framework can be applied to study protein sequences of this virus under this evolutionary assumption. We test our hypothesis by combining a standard Poisson process with the Markov model. The Poisson process models the occurrences of mutations in a given time interval, and the Markov model estimates the probabilities of changes to the genetic distances due to mutations. We show that it is highly unlikely that very similar sequences would arise long after the original sequence. Given the observations of several pairs of very similar sequences separated by several decades, we conclude that there must be some reservoir or evolutionary mechanism that is capable of preserving old virus strains, allowing them to reappear after extended time intervals.

BIOKDD '08 Las Vegas, Nevada USA
 ISBN 978-1-60558-302-0 ...$5.00.

2. MATERIALS AND METHODS

2.1 Protein Sequence Data and Processing

The HA protein is the major surface antigen of the influenza virus. Its role is to bind to host cell receptors promoting fusion between the viron envelope and the host cell [20]. Influenza A virus HA genes have been classified into 16 subtypes (H1-H16) according to their antigenic properties. This HA protein is cleaved into two peptide chains HA1 and HA2 respectively when matured [19]. The HA2 chain has been found to vary less and is more conserved compared to HA1 chain [10]. The HA1 chain is 329 residues long and is the immunogenic part of HA protein. Past studies have shown that HA1 is undergoing continual diversify-

ing change [5, 14] and is the most variable portion of the influenza genome[16].

Using the NCBI Influenza database available online [23], we have collected 3439 influenza virus type A protein sequences deposited before December, 2007 (excluding identical sequences and lab strains/NIAID FLU project). This collection of protein sequences contains isolates from around the globe and from a diverse range of hosts. We used protein sequences because they were known to give more reliable results than nucleotide sequences when constructing evolutionary history [19]. Each of the 3439 sequences has a unique annotation which contains the host organism, the strain number, the year of isolation, subtype, and protein name. We aligned all sequences to a consensus sequence using the NCBI alignment tool. According to the study presented by [16], a uniform consensus strain tends to circulate for some time, since the mutations that occur during replication do not become fixed in the early stages of circulating. The aligned sequence data were then used with a genetic distance function to determine the pairwise genetic distance (including gaps) of the sequences.

The genetic distance between two sequences can be thought of as the "edit" distance, which is the number of single letter changes needed to transform one sequence to the other. This yields a simple scoring function assigning a zero to a matching amino acid base and a one to a mismatch. The sum of all mismatches is usually called the Hamming distance (k) or Hamming score for the pairwise sequence comparison. For comparison of very similar biological sequences, this Hamming distance can be used under the assumption that the observed difference between a pair of sites represents one mutation [3]. The present study could also be carried out using BLAST or any alignment algorithm, but as considerably greater expense. In [15], Hamming distance was successfully used to find interesting clusters of IV HA sequences and to predict vaccine strains with good results. Hamming distance as genetic distance between viruses has also been used effectively in modeling influenza viruses [18]. In our study, we compute the Hamming distance based on a consensus alignment to account for the small number of insertions and deletions. We then store the pairwise Hamming distance scores of HA gene in a pairwise affinity matrix and identify virus sequence pairs sharing high sequence similarity (at least 90 percent) but separated by a long time gap.

2.2 Markov model

We model all mutations as the combination of several single point mutations and use a Poisson process to model the mutation rate. The Poisson process naturally admits more complex mutations, treating them as several single point mutations occurring in rapid succession. Then we build a compact Markov model to model the mutations themselves. Markov models have proven to be a powerful tool for phylogenetic inference and hypothesis testing when modeling transitions between amino acid states. Modeling amino acid transitions is complex since proteins are made of twenty amino acids. Because of this, we take a very different approach in building our Markov model. We are trying to avoid a Markov chain where each sequence is a state because this would give rise to an exponentially large number of states (20^n where n is the number of sites). In our Markov model, we collect into a single state H_k all the protein sequences

$$\begin{pmatrix} 0 & 1 & & & & \\ x_1 & y_1 & z_1 & & & 0 \\ & \ddots & \ddots & \ddots & & \\ 0 & & x_{n-1} & y_{n-1} & z_{n-1} & \\ & & & x_n & y_n & \end{pmatrix}$$

Figure 1: Markov transition matrix

at given Hamming distance k from the starting sequence $s_0 \in H_0$. The starting sequence s_0 can be chosen either as the earliest isolated sequence or the most recent one as long as a large time gap is observed when comparing to other sequences. Our Markov model assigns the probability of an arbitrary HA sequence $s_1 \in H_k$ mutating into a different HA sequence $s_2 \in H_l$ through a single point mutation, where l must be one of $k-1, k, k+1$.

Previous studies [4, 13] have shown that to better fit the model, active sites should be excluded in the analysis under the neutral theory framework. Here we have taken the same approach where we have limited the mutations captured by our Markov chain to the HA1 domain consisting of $n = 329$ sites, since this region is less conserved than the HA2 region [15, 14]. Therefore, our Markov model has only $n+1 = 330$ states instead of the 20^n states it would have if we kept each state and each possible transition separate.

Formally, consider a finite set of states labeled $\{H_0, ..., H_n\}$. In order to keep the Markov chain to a manageable size, we group all the sequences within Hamming distance of k from a start sequence into a single "super state" H_k. At each transition, we assume a single point mutation occurs, and that this mutation of amino acid replacement exhibits uniform rate of evolution throughout long periods of evolutionary time [25]. This assumption is particularly consistent with the concept of "molecular evolutionary clock" and is central to the neutral theory [1, 7, 22]. Because of the high rate at which RNA viruses evolve, it has been observed that these sequences show the typical pattern of neutral evolution [7].

We denote by a the size of the alphabet of amino acids, in our case 20. For a sequence $s_1 \in H_k$, there is a probability k/n that the mutation occurs in one of the k positions where s_1 differs from s_0, and if this change occurs, there is a $1/(a-1)$ chance that the new amino acid in this position will match that in the same position of s_0. Hence the probability x_k of a transition from H_k to H_{k-1} is $x_k = \frac{k}{n} \cdot \frac{1}{a-1}$. Similar reasoning yields the probability y_k that a transition will remain at the same Hamming distance: $y_k = \frac{k}{n} \cdot \frac{a-2}{a-1}$. The probability that mutation will be in one of the $n-k$ sites that still match s_0 is $z_k = 1 - \frac{k}{n}$, corresponding to a transition from H_k to H_{k+1}. The resulting probabilities x_k, y_k, z_k are assembled into a Markov transition matrix M shown in Figure 1. The entries in each row of M add up to 1.

Using this model, we can compute the probability q_t that a virus will have a Hamming distance at most k from the initial source sequence after t mutations. We give the general form of how to compute the above probability. We let $v_t = (v_{t0}, v_{t1}, \ldots, v_{tn})$ be the row vector of probabilities of being in state $H_0, H_1, \ldots, H_n$, respectively, after t mutations. At $t = 0$ we are in state H_0 consisting of just the initial sequence. This is represented by the row vector

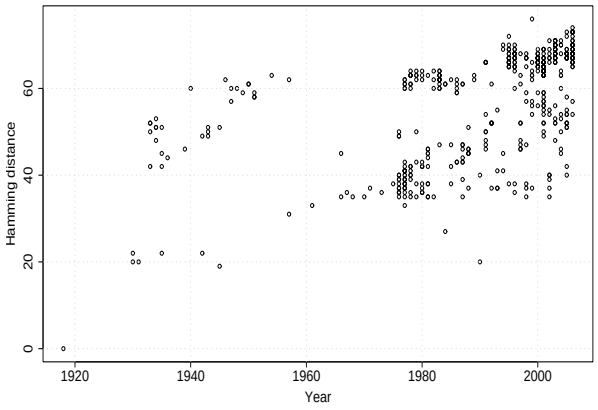

Figure 2: H1 subtype pairwise Hamming distance plot

Table 1: H1N1 subtype long time gap strains (Rate: 2×10^{-3} per site per year). H = Hamming distance, Y = Year, EG = Expected number of mutations.

Strain	H	Y	EG	$\mathcal{P}$-value
AAD17229: A/South Carolina/1/1918	0	0	0	source sequence
AAA91616: A/swine/St-Hyacinthe/148/1990(H1N1)	20	72	47.3	6.3499e-06

$v_0 = (1, 0, 0, ..., 0)$. Then the vector of probabilities after $t+1$ mutations is related to the probabilities after t mutations by $v_{t+1} = v_t * M$. The probability of being at most distance k from s_0 after t mutations is the sum of the first $k+1$ components of v_t: $q_t(k) = \sum_{i=0}^{k} v_{ti}$.

The above analysis counts events consisting of a single mutation. The mutation rate is modeled by a Poisson process [4, 11]. This includes the possibility that no mutation or several mutations take place in a given time interval, assuming all sites undergo the same substitution rate. This assumes that the probability of a mutation in a given time interval depends only on the length of the interval but is independent of the behavior outside the time interval. If λ is the average number of mutations in a time interval of 1 year, then the probability that t mutations occur in any time interval of length Y is given by $p_t(Y) = \frac{(Y\lambda)^t}{t!}e^{-Y\lambda}$. The Poisson process models when mutations occur, and the Markov model models the nature of the mutations. Combining these two models yields the probability $P_\kappa(Y)$ that after Y years a sequence would appear with a genetic distance from s_0 of κ, namely $P_\kappa(Y) = \sum_{t=0}^{\infty} p_t(Y) \cdot q_t(\kappa)$.

3. RESULTS AND DISCUSSION

We first identified viruses having very close genetic distance but with large time gap. Figure 2 shows the H1 subtype HA1 domain pairwise sequence genetic distance plotted against time of isolation in year. The genetic distance corresponds to the Hamming distance including gaps. Tables 1 and 2 show viruses sharing very high sequence similarity but with large time gap. We used the amino acid substitution rate of $r = 2\times10^{-3}$ per site per year for H1 and H2 subtype viruses, estimated using the entire region of the HA gene and

Table 2: H2 subtype long time gap strains

Strain	H	Y	EG	$\mathcal{P}$-value
AAY28987: A/Human/ Canada/720/2005(H2N2)	0	0	0	source sequence
AAA64365: A/RI/5+/ 1957(H2N2)	6	48	31.5	7.807e-09
AAA64363: A/RI/5-/ 1957(H2N2)	3	48	31.5	1.206e-11
AAA64366: A/Singapore /1/1957(H2N2)	5	48	31.5	1.155e-09
AAA43185:A/Human/ Japan/305/1957(H2N2)	5	48	31.5	1.155e-09

assuming that the molecular clock is followed [19] throughout evolutionary history. This yields an annual mutation rate of $\lambda = nr = 329\cdot2\times10^{-3} = 0.658$. We give two examples of unlikely similarities over long time gaps in table 1 and 2. Each table includes the accession number "Accession", strain name "Strain", the Hamming distance "H" (calculated from the first strain), expected number of mutations "EG", the year difference "Y", and the $\mathcal{P}$-value, the probability that this Hamming distance (or less) would be observed after the given time interval as predicted by our model. Using the pandemic strain A/South Carolina/1/1918 and A/swine/St-Hyacinthe/148/1990(H1N1) from Table 1, the interpretation of the result is that after 72 years, the expected number of mutations is 47.3 and the probability of being within a Hamming distance of 20 of the original source sequence is 6.35×10^{-6}. A very recent published research study [22] employing the state-of-the-art Bayesian Markov chain Monte Carlo [2] which allows for substitution rate variation and maximum likelihood phylogenetic methods indicates that this A/swine/St-Hyacinthe/148/1990(H1N1) virus is a contaminant from the A/swine/1930 strain. The genetic distance of the pandemic strain to the A/swine/1930 strain is 22. The genetic distance of A/swine/1930 to A/swine/St-Hyacinthe/148/1990(H1N1) is only 3 indicating that these two strains are virtually identical. From table 2, we see that A/Human/Canada/720/2005(H2N2) strain isolated in 2005 is exceptionally similar to the two asian pandemic strains A/Singapore/1/1957(H2N2) and A/Human/Japan/305/ 1957(H2N2) in terms of the genetic distance. These two pandemic strains were human transmissible and currently no influenza vaccines contained the H2N2 virus [21]. This reappearance of the highly pathogenic H2N2 virus could cause a potential pandemic as current population is not immunized against this strain of virus. The origin of the A/Human/Canada/720/ 2005(H2N2) strain was traced back to human error at a laboratory distributing virus samples for training purposes and the distributed strains were quickly destroyed at all receiving laboratories [21].

To check how our model matches the data, we show the predicted distribution of Hamming distances in Figure 3 based on a time interval of $Y = 49$ and annual mutation rate of $nr = 0.658$ for the H2 subtype. The peak of the curve indicates that with high probability, roughly 30-40 mutation events would have taken place. This tells us that we should expect to see the majority of H2 sequence pairs with Hamming distances in the vicinity of 40 given the length of time interval equals 49 years base on Poisson process assumption. We compare this to the actual distribution of Hamming distances found in the H2 subtype data shown in Figure 4 over the range of data available (from 1957 through 2006 or a span of 49 years). Figure 4 shows that the majority

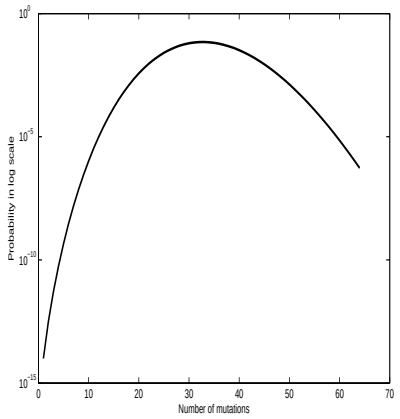

Figure 3: Poisson process distribution plot

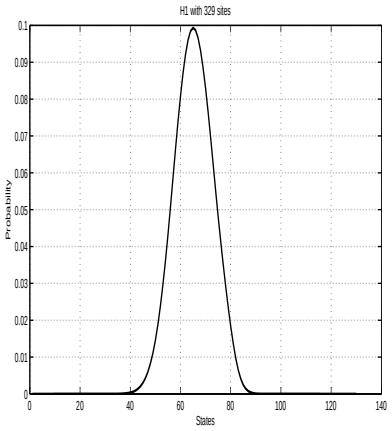

Figure 5: Model prediction plot

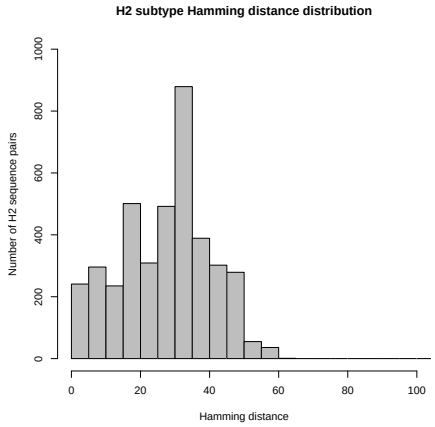

Figure 4: H2 subtype histogram plot

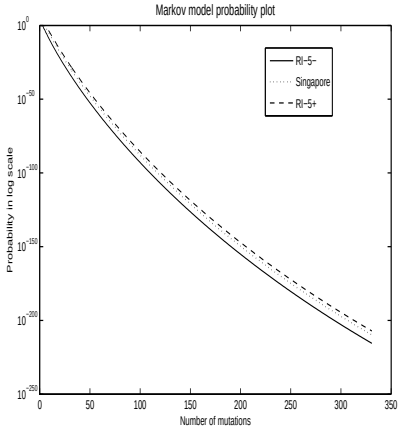

Figure 6: H2 strains probability plot

of the H2 sequence pairs have Hamming distances around 35, which matches the Poisson process prediction. Figure 6 illustrates how the probability values of 3 H2 strains in Table 2 are rapidly dropping against the expected number of mutations from the Markov model calculation. Figure 5 shows the predicted distribution within the time interval of 70-85 years from the combined Poisson process and Markov chain model using H1 subtype HA1 sequences. The curve shows that with high probability most sequences should be in states H_{60} to H_{70}. This reflects what is observed in figure 2 and figure 7 where most sequences have Hamming distance around 60-70. This suggests that our model is able to capture the overall evolutionary behavior of the influenza virus according to a molecular clock, leading to a natural increase in the genetic distance as time passes, consistent with [1].

4. CONCLUSIONS

The extensive genetic diversity of influenza A viruses through genetic drift and reassortment in the past century has resulted in many new strains being produced. However, H1, H2, and H3 subtypes strains have displayed cyclic behavior resulting in influenza pandemics [6]. In the present study, we applied neutral evolution theory to influenza virus HA protein sequences to investigate the evolutionary dynamics of the virus. Using the combination of a Poisson model with a novel Markov model, we were able to calculate the probability values of finding a very similar sequence composition separated by a large time gap. We have so far been able to identify several anomalies due to laboratory artifacts or human error. This finding is promising since we have yet to apply it in a full scale comprehensive analysis of all 16 subtypes of the virus. However, judging by the extremely low probability values obtained for some observed sample strains, we conclude that there may be one or more sources of various strains of the virus in which they are preserved over long time periods. The existence of reservoirs preserving viruses for decades cannot be completely eliminated.

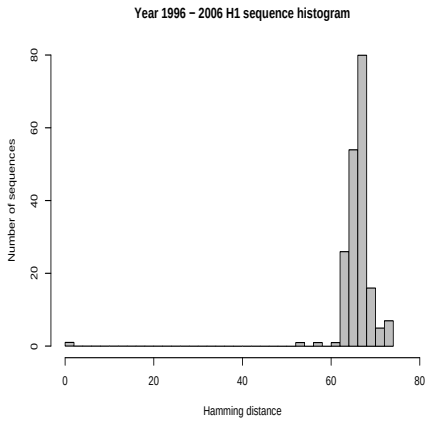

Figure 7: Histogram of H1 from 1996-2006

5. FUTURE WORK

For future work, our immediate next steps are: (1) apply our model to nucleotide sequences which allows us to compare our model with other existing models that study nucleotide sequences of the virus, and (2) use a more robust distance function in which we can incorporate antigenic distance information to the model. The evolutionary modeling of influenza virus has primarily been based on models using nucleotides substitution models and phylogenetic analysis. Our approach is different in that we demonstrated that by applying the same theoretical concept, we can instead model the differences between viral protein sequences. A key advantage of modeling the differences between sequences is that the distance function can be further refined so that additional genetic information can be incorporated into the model. However, it is imperative that we compare our model to existing models where nucleotide sequences are used and to provide a rigorous statistical framework in support of our new Markov model.

Incorporating antigenic distance information is vital due to the fact that vaccine strain selection is largely based on the antigenic differences between circulating strains and influenza viruses are antigenically variable in each influenza season. The antigenic distance map, originally proposed by Lapedes and Farber [9], is a geometric interpretation of Hemagglutination Inhibition (HI) binding assay data wheres a point is assigned in a two dimensional grid between each antigen and antiserum and this distance reflects the direct HI measurement. The antigenic distance measurement can be included in the genetic distance function to find a total distance value. Further, HI binding assay data is generated through the binding of individual viral protein to red blood cells[6], this implies that a pairwise alignment scheme for sequence comparison can be used to capture each sequence's compositional characteristic.

6. ACKNOWLEDGEMENT

This work was partially supported by NSF grant 0534286.

7. REFERENCES

[1] R. Chen and E. C. Holmes. Avian influenza virus exhibits rapid evolutionary dynamics. *Mol. Biol. Evol.*, 23:2336–2341, 2006.

[2] A. Drummond, G. Nicholls, A. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161:1307–1320, 2002.

[3] I. Eidhammer, I. Jonassen, and W. Taylor. *Protein Bioinformatics: An algorithmic approach to sequence and structure analysis*. John Wiley and Sons, 2004.

[4] W. Fitch and E. Margoliash. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as model case. *Biochemical Genetics*, 1(1):65–71, June 1967.

[5] W. M. Fitch, J. M. E. Leiter, X. Li, and R. Palese. Positive darwinian evolution in human influenza a viruses. *Proc. Natl. Acad. Sci. USA*, 88:4270–4274, 1991.

[6] S. J. Flint, L. Enquist, V. Racaniello, and A. Skalka. *Principles of Virology*. ASM press, 2004.

[7] T. Gojobori, E. Moriyama, and M. Kimura. Molecular clock of viral evolution, and the neutral theory. *Proc Natl Acad. Sci. USA*, 87:10015–10018, 1990.

[8] F. P. Kelly. *Reversibility and Stochastic Networks*. John Wiley and Sons, 1979.

[9] A. Lapedes and R. Farber. The geometry of shape space: Application to influenza. *Journal of Theor. Biol.*, 212(1):57–69, September 2001.

[10] W. Laver, G. Air, R. Webster, W. Gerhard, C. Ward, and T. Dopheide. The antigenic sites on influenza virus hemagglutinin. studies on their structure and variation in influenza virus. *Dev. Cell Biol*, 5:295–307, 1980.

[11] M. Nei and S. Kumar. *Molecular evolution and phylogenetics*. Oxford University Press, Oxford, New York, 2000.

[12] T. Ohta and M. Kimura. On the constancy of the evolutionary rate of cistrons. *J Mol Evol*, 1:18–25, 1971.

[13] M. Plass and E. Eyras. Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. *BMC Evol. Biol*, 6(50), June 2006.

[14] J. B. Plotkin and J. Dushoff. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza a virus. *Proc Natl Acad Sci USA*, 100(12):7152–7157, June 2003.

[15] J. B. Plotkin, J. Dushoff, and S. A. Levin. Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proc Natl Acad Sci USA*, 99(9):6263–6268, April 2002.

[16] A. H. Reid, T. A. Janczewski, R. Lourens, A. J. Elliot, R. Daniels, C. L. Berry, J. S. Oxford, and J. K. Taubenberger. 1918 influenza pandemic caused by highly conserved viruses with two receptor-binding variants. *Emerging Infectious Diseases*, 9(10), 2003.

[17] S. A. Sawyer. On the past history of an allele now known to have frequency p. *J Appl Probab*, 14:439–450, 1977.

[18] D. J. Smith, F. Forrest, D. H. Ackley, and A. S. Perelson. Variable efficacy of repeated annual influenza vaccination. *Proc Natl Acad Sci USA*,