Probability Analysis and Modeling of Influenza Type A virus Hemagglutinin Gene with Markov Model

HamChing Lam*and Daniel Boley†

Abstract

Under a neutral evolution framework, genetic drift evolves independently of one another and the mutation event can be modeled as a Poisson process. We derive a Markov model and using the assumption of a standard Poisson process to investigate the likelihood of finding highly similar influenza viruses separated by a long time gap. Our Markov model is based on using Hamming distance as the pairwise sequence comparison scheme. In order to keep the order of the Markov chain to a manageable size, we defined a "super state" where sequences are grouped within a certain Hamming distance. Through this model, we estimate the probability of observing highly similar influenza viruses over long time gap. We conclude that in a neutral evolutionary environment the chance of observing this is extremely low. This leads us to believe that there exists some mechanism not currently modeled that helps preserve virus sequences over long time periods.

Keywords: Influenza, Markov Model, Neutral evolution

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 [15] 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. 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 [15]. Influenza A virus HA genes have been classified into 16 subtypes (H1-H16) according to their antigenic properties. Influenza viral HA protein is cleaved into two peptide chains HA1 and HA2 respectively when matured [8]. The HA2 chain has been found to vary less and is more conserved compared to HA1 chain [5]. The HA1 chain is 329 residues long and is the immunogenic part of HA protein. Past studies have shown that HA1 is undergoing continual diversifying change [3, 12]. In this study, we utilize a Poisson process coupled with a Markov model under the assumption that genetic drift is acting in a neutral evolutionary framework [4] and each site evolves independently of one another [10, 13, 6]. Under this assumption, 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.

2 Materials and Methods

2.1 Sequence Data

Using NCBI Influenza database available online, *The Influenza Virus Resource at the National Center for Biotechnology Information* [1], we have collected 3439 influenza virus type A protein sequences (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. So far 16 subtypes (H1-H16) of influenza virus A HA genes have been classified in the past century. We used protein sequences because they were known to give more reliable results

*Dept. of Computer Science & Eng., Univ. of Minnesota, Minneapolis, MN 55455. Email: hamching@cs.umn.edu

†Department of Computer Science, University of Minnesota, Minneapolis, MN 55455. Email: boley@cs.umn.edu

than nucleotide sequences when constructing evolutionary history [8]. All the sequences have been preprocessed to eliminate gaps, and each sequence is 566 bases in length. 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.

2.2 Pairwise sequence analysis

Our pairwise sequence analysis is based on the degree of similarity between each virus sequence. We first establish a distance function in order to measure the similarity between two protein sequences. A 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 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 [2]. The present study could also be carried out using BLAST or any alignment algorithm, but as considerably greater expense. In [11], Hamming distance was successfully used to find interesting clusters of IV HA sequences and to predict vaccine strains with good results. Hamming distance as antigenic distance between viruses has also been used effectively in modeling influenza viruses [14] . In our study, we store the pairwise Hamming distance scores of HA1 domain 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. We also include multiple sequences that share very high sequence similarity with long time gap in our results section.

2.3 Markov model

We model all mutations in the sense of single nucleotide polymorphism (SNP) and use a Poisson process to model the mutation rate and then build a 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 amino acids at given Hamming distance k from the starting sequence $s_0 \in H_0$. Our Markov model estimates 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 $k-1, k, k+1$. We use a simple model of limiting 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 [11, 12]. 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 $\{H_0, H_1, ..., H_n\}$. 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 could occur in any site with equal probability. We denote by a the size of the alphabet of letters, 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 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 probabil-

ity q_t that a virus can have at most k Hamming distance away from its initial state 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 $v_0 = (1, 0, 0, ..., 0)$. 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 k distance 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 mutation. To model the probability of no mutation taking place in a given time interval, we use a Poisson process [7]. 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 $p_t(Y) = e^{-Y\lambda}\frac{(Y\lambda)^t}{t!}$.

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 Hamming distance from s_0 of κ, namely $P_\kappa(Y) = \sum_{t=0}^{\infty} p_t(Y) \cdot q_t(\kappa)$. Thus, the probability that a sequence would appear with at most k Hamming distance from s_0 is $\mathcal{P}_k(Y) = \sum_{\kappa=0}^{k} P_\kappa(Y)$.

3 Results and Discussion

We selected viruses sharing very similar sequence composition but with large time gap. We used the amino acid substitution rate of $r = 2 \times 10^{-3}$ per site per year for H1 and H2 subtype viruses, estimated using the entire region of the HA gene [8]. This yields an annual mutation rate of $\lambda = nr = 329 \cdot 2 \times 10^{-3} = 0.658$. Tables 1 and 2 show viruses sharing very high sequence similarity but with large time gap. Each table includes the accession number "Accession", strain name "Strain", the Hamming distance "H" (calculated using the first strain), expected number of mutations "EG", the year difference "Y", and the $\mathcal{P}$-value. Taking the strains A/swine/St-Hyacinthe/148/1990(H1N1) and A/South Carolina/1/1918 from Table 1 as an example, 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}. Figure 2 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.

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 $nr = 0.658$ of 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 of the H2 sequence pairs have Hamming distances around 35, which matches the Poisson process prediction. We have excluded the few sequences with Hamming distance over 300 (almost the length of the entire subsequence we are considering) in the figure. The instances of Hamming distances around 300, less than 10% of all the sequences, can be attributed to the fact that the strains are from different hosts and differences in the regions of the H2 pair being selected. Some may share the same serological reaction and hence the same potential signature epitope sites in the HA1 region even with different sequence composition.

4 Conclusions

The extensive genetic diversity of influenza A viruses through genetic drift and reassortment in the past century resulting in many new strains being generated. In this study, under the neutral evolution framework, we have illustrated the "unlikeliness" of

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

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

finding highly similar influenza virus sequences after a long time gap using a Markov model with Hamming distance metric under the assumption of a standard Poisson process. Judging by the extremely low probability values obtained of finding highly similar influenza viruses over long periods, we conclude that reservoir of viruses capable of preserving them over long period of time exist. Our model is able to capture the dynamic of the data being studied, and is sufficient to show that some underlying mechanism exists that preserves viruses even in the sequence region most subject to active mutation, even if the mechanism cannot be identified. The model can be readily applied to study the internal and external proteins of influenza virus across different subtypes.

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

Figure 1: Markov transition matrix

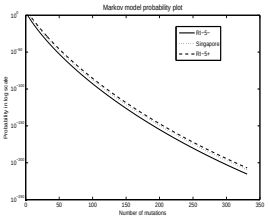

Figure 2: H2 subtype probability plot

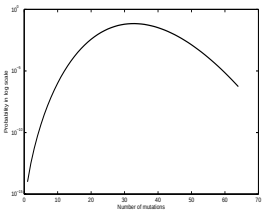

Figure 3: Poisson process distribution plot

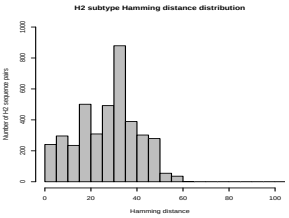

Figure 4: H2 subtype histogram plot

References

[1] Bao Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. The Influenza Virus Resource at the National Center for Biotechnology Information. J. Virol. 2008 Jan;82(2):596-601.

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

[3] Fitch W.M, Leiter J M E, Li X, and Palese R., Positive Darwinian Evolution in Human Influenza A Viruses. Proc. Natl. Acad. Sci. USA Vol. 88, 1991. pp. 4270-4274

[4] Gojobori T., Moriyama E N. and Kimura M., Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad. Sci. USA, Vol 87 1990. p.10015-10018

[5] Laver W., Air G., Webster R., Gerhard W., Ward C., Dopheide T., Structure and Variation in Influenza Virus, Elsevier North Holland, Inc., 1980, p. 295-306.

[6] Kelly F. P., Reversibility and Stochastic Networks. John Wiley and Sons, Chichester, 1979.

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

[8] Suzuki Y and Nei M., Origin and Evolution of Influenza Virus Hemagglutinin Genes. Mol. Biol. Evol. 19(4) 2002, p.501-509.

[9] Nelson M and Holmes E C., The evolution of epidemic influenza. Nature reviews, Genetics. Vol 8 2007. p.196-205

[10] Ohta T. and Kimura M., On the constancy of the evolutionary rate of cistrons. J Mol Evol, Vol 1 1971. p.18-25.

[11] Plotkin J B., Dushoff J., and Levin S. A., Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci U S A. April 30, vol 99, no. 9, 2002. p. 6263-6268.

[12] Plotkin J B. and Dushoff J., Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc Natl Acad Sci U S A. 2003 Jun 10, vol 100, no.12, p. 7152-7.

[13] Sawyer S. A., On the past history of an allele now known to have frequency p. J Appl Probab, Vol 14 1977. p.439-450.

[14] Smith D J., Forrest F., Ackley D H., and Perelson A S., Variable efficacy of repeated annual influenza vaccination. Proc Natl Acad Sci U S A. 1999 November 23; 96(24). p. 14001-14006.

[15] Webster R G., Bean W J., Gorman O T., Chambers T M., and Kawaoka Y. Evolution and Ecology of Influenza A Viruses. Microbiological Reviews, Mar. 1992, p.152-179.