# Epidemiology as Related to the Phylogenetic Analysis of the Evolution of the Influenza Virus

MTBI-02-03M

Amanda Criner[1],Brandon Hale[2], Lorena Morales-Paredes[3], David Roy Estrella Segura[4], and Ariel Cintron-Arias[5]

[1] Department of Mathematics, University of Maine, Orono, ME

[2] Departments of Mathematics and Biology, Murray State University, Murray, KY

[3] Department of Mathematics, University of Alabama at Huntsville, Huntsville, AL

[4] Department of Mathematics and Statistics, Arizona State University, Tempe, AZ

[5] Department of Mathematics, Cornell University, Ithaca, NY

**Abstract**

The evolution of the influenza virus is characterized by continual changes to its surface structures due to antigenic drift and antigenic shift. The host immune system must alter antibodies in response to the ever-changing virus allowing for the persistence of influenza in a host population. The spread of related strains through a susceptible population with regard to their phylogenetic distance from a parental strain during a season is examined, as well as the within host dynamics. Very little work has been done to integrate phylogenetic analysis of evolution with the epidemiological spread of the influenza virus. In this study, an attempt is made to couple these two scale by using infection rates that are defined as functions of phylogenetic distances between strains. Competition between strains is focused on and strain prevalence for outbreaks during several seasons (2000-2004, inclusive) is examined at various levels: global, regional, and for New York City. Coexistence is found to only be possible between very similar strains, otherwise competitive exclusion or extinction of all strains occurs. Stochastic simulations at the cellular level indicate that the immune system is most effective when the virus has little variability, so the rapid mutation of influenza is an effective strategy in evading the immune system. Similar simulations for the population level show that a strain's prevalence depends largely on the effect on the antigenic structure as a result of the locations of amino acid mutations.

## 1 Introduction

Influenza, commonly known as the flu, is caused by viruses that infect the respiratory tract. Typical symptoms of influenza include fever, cough, sore throat, runny or stuffy

nose, as well as a headache, muscle aches, and sometimes fatigue. The Centers for Disease Control (CDC) note that every year in the United States, an average of 5% to 20% of the population gets the flu, more than 200,000 people are hospitalized from flu complications, and about 36,000 people of those infected die from the disease. Influenza costs more than 20 billion dollars per year in treatment costs and lost productivity in the US. The virus spreads directly from person to person via respiratory droplets released by coughing and sneezing [4], and it can survive outside the body for several hours, especially in cold weather [10] or indirectly via inanimate objects (fomites) touched by infected individuals [12]. Some of the major pandemics of the last century include the "Spanish Flu" of $1918 - 1919$ which was responsible for 20 million deaths worldwide and 500,000 death in the US, the "Asian Flu" pandemic of $1957 - 58$ responsible for 70,000 deaths in the US and the "Hong-Kong flu" pandemic of $1968 - 69$ which resulted in 34,000 deaths in the US.[4]

The influenza virus is a member of the Orthomyxoviridae family, which is characterized by segmented negative-sense single stranded RNA molecules encased within a protein coat [10]. The influenza virus, like other viruses, can only reproduce by seizing the protein manufacturing abilities of a host cell. The virus itself does not directly kill the cell, since new viruses are released through budding rather than cell membrane rupturing as is the case with many viruses [10]. However, since the infected host cell cannot produce vital proteins needed for metabolism, it does eventually die due to starvation.

Influenza is a particularly interesting virus in that it has the ability to undergo rapid antigenic mutation [7]. The immune response against viruses such as influenza involves the use of antibodies against the viral antigen. Antigens are proteins carried by viruses that serve as targets for antibody binding. As shown in Figure 1, when a virus infects a human host cell, the antigens are displayed on cell surface proteins called human leukocyte antigens (HLAs). The HLA-antigen complex is then detected by CD4+ T-helper cells, which release cytokines that instruct B cells to begin producing antibodies. The infected cell produces new viruses until an effective antibody is produced, which allows for an immune response to be initiated.

The antibodies mark virally-infected cells for destruction by the immune system's CD8+ cytotoxic T lymphocytes (CTLs). After an infection, memory B cells continue to carry the antibodies, creating an arsenal for use against future infections. Should the same virus infect the host, the memory B cells can immediately create the same antibody to fight the virus (See Fig. 2, Case I). Since CTLs will only attack an infected cell that is marked with an antibody, the ability of the influenza virus to rapidly alter its antigen forces the immune system to alter its antibodies. The immune system utilizes a specialized B cell selection process known as somatic hypermutation to create new antibodies. The immune system selects B cells with antibodies that are more similar to the new antigen. The antibody selected is better than the original but still does not bind with the antigens perfectly, so the selection process repeats. This refinement is repeated until an antibody that is suitable for attaching the antigen is produced, so the length of time required for this process depends on the difference between the antigens of a new influenza virus as

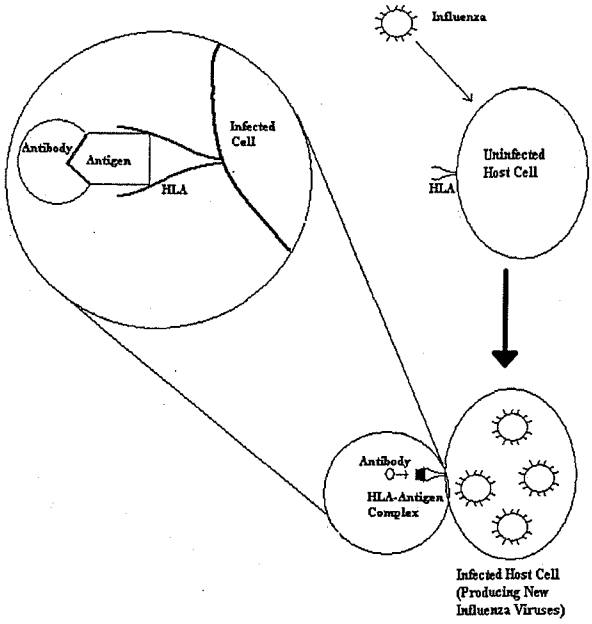



Figure 1: Model of Influenza Infection

opposed to the strain encountered previously. Due to this somatic hypermutation process, an antibody against a slightly altered influenza virus can be produced relatively quickly (Fig. 2, Case II), but more time is required to produce an antibody against a very different strain (Fig. 2, Case III).

The genetic variation between influenza strains can be represented using phylogenetic trees. A phylogenetic tree is a graphical representation of the ancestor-descendant relationship between organisms created by examining differences in nucleic or amino acid sequences [11]. Figure 2 shows the trees which correspond to the three previously mentioned cases. The genetic distances between the original influenza strain and the new strain can be calculated from the branch lengths of a phylogenetic tree. Notice that the genetic distances are horizontal only. The vertical direction represents evolutionary splits from one viral strain into two distinct new strains. Each split produces an additional genetic distance from the original strain. In Case I of Figure 2, no splits have occurred, so there is no genetic difference between the original and new strain. In Case II, only one split has occurred, so the genetic distance between the original and new strains is fairly low. Three splits have occurred in Case III, creating a much larger genetic distance between the original and new strains. Phylogenetic trees can become quite complex, as shown in Figure 3 for a group of Influenza A strain $H3N2$ variants.

The evolutionary progression of the influenza A virus is examined at both the cellular and population levels. At the cellular level, the evolution of the virus results from

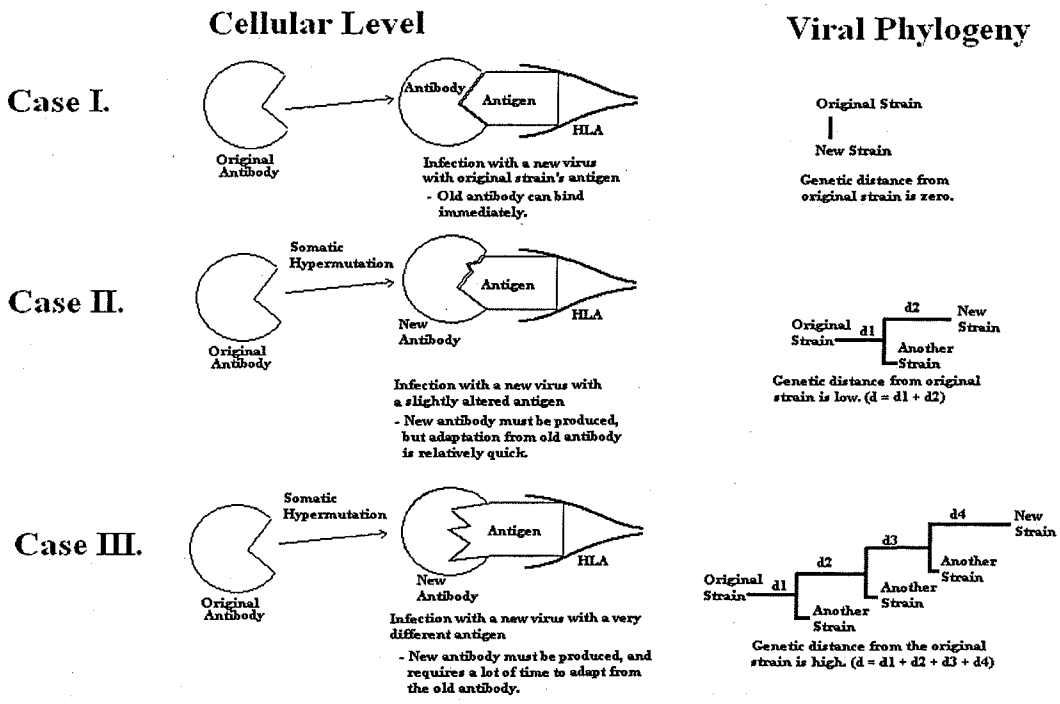



Figure 2: Model of Antibody Adaption

intense selective pressures by the immune system. At the population level, the spread of the various mutated strains can be observed to determine how the prevalence of the strains changes over time. Therefore, two models interplay in the relationship between the selective evolution of the virus and the spread of mutant viruses through the population. The evolution occurs at the cellular level within hosts, while the spread between hosts is modeled at the population level. Both are presented in this paper with the goal of modeling and capturing the epidemiological and evolutionary dynamics of the influenza virus. Linking these two levels is quite difficult, since the application of evolutionary virulence to epidemiology is still in its infancy [8]. As a result, little theory linking the two levels has been developed[23], though strides have been made by Perelson *et al.* with studies and simulations of HIV infection.[18]

In order to understand the dynamics of influenza evolution, a cellular-level model is examined in Section 2. This is followed by an introduction to evolutionary analysis with phylogenetic trees in Section 3. Section 4 presents statistical analyses of phylogenetic trees for various levels (global, regional, and for New York City), which are used to make

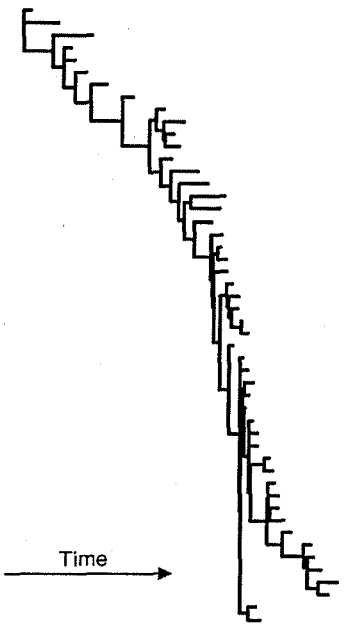



Figure 3: Phylogeny of a human influenza A virus strain $H3N2$

predictions of the most prevalent strains in each region. The genetic distances from the phylogenetic trees are then incorporated into a population-level S-I-R model in Section 5. A discussion of the results is given in Section 7.

# 2 The Cellular-Level Model

In this section, a cellular-level model is presented to examine the selection process acting upon the influenza virus as a result of a single host's immune system.

## 2.1 Deterministic Model

For an individual host, let $X$ be the population of uninfected cells, $V_i$ be the population of the viral strain $i$, $Y_i$ be population of host cells infected with viral strain $i$, and $Z_i$ be the population of the Cytotoxic T Lymphocytes (CTLs) of the immune system, which attack host cells infected with viral strain $i$.

Table 1: Parameters for Cellular-Level Deterministic Model

| Parameters at Host Level | Description |
|---|---|
| $\Lambda$ | Rate at which uninfected cells are produced per capita |
| $d$ | Death rate of the uninfected cells per capita |
| $\beta_i$ | Rate at which an uninfected host cell becomes infected by viral strain $i$ |
| $a$ | Death rate of infected cells per capita |
| $p$ | Rate at which infected cells are destroyed by CTL's per capita |
| $k_i$ | Rate at which virus $i$ is produced per capita |
| $\mu$ | Decline of rate of virus production |
| $c_i$ | Rate of CTL proliferation in response to antigen from strain $i$ |
| $b$ | Rate of decay of CTL (in absence of stimulation) |
| $m_{ji}$ | Mutation rate of virus $j$ into virus $i$ |

Table 2: Variables for Cellular-Level Deterministic Model

| Variables at Host Level | Description |
|---|---|
| $X(t)$ | Population of uninfected host cells at time $t$ |
| $Y_i(t)$ | Population of host cells infected with viral strain $i$ at time $t$ |
| $V_i(t)$ | Population of the viral strain $i$ at time $t$ |
| $Z_i(t)$ | Population of the Cytotoxic T Lymphocytes (CTL) of the immune system that attack host cells infected with viral strain $i$ at time $t$ |

$(X, Y_i, V_i, Z_i)_{i=1:n} \in \mathbb{R}^{3n+1}$

For $1 \leq i \leq n$, then the following system has dimension $3n + 1$:

$$\begin{cases} \dot{X} = & \Lambda - dX - \sum_{j=1}^{n} \beta_j V_j X \\ \dot{Y}_i = & \beta_i V_i X - (a + pZ_i) Y_i \\ \dot{V}_i = & \sum_{j=1}^{n} m_{ji} k_j Y_j - \mu V_i \\ \dot{Z}_i = & (c_i Y_i - b) Z_i \end{cases} \tag{1}$$

As shown in Figure 4, the model begins with uninfected host cells $X$. New cells enter this class at a rate of $\Lambda$, and leave the class in the form of a natural death rate of $d$, or by infection by an influenza virus. The host cells enter the $Y_i$ class upon infection by any one of the $n$ strains of viruses $V_i$ at a rate $\beta$, which is different for each strain. The infected host cells die at a a death rate of $a$, or are destroyed by the CTLs at a rate of $p$.

The class of virus particles includes $n$ strains of the influenza virus. A virus mutates from the strain $i$ class to the strain $j$ class at a mutation rate of $m_{ji}$, once during the cycle of a single virus. The rate $k$ defines the strain-specific rate at which the viruses are produced. The viruses become inviable at a rate of $\mu$.

The CTLs enter the $Z_i$ class as a result of activation due to the presence of infected cells $Y_i$ using a scaling constant $c$. The CTLs cease responding to infection at a rate $b$.

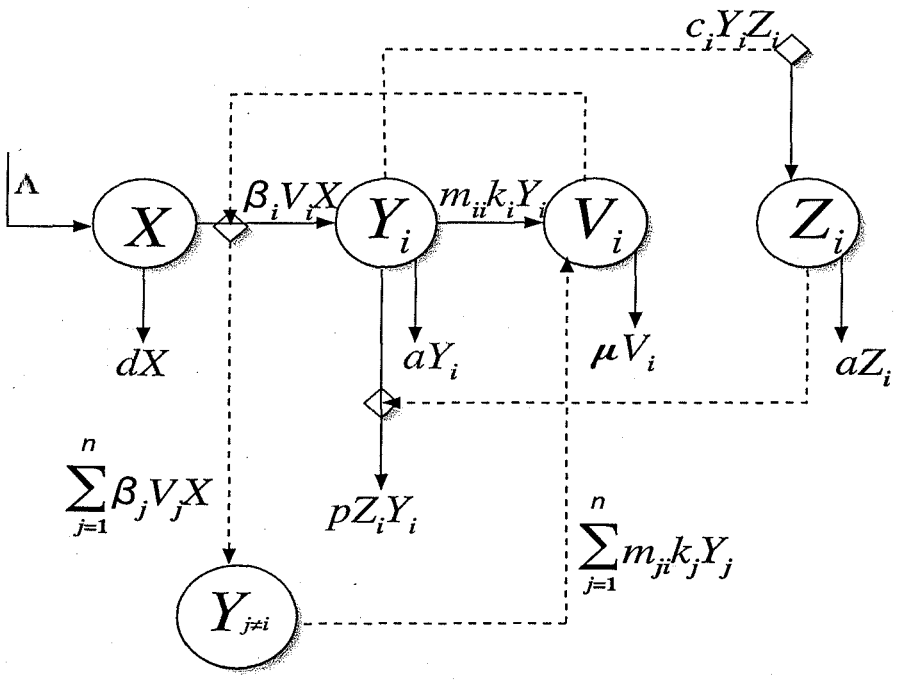



Figure 4: Within-Host Model

In the case of the host-pathogen level model, at least three equilibria exist: the disease-free equilibrium, the coexistence equilibrium, and the immune-free equilibrium.

## 2.2 Disease-Free Equilibrium $\Re_0$

The disease-free equilibrium (DFE) for the cellular-level model occurs when the virus is absent (and in turn the infected cells and CTLs as well). Therefore, to find the disease free equilibrium (DFE) for the two-strain case, ($X$, $Y_1$, $Y_2$, $V_1$, $V_2$, $Z_1$, $Z_2$) is determined with all values except $X$ set to equal zero. After solving the equations it can be seen that the value of X is $\frac{\Lambda}{d}$. Therefore, the DFE exists at:

$$(\frac{\Lambda}{d}, 0, 0, 0, 0, 0, 0)$$

The DFE can be used to calculate the basic reproduction number $\Re_0$ for the system given in (1). We are able to find the general $\Re$ for n-strains by observing the patterns

that emerge in the 1-strain system, compared to the 2-strain system. The computation is included in the Appendix, Section 8.1.1.

For n=1, we get

$$\Re_{0_1} = \sqrt{\frac{\beta_1}{\mu}\frac{\Lambda}{d}m_{11}\frac{k_1}{a}}$$

where

$$\Re_{0_i} = \sqrt{\frac{\beta_i}{\mu}\frac{\Lambda}{d}m_{ii}\frac{k_i}{a}}$$

For n=2, we get

$$\Re_0 = \sqrt{\frac{1}{2}\left[\frac{\beta_1}{\mu}\frac{\Lambda}{d}m_{11}\frac{k_1}{a}\right] + \left[\frac{\beta_2}{\mu}\frac{\Lambda}{d}m_{22}\frac{k_2}{a}\right] + \sqrt{\left[\frac{\beta_1}{\mu}\frac{\Lambda}{d}m_{11}\frac{k_1}{a} - \frac{\beta_2}{\mu}\frac{\Lambda}{d}m_{22}\frac{k_2}{a}\right]^2 + 4\frac{\beta_1}{\mu}\frac{\Lambda}{d}\frac{\beta_2}{\mu}\frac{\Lambda}{d}m_{12}\frac{k_1}{a}m_{21}\frac{k_2}{a}}}$$

If we define

$$q = \frac{\beta_2 m_{22} k_2}{\beta_1 m_{11} k_1}$$

a measure of strain two's reproductive capacity relative to strain one.

Then we can write $\Re_0$ more simply as:

$$\Re_o = \sqrt{\frac{1}{2}\left(\frac{\beta_1}{\mu}\frac{\Lambda}{d}m_{11}\frac{k_1}{a}\right)\left[1 + q + \sqrt{(1-q)^2 + 4q\frac{m_{12}m_{21}}{m_{11}m_{22}}}\right]}$$

Note:

$$\Re_0 > \sqrt{\frac{1}{2}}\Re_{o_1}\sqrt{1 + q + |1-q|} = \Re_{0_1}\sqrt{\max(1,q)} = \max(\Re_{0_1}, \Re_{0_2})$$

For a general $n$, $\Re_0 >$ max $\Re_{0_i}$
$\Re_0$ is greater because the contribution mutations make, $m_{12}$ and $m_{21}$.

## 2.3 Co-existence Equilibrium

The co-existent equilibrium for n-strain model occurs when all strains are able to attain a stable coexistence. To find this equilibrium, each of the derivative equations (1), are set equal to zero and then solved for values of $X$, $Y_i$, $V_i$, and $Z_i$. Doing so yields the following components for the co-existence equilibrium:

$$(X^\infty, Y_i^\infty, V_i^\infty, Z_i^\infty)_{i=1}^n \in \mathbb{R}^{3n+1}$$

where,

$$\begin{aligned} X^{\infty} &= \frac{\Lambda}{d+\sum_{j=1}^{n}(\frac{\beta_i}{\mu}\sum_{i=1}^{n}k_i m_{ij}(\frac{b}{c_i}))} \\ Y_i^{\infty} &= \frac{b}{c_i} \\ V_i^{\infty} &= \frac{\sum_{j=1}^{n}k_j m_{ji}\frac{b}{c_j}}{\mu} \\ Z_i^{\infty} &= (\frac{\beta_i}{\mu})(\frac{X^{\infty}}{p})[(\sum_{j=1}^{n}k_i m_{ji}\frac{c_i}{c_j}) - a] \end{aligned} \tag{2}$$

The coordinates of the co-existence equilibrium are positive (and therefore realistic) as long as this restriction is met:

$$\sum_{j=1}^{n}(\frac{k_j}{a}m_{ji}\frac{c_i}{c_j}) > 1 \quad \forall\, i$$

Figure 5 displays the stability of the co-existence equilibrium based on computer simulations.

From this picture generated by the simulation we can see that this equilibrium is stable. This is because of the fact that both the $Y_1$ and $Y_2$ level off to a specific value. For the $Y_1$ the value is approximately and for the $Y_2$ the value is very close to 0

## 2.4 Immune-Free Equilibrium

There is also a case in which the host's immune system does not respond to the infection. Such an equilibrium would occur in a host with AIDS or another condition which severely compromises the immune system. This immune-free equilibrium occurs (IFE) when the CTL levels ($Z_i$) remain at zero, regardless of the infection size. The IFE is given by Equation (3) for any strain $i$.

$$\begin{aligned} X^{\infty} &= \frac{\Lambda}{d+\sum_k \beta_k V_k^{\infty}} \\ Y_i^{\infty} &= \frac{\beta_i}{a}X^{\infty}V_i^{\infty} \\ Z_i^{\infty} &= 0 \;\; \forall\, i \end{aligned} \tag{3}$$

See Appendix, Section 8.1.2, for computation.

The biological meaning of these equilibria can be interpreted as failure to produce effective antibodies against particular strains,this is realistic, since CTLs cannot attack infected cells unless they are marked with antibodies. This may also reflect the antigenic

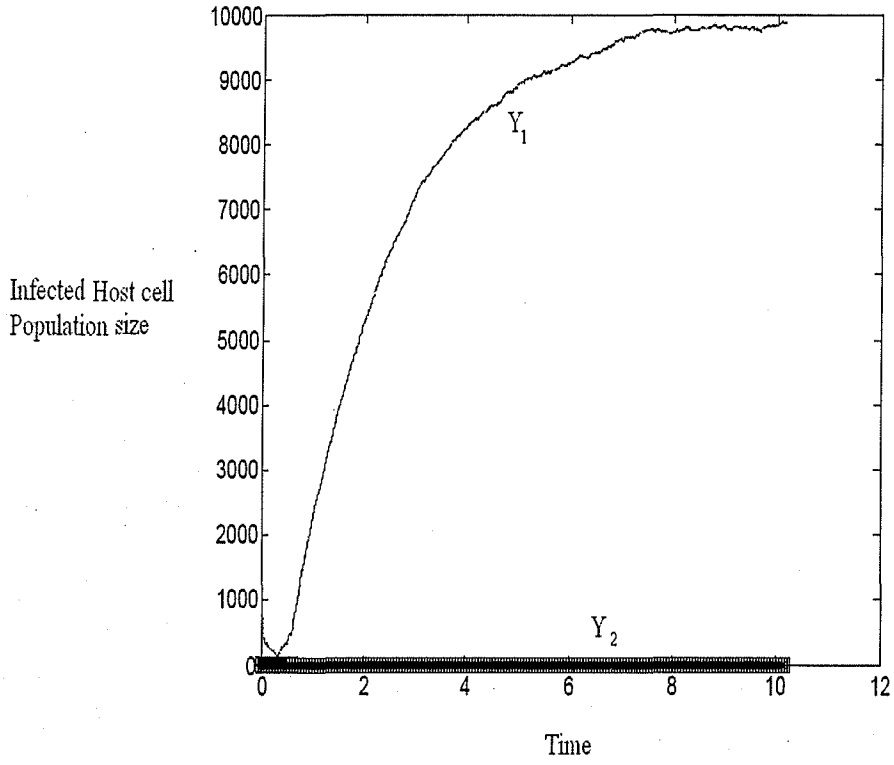



Figure 5: Simulated Stability of the Co-existent Equilibrium

distance hypothesis proposed by Smith *et al.*[21], which results from a study of influenza vaccines.

In some cases a vaccine against a strain may not result in an immune response if the antigenic structure of the vaccine was very similar to that of a vaccine administered in a previous year. This is due to cross-reactivities of the antibodies, so the antibodies created in response to the first vaccine prevent new antibody production against the second vaccine.

## 2.5 Competitive Exclusion

Competitive exclusion for the two-strain case occurs when one strain infects the entire host while the other dies out. Begin by setting all the equations for strain 2 equal to zero, the competitive-exclusion equilibrium should be found.

Let:

$$Y_1 > 0, \quad Y_2 = 0; \quad V_1 > 0, \quad V_2 = 0; \quad Z_1 > 0, \quad \text{and} \quad Z_2 = 0$$

In addition, if strain 2 is out-competed, then $\dot{V}_2 = 0$, so

Competitive exclusion of strain $j$ can only occur if no other strain mutates in to it, i.e., $m_{ji} = 0 \quad \forall\, i \neq j$.

Biologically, these conditional make sense. Biological fitness is a term used to describe an organism's ability to persist and reproduce. Viruses require invading and seizing a host cell's protein manufacturing abilities to replicate, while avoiding the immune system long enough to accomplish this. The more efficient a virus is at evading the immune system, the higher its fitness value. Therefore, very similar viruses would have very similar fitness values, and thus be equally evasive. However, if one virus is more fit than the other, the fitness of the first virus will be higher than the fitness of the second virus, the first would be able to replicate for a much longer period of time than the second virus, which would be eliminated by the immune system.

From the analysis of the cellular level model we are able to say many things. As long as $\Re_0 < 1$ the immune system is able to lower the rate of production of secondary infections. This means that the disease will die out. When $\Re_0 > 1$ we have three possible cases. In the first case we see that multiple viruses are able to infect the host and coexist. The second case shows the situation in which the host's immune system does not respond to the infected cells. This is biologically meaningful when the host's immune system is severely compromised like in AIDS patients. In the third case we see that the only situation in which one strain will out compete the other strain is when there is no mutation. This would happen because the immune system would select against the strain with lower fitness. Ultimately the evolution of the influenza virus is driven by the virus's mutation and selection by the immune system.

## 2.6 Stochastic Model

The consideration of the deterministic version of this model is especially useful for analysis and consideration of the average behavior of the interaction between healthy cells, cells infected by influenza, free influenza virus and Cytotoxic T Lymphocytes. A stochastic interpretation of this behavior is useful in the consideration of the distribution events.

This approach is useful biologically for many reasons because biological processes do not occur at identically distributed intervals. Also, it gives insight into the possible outcomes of these interactions, or the distribution of cellular events during different infections. Varying the parameters, gives information about the distribution of cellular events over

different strains (by varying virus and infected cell parameters) and different hosts (by varying the CTL parameters). Using this simulations, we can look at the effect of the mutation matrix on the course of infection.

The behavior was modelled stochastically, and analyzed through simulations that were written as continuous time Markov chain models [1]. The simulations were run over multiple realizations with varying parameters. The cellular level model can be characterized with stochastic effects by Equation (9) [1],[9]

The cellular level has $3n + 1$ classes including: One susceptible class, $X$, which corresponds to the healthy cell population; $n$ infected classes, $Y_i$, which correspond to cells infected by each viral strain $i$; $n$ infective classes, $V_i$, which correspond to the viral load of strain $i$; and $n$ CTL classes, $Z_i$, which corresponds to the population of Cytotoxic T Lymphocytes acting against cells infected with viral strain $i$. There are, not including initial conditions, $n^2 + 3n + 6$ parameters: $\Lambda$ is the constant rate of birth of healthy cells; $d$ is the proportional rate of natural death of healthy cells; $\overrightarrow{\beta}$ which is a $1 \times n$ array of infection rates of each free virus; $a$, the natural death rate of infected cells; $p$, the rate that CTL's kill infected cells; $\overrightarrow{k}$, which is a $1 \times n$ array of replication rates of each strain within an infected cell; a matrix $m$ of mutation rates, where $m_{ji}$ is the rate of mutation from strain $j$ to strain $i$; $\mu$ is the clearance rate of free virus; $\overrightarrow{c}$ is the strain specific birth of CTL in response to infected cells; and $b$ is the natural loss of productivity of CTL's.

Each simulation was run over 25 realizations, with $1.5 \times 10^6$ time steps. The parameter $\mu$ was estimated as 3.1 days$^{-1}$ from an e-mail correspondence with Alan Perelson[19] $X_0$ was set at $10^6$ cells. $\Lambda$ and $d$ were set at $5 \times 10^5$ cells $\times$ days $^{-1}$, and .5 days $^{-1}$ respectively, so that $X_0$ was at disease free equilibrium. All other parameters were set constant except for $\overrightarrow{k}$, $\overrightarrow{\beta}$, and $\overrightarrow{c}$. $\overrightarrow{k}$ and $\overrightarrow{\beta}$ were chosen from uniform random distributions such that all of the decoupled $\Re_{i0}$ would be significantly greater than one. Using the bounds of those distributions, $\overrightarrow{c}$ was selected such that endemic equilibria would exist under all conditions for each strain. Selecting each $\beta_i \& k_i$ from a uniform distributions for each realization corresponds to the varying ability of free virus strains to infect healthy cells and replicate. Choosing each $c_i$ from a uniform random distribution in each realization corresponds to variability in the hosts immune response time. Two simulations were run over 25 realizations. One used a tridiagonal mutation matrix with the main diagonal equal to .79618 and the upper and lower subdiagonals .10191 with wrapped boundaries which were estimated from The Encyclopedia of Virology [10]. The other simulation used matrices with entries chosen from a uniform random distribution and then normalized so that the sum of each row is one. A counter variable was included in the programs to count infection events by strain. From that we determined the final number of total infections produced by each strain during each simulation.

The results of the simulations include the total number of cells infected by each strain during each realization. For each strain, the mean, standard deviation and median of its total infection size were calculated for the 25 realizations. These values were averaged over all strains to discern the difference from the mutation matrices used in the two simulations.

| Realization | Mean | StDev | Median | Realization | Mean | StDev | Median |
|---|---|---|---|---|---|---|---|
| 1 | 89833 | 72860 | 51845 | 1 | 31364 | 15320 | 30804 |
| 2 | 89025 | 59653 | 70738 | 2 | 34899 | 15862 | 33504 |
| 3 | 86898 | 65190 | 64376 | 3 | 27881 | 19597 | 23432 |
| 4 | 89152 | 51798 | 80593 | 4 | 30063 | 14344 | 29541 |
| 5 | 90139 | 61628 | 69737 | 5 | 28545 | 12458 | 24804 |
| 6 | 89365 | 53329 | 84295 | 6 | 27919 | 19680 | 17932 |
| 7 | 88609 | 70391 | 60996 | 7 | 31902 | 12693 | 35935 |
| 8 | 90330 | 67642 | 70215 | 8 | 27008 | 18674 | 17123 |
| 9 | 90230 | 53039 | 64619 | 9 | 32886 | 18277 | 32377 |
| 10 | 89458 | 80660 | 56073 | 10 | 37073 | 14415 | 35432 |
| 11 | 90794 | 64054 | 78210 | 11 | 27156 | 19712 | 27893 |
| 12 | 90885 | 40065 | 85467 | 12 | 25612 | 17012 | 22585 |
| 13 | 89744 | 67532 | 83929 | 13 | 33806 | 18013 | 28349 |
| 14 | 90798 | 78766 | 43387 | 14 | 36407 | 18347 | 34774 |
| 15 | 90596 | 78593 | 60164 | 15 | 29756 | 18136 | 28431 |
| 16 | 90110 | 63754 | 66061 | 16 | 32156 | 18634 | 28179 |
| 17 | 90117 | 54239 | 59644 | 17 | 35072 | 19840 | 27269 |
| 18 | 89909 | 66149 | 84799 | 18 | 24692 | 12369 | 19424 |
| 19 | 86477 | 56290 | 66031 | 19 | 36310 | 14008 | 37330 |
| 20 | 90817 | 56643 | 83634 | 20 | 32367 | 21308 | 30462 |
| 21 | 88778 | 70507 | 50510 | 21 | 30837 | 14424 | 34078 |
| 22 | 84476 | 60525 | 64820 | 22 | 34994 | 9819 | 32536 |
| 23 | 86011 | 72221 | 56283 | 23 | 29108 | 15402 | 24313 |
| 24 | 90746 | 59605 | 77632 | 24 | 31078 | 13247 | 29702 |
| 25 | 89438 | 63988 | 74125 | 25 | 29295 | 20307 | 18447 |

Table 3: Descriptive statistics of realizations over strains: Random Mutation Matrix (R), Tridiagonal Mutation Matrix (L)

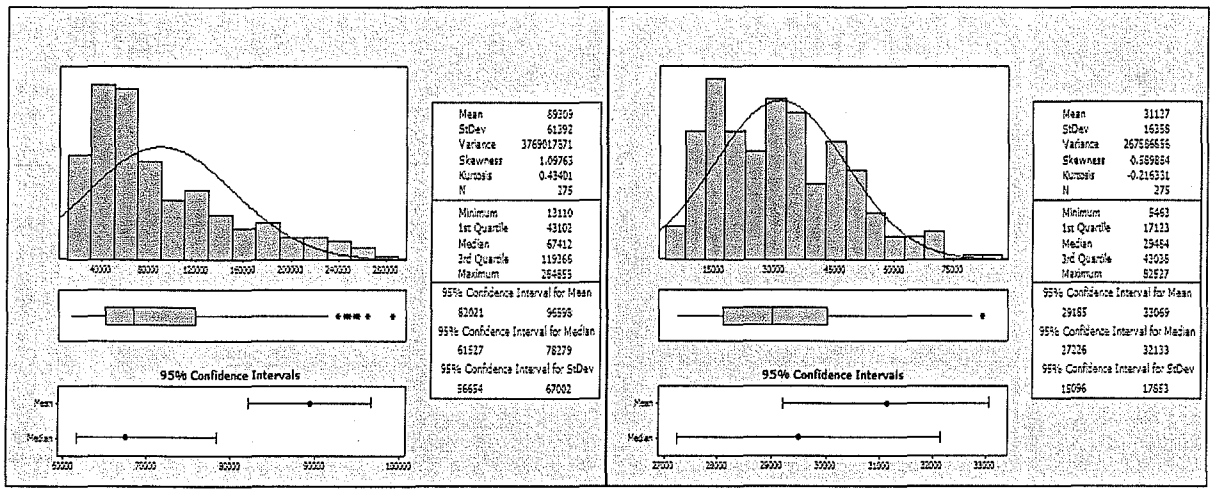



Figure 6: Descriptive statistics for total number of infected cells of each strain during each realization: constant, random mutation matrix (R), tridiagonal mutation matrix (L)

The simulations with the constant mutation matrix had strain means varying from 25,617 to 37740 with a mean of 31,127. The standard deviations varied from 12,936 to 19,426 with a mean of 16,123. The medians varied from 22,441 to 36,357 with a mean of 28,886. The simulations with a random mutation matrix had means varying from 40,377 to 201,626 with a mean of 89,309. The standard deviations range from 16,849 to 57,822 with a mean of 33,716. The median varied from 32,555 to 223,354 with a mean of 90,515. It is worthwhile to note that the 95% confidence intervals for the mean of the means and standard of deviations do not intersect.

The random mutation matrix produced larger final infection counts and final infection count variability amongst strains. This implies that the immune system is most effective when the virus has little variability. It also implies that viral mutation is an effective evasive strategy in the presence of cytotoxic T lymphocytes.

# 3 Phylogenetic Evolution

## 3.1 Phylogenetic Trees

A phylogenetic tree is a graphical representation depicting the ancestor-descendant relationship between organisms by the examination of differences in nucleic or amino acid sequences. The descendants are located at the tips of the tree branches, and splits between the branches can be traced back to their unobservable ancestor [11]. This allows for reconstruction and comparison of the evolutionary history of species which are among the most important biological and genetic subjects. Phylogenetic trees can shed light on which of

an organism's features are changing rapidly and allows for the calculation of residues that provide evidence of natural selection operating on the species.

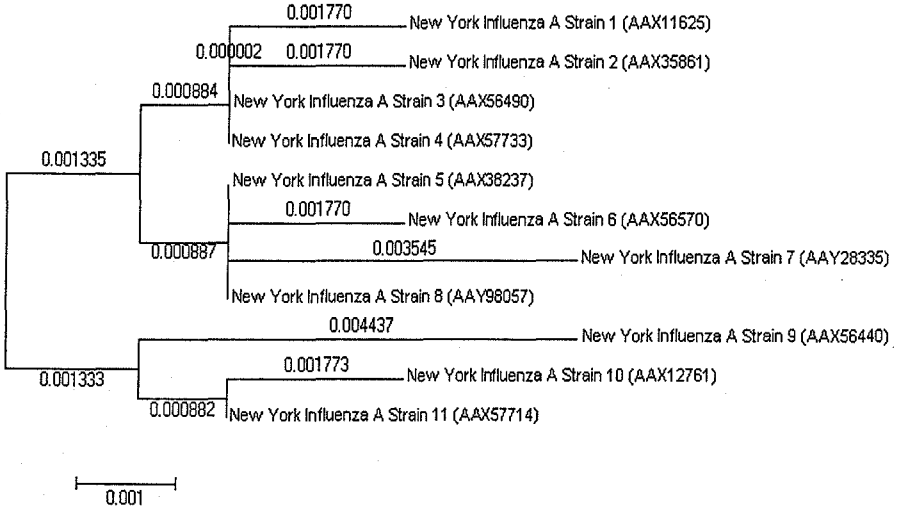



Figure 7: Influenza A $H3N2$ phylogenetic tree for New York during 2002

Influenza A viruses consist of eight RNA segments wrapped in a protein coat. Of the eight RNA segments, two code for proteins known as hemagglutinin (HA) and neuraminidase (NA), which aid in the invasion of cells. These proteins become the major antigens which the immune system recognizes, and thus they are used to name influenza strains. For example, Influenza A type H3N2 carries a type three hemagglutinin and type two neuraminidase. Studies have indicated that production of antibodies against neuraminidase are rare due to the small size of NA, so the immune response depends mostly on antibodies against HA [2]. Therefore, a mutated strain will refer to changes occurring in hemagglutinin only, and all phylogenetic trees to follow will track amino acid changes in the HA protein among influenza A viruses.

## 3.2 Statistical Methods and Algorithms

Several different methods exist for creating and testing the reliability of a phylogenetic tree. The Neighbor-Joining method was chosen since it does not require the assumption of a constant rate of evolution, and it is relatively fast computationally. The first step in the algorithm is converting the DNA or protein sequences into a distance matrix that represents the evolutionary distance between sequences. The NJ method is known to do well with data that have diverged recently. This assumption holds true for influenza since yearly data for influenza from within the last decade is being considered [11]. All phylogenetic trees were created by using Molecular Evolutionary Genetics Analysis (MEGA),

version 3.0 [13], from sequences obtained from the National Centers for Biotechnology Information (NCBI) Influenza Virus Resource [17] and the Los Alamos National Laboratory (LANL) Influenza Sequence Database [15].

It is important to consider how strongly the data supports the relationships depicted in the phylogenetic trees. One of the most commonly used tests of the reliability of an inferred tree is the bootstrap test, which is evaluated using the bootstrap resampling technique. Consider $m$ sequences, each with $n$ nucleotides or amino acids, a phylogenetic tree can be reconstructed. From each sequence, $n$ nucleotides are randomly chosen with replacements, giving rise to $m$ rows of $n$ columns each. These now constitute a new set of sequences. A tree is then reconstructed with these new sequences using the same tree building method as before. Next the topology of this tree is compared to that of the original tree. Each interior branch of the original tree that is different from the bootstrap tree the sequences it partitions is given a score of 0; all other interior branches are given the value 1. This procedure of resampling the sites and the subsequent tree reconstruction is repeated several hundred times, and the percentage of times each interior branch is given a value of 1 is noted. This is known as the bootstrap value. In the phylogenetic trees produced for the data set, the number of replicates were specified as well as the seed for the pseudorandom number generator. In each bootstrap replicate, the desired quantity is estimated and the standard deviation of the original values are computed.

### 3.3 Phylogenetic Distance Coupled to Epidemiology

Global spatio-temporal strain dynamics determine phylogeny. By utilizing the genetic distances obtained from the phylogenetic trees, the effects of immune system selection at the cellular level can be observed at the population level by incorporating these distances into an epidemiological S-I-R model. This allows for the spread of the various strains to be tracked through a population of susceptible individuals. Each strain creates a separate infective class, which leads to a corresponding resistant class. It should be noted that some degree of immunity toward every strain is incurred by resistance to any strain. However, for simplicity, this variable cross-immunity will be neglected and left for future versions of the model. Instead, immunity will be considered to be complete and cover all strains.

## 4 Statistical Analysis

In this section, influenza outbreaks across the globe over a three-year period from 2002 to 2004 inclusive are considered. Studies have indicated that influenza A with type 3 hemagglutinin ($H3$) evolve more rapidly than those bearing type 1 ($H1$) [3]. Of those carrying type 3 hemagglutinin, strain $H3N2$ is most prevalent in outbreaks in humans [4]. Influenza B evolves more slowly than either A subtype, and influenza C accounts for only a small percentage of outbreaks [3].

## 4.1 Global Data for Influenza A $H3N2$ in the year 2002

When doing any kind of statistical analysis, the question of normality must be addressed. The normal distribution is one that appears in a variety of statistical applications. One reason for this is the central limit theorem. This theorem states that the sums of random variables are approximately normally distributed if the number of observations is large. If the data are found to be normally distributed, then many different kind of statistical procedures, such as the t-test, can still be used. The global data for 2002 obtained from the World Health Organization (WHO)[24] is presented in the Appendix. Figure 7 shows a summary of the this data with four graphs: histogram of data with an overlaid normal curve, boxplot, 95% confidence intervals for $\mu$, and 95% confidence intervals for the median.

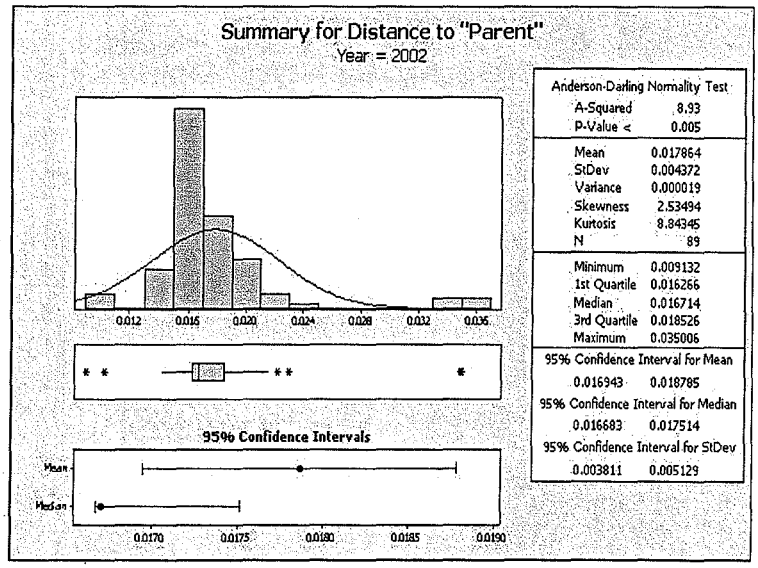



Figure 8: Summary Statistics for 2002

This graphical summary also displays a table of the Anderson-Darling Normality Test, descriptive statistics and confidence intervals for mean $\mu$, standard deviation $\sigma$, and the median. To test for normality, a conservative $\alpha$ is chosen, in this case 0.05. The data are normal if p-value $> \alpha$. Since p-value $< 0.005 < \alpha = 0.05$, the null hypothesis ($H_0$) is rejected, so there would appear to be a problem with normality. This can be seen in the histogram of data with an overlaid normal curve where it does not really fit.

The data appears to follow non-normal distribution, and Figure 8 shows a histogram of the 2002 data overlaid with a gamma distribution curve. The gamma-distribution parameter ($\gamma$), or shape parameter of, the curve can be interpreted biologically as an estimate of the transfer rate $\beta$ of the virus as it spreads through a population. In addition, the frequency shown on the histogram indicates that large genetic mutations are usually rare events. This is further supported by the fact that antigenic shifts (that is, the appearance

of totally new hemagglutinin types) are rare events [12].

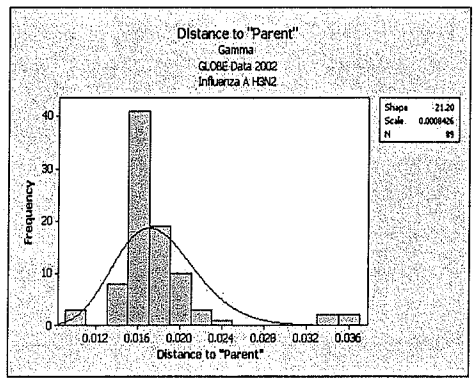



Figure 9: Global Data overlaid with Gamma Distribution 2002

The data appears to follow a gamma distribution, implying the waiting times between Poisson distributed events are relevant. Biologically this makes sense because rates are most likely heterogeneous, and highly dependent upon biochemical properties of the antigens and antibodies which must be accounted for at the population level. This is especially true for creating phylogenetic trees for viral strains that mutate at different rates in different locations along the antigen sequence.

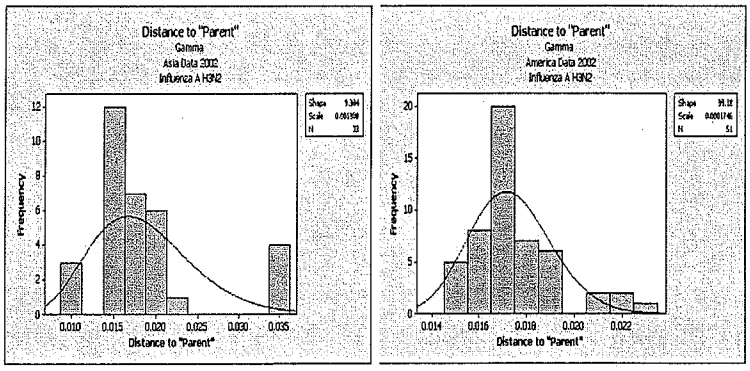



Figure 10: Regional Data overlaid with Gamma Distribution 2002

On a regional level (Figure 9) this can be observed, because the immunological histories of hosts differ due to biological and socio-economic factors. For instance, all subtypes of

influenza A viruses have been found in aquatic birds in China, where current agricultural practices put large numbers of people in close proximity to farm animals, including ducks and pigs. These conditions favor the generation and spread of new viral strains and provide a possible medium (pigs) for the genetic reassortment of human and avian viruses [20]. In fact the serotype under study in this paper ($H3N2$) originally appeared as an antigenic shift that resulted in a pandemic in 1968 – 1969 known as the "Hong-Kong Flu." This pandemic was responsible for 34,000 deaths in the United States of America alone.

In the phylogenetic tree approach, interest lies in identifying the most distant strand from its "parent". One way to identify outliers is the plot of residuals versus fits. This plot shows a random pattern of residuals on both sides of 0. If a point lies far from the majority of the rest of the points, it may be an outlier. In addition, there should not be any recognizable patterns in the residual plot. The residual plot shown in Figure 10 may indicate error that is not random.

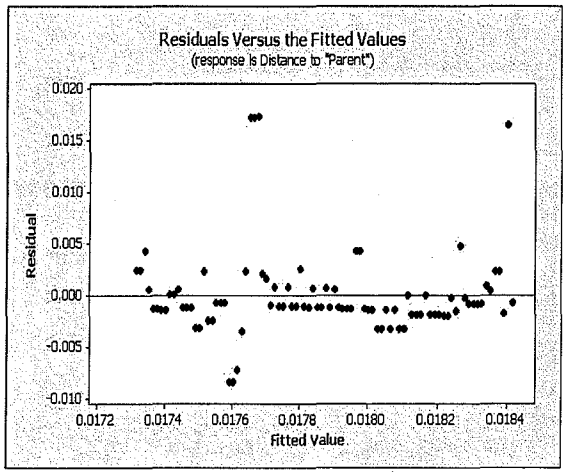



Figure 11: Residuals versus the Fitted Values 2002

It can be readily seen that there is no discernible pattern and thus no problem with linearity or constant variance. More interesting, though, is how evident the outliers are in the positive direction compared to the rest of the data points. There appear to be around four points that are extreme in the positive case, which coincides with the observable distance from the phylogenetic trees produced.

## 4.2 Influenza A $H3N2$ for the year 2003

Similar statistical analysis as done above can be employed on other data with similar results. Figure 11 summarizes the statistics for data from the year 2003 (also given in the

Appendix). Figures 12 overlays a gamma distribution onto the histograms for the global data. Figure 13 shows the data for the American and Asian regions. Figure 14 shows the residuals for the global data.

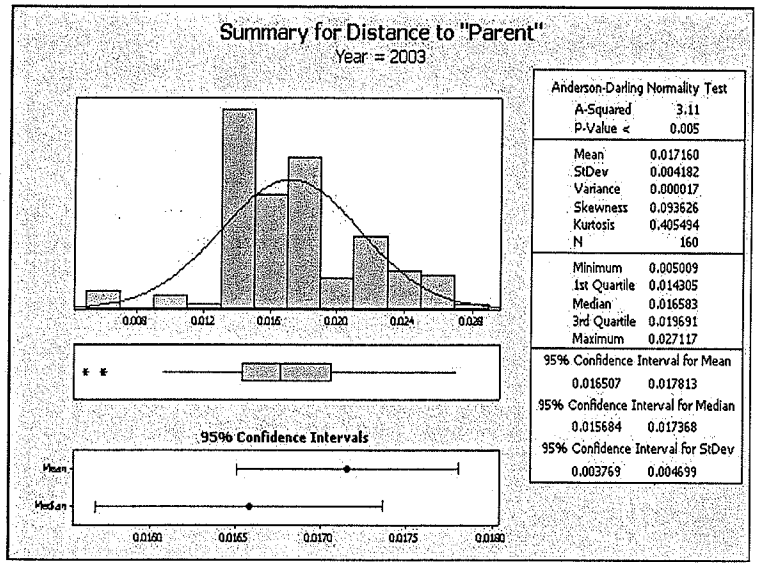



Figure 12: Summary Statistics for 2003

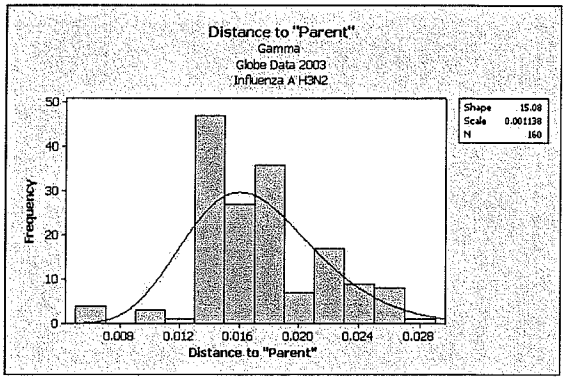



Figure 13: Global Data overlaid with Gamma Distribution 2003

It should be noted that other regions, including Africa, Europe, and Oceania, also reported influenza cases. However, since the number of unique strains was less than ten,

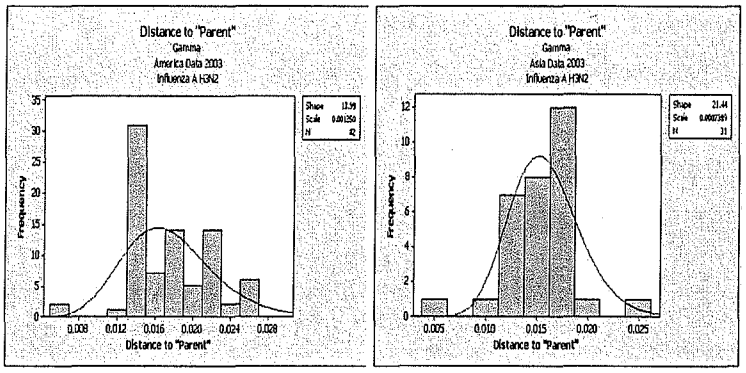



Figure 14: Regional Data overlaid with Gamma Distribution 2003

these regions were not included in this study, but this data are given in the Appendix. This brings up a very important biological fact, since hosts in different regions of the world have encountered different strains of influenza, both during the years studied as well as during previous years. Therefore, people in different regions would have different antibodies, and therefore would have different immunological histories. At the population level, this geographical isolation of strains should be considered using a weighting factor $b_i$ to account for variations in immunological history.

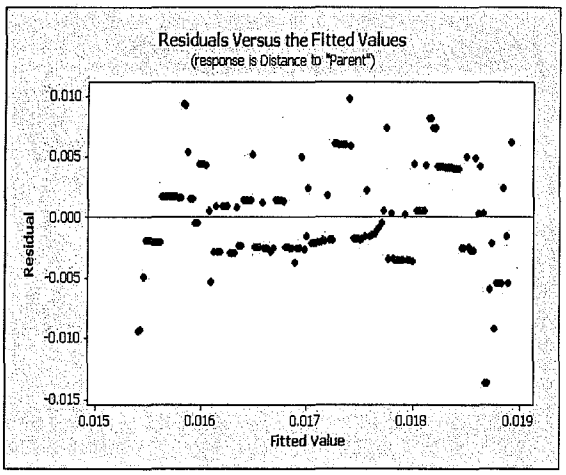



Figure 15: Residuals versus the Fitted Values 2003

## 4.3 Influenza A $H3N2$ for the year 2004

North America, specifically the United States of America, did not report more than ten unique strains during 2004. For this reason North America was not included in the regional study for that year. Data obtained from the Centers for Disease Control and Prevention (CDC) indicated that influenza activity occurred at low levels from October to mid-December, steadily increased during January and peaked in mid-February during the 2004-05 U.S. season. [4]

Figure 15 shows the global data (also given in the Appendix) statistics for 2004, Figure 16 displays regional data for Oceania and Asia, and Figure 17 displays the residuals for the global data.

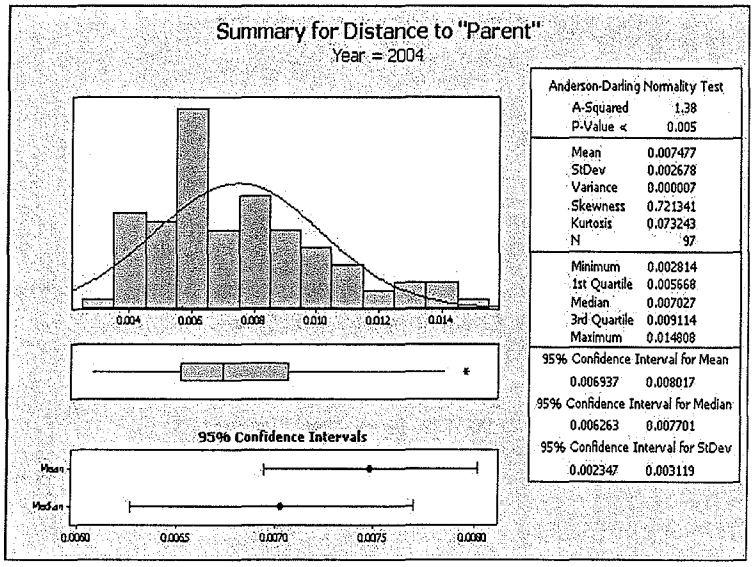



Figure 16: Summary Statistics for 2004

## 4.4 Summary of Statistical Analysis

This analysis has employed standard statistical techniques to explain the interaction and underlying factors of data collected over a three period for Influenza A $H3N2$. Among things tested for were normality, perceived distribution and outliers. It was found that none of the data sets were normal with an alpha level of 0.05. The data did follow a gamma

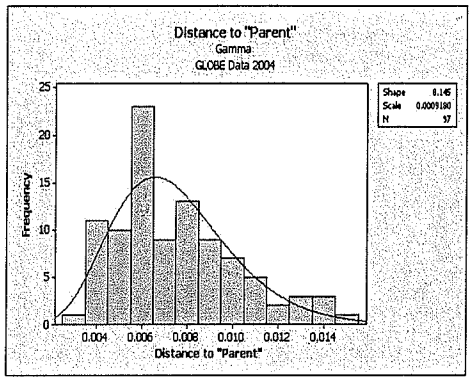



Figure 17: Global Data overlaid with Gamma Distribution 2004

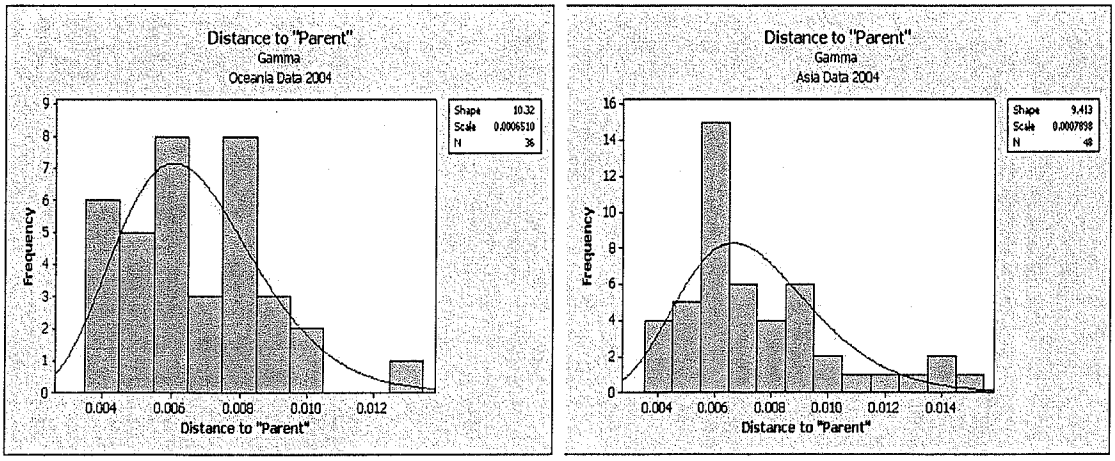



Figure 18: Regional Data overlaid with Gamma Distribution 2004

distribution better than a normal or exponential, which coincides with the biological aspect of the data, specifically the stochasticity of waiting times between Poisson distributed events.

## 4.5 Most Prevalent Strains per Region per Year

Having presented the statistical data for a three-year period $2002 - 2004$, the attention is returned to the phylogenetic trees and the strain distance from an ancestral strain. Having established a phylogenetic tree for 2002, and implementing the weighted distances, somewhere among the estimates should contain the dominating strain for the next year, 2003.

Considering the 10 most distant strains of 2002, the most prevalent ($H3N2$) strain

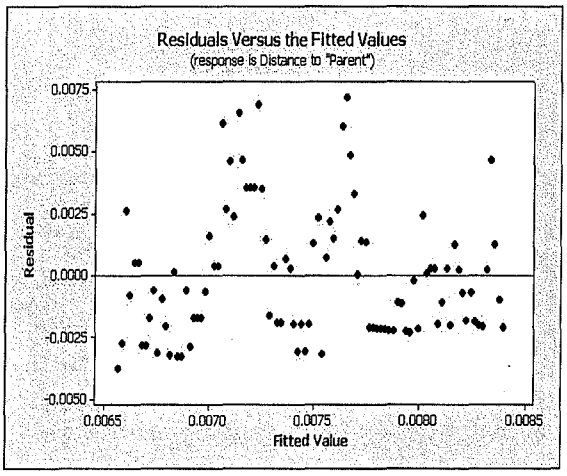



Figure 19: Residuals versus the Fitted Values 2004

for 2003 was found to be $A/Fujian/411/2002$ which accounted for 88.8% of the antigenic characterization of viral isolates [24]. The top four strains were from the Republic of Korea, where its geographical location and socialized health care most likely kept those strains from spreading. Positions five, six and seven differ from $A/Fujian/411/2002(H3N2)$ within a hundredth of a decimal place. This discrepancy may also be attributed to how the trees were constructed and the fact that no true tree exists, but rather estimations of them are produced. So the hypothesis that phylogenetic distance can be an estimator of dominance was validated. Turning our attention to the 2003 season, 22% of the 709 influenza A $H3N2$ isolates were antigenically similar to $A/Wyoming/3/2003$ [24], but a new strain emerged as the dominant one in 2004: $A/California/7/2004$ [24]. However this strain has not yet been sequenced and therefore could not be included in the analysis nor the construction of the phylogenetic tree. In this estimation, a snapshot of the influenza persistence was captured, yet estimations were hindered due to lack of information and access to all strains. Similarly, the estimation of the most influential strain for the 2005 season may be skewed due to lack of the sequencing of $A/California/7/2004$. However, a list of possible dominant strains was developed based on the 2004 analysis. Figure 20 shows a radial phylogenetic tree of the estimation for 2004.

# 5 Population-Level Model

The cellular level model allowed for the selection of single influenza virus by the immune system to be observed. This underlying mechanism is an important factor in understanding the evolution of the virus over time. The phylogenetic trees allow for the determination of

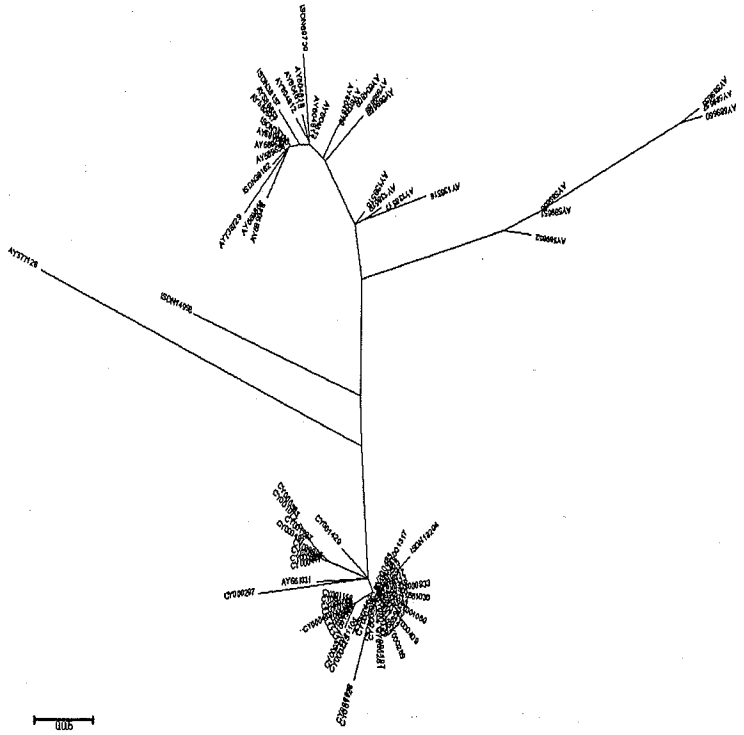


Figure 20: Radial Phylogenetic Tree of Global 2004 Data

genetic distances among different strains. Now that the in-host dynamics are understood and the genetic distances have been determined, it is possible to examine the epidemiological effects the viral evolution as the strains spread through a population of susceptible individuals.

## 5.1 The S-I-R Model

Consider $S(t)$ as a population of susceptible hosts at time $t$. The population can be kept constant for simplicity, that is the birth rate equals the death rate (both are $\mu$), and the same death rate is used for every class. The susceptibles become infected at a strain-specific rate $\beta_i$, and enter an infective $I_i(t)$ class for strain $i$. Infected individuals recover at a strain-specific rate $\gamma_i$, and enter a resistant class $R_i(t)$.

Table 4: Variables for S-I-R Population Model

| Variables at Population Level | Description |
|---|---|
| $S(t)$ | Susceptible hosts at time $t$ |
| $I_i(t)$ | Hosts infected with strain $i$ at time $t$ |
| $R_i(t)$ | Hosts recovered from strain $i$ infection at time $t$ |

Table 5: Parameters for S-I-R Population Model

| Parameters at Population Level | Description |
|---|---|
| $\mu$ | Birth and death rates of hosts per capita |
| $\beta_i$ | Rate of infection of susceptible hosts from strain $i$ |
| $\gamma_i$ | Recovery rate of infected hosts from strain $i$ infection |

$$\begin{cases} \dot{S} = & \mu N - \left[\sum_{k=1}^{n} \beta_k \frac{SI_k}{N}\right] - \mu S \\ \dot{I}_i = & \beta_i \frac{SI_i}{N} - \gamma_i I_i - \mu I_i \\ \dot{R}_i = & \gamma_i I_i - \mu R_i \end{cases} \tag{4}$$

In order to mimic the immune system selection at the cellular level, $\beta_i$ must be altered in such a way that the transmission rates incorporate the genetic distance of strain $i$ from the ancestral strain, which is represented as $d_i$. This can be achieved by considering the genetic distance of strain $i$ versus all strains, as shown in Eqn. (5). The constant $b_i$ is a strain-specific birth rate, included to reflect the relative fitness of the different viral strains. However, the ratio of genetic distances alone does not determine which strain dominates during a particular season. In fact, the histograms presented in Section 4 indicate that the dominant strain during a given year is often only a moderate genetic distance from the dominant strain of the previous year. This is likely a reflection of the effects of the positions of mutations in the amino acid sequence.

Research has shown that antibodies toward hemagglutinin target five particular antigenic sites comprised of 131 amino acids combined [14]. However, various searches on the LANL and NCBI Influenza Databases [15],[17] returned hemagglutinin chain lengths often exceeding 1700 amino acids. As a result, significant genetic distances alone do not necessarily reflect significant antigenic changes. In addition, certain amino acid substitutions may not significantly change the three-dimensional shape of the antigen [2]. These are known as synonymous mutations and do not change the antigenicity of the hemagglutinin protein, even when the mutation occurs within one of the five antigenic sites. In order to incorporate these important biological concepts, it is necessary to weight the genetic distance $d_i$ values to convert them into antigenic distances using strain-specific weighting constants $\omega_i$. These weighting constants are defined as some currently unknown function

relating the genetic distance to the antigenic distance of strain $i$ from the unknown ancestral strain. A higher $\omega_i$ indicates a mutation within one of the five antigenic sites, while a lower $\omega_i$ represents either a mutation outside of these sites or a synonymous mutation. An $\omega_i = 0$ would represent a synonymous mutation (one which has no effect on the antigenic structure).

$$\beta_i = b_i \left[ \frac{\omega_i d_i}{\sum_k \omega_k d_k} \right] \tag{5}$$

where $\omega_i = f(\text{antigenic distance, genetic distance } (d_i))$

Figure 21 shows the diagram that corresponds to the host-level population model.

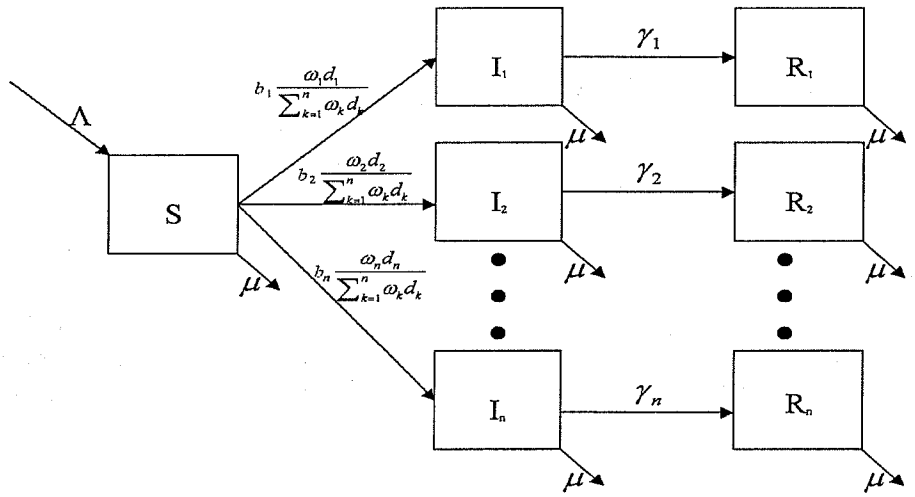



Figure 21: S-I-R Population Model

Two equilibria exist for the population-level model: the disease-free equilibrium and competitive exclusion equilibrium.

## 5.2 Disease-Free Equilibrium and Calculation of $\Re_0$

The disease-free equilibrium (DFE) occurs when no influenza strains are circulating through the population. In order to calculate the DFE for the population-level model,

$$\begin{aligned} \text{Let } I_i = 0 \quad & \forall i \\ & \Longrightarrow 0 = \mu N - \mu S \\ & \Longrightarrow S^{\infty} = N \\ & \Longrightarrow R_i = 0 \end{aligned}$$

The DFE exists at $(N, \overrightarrow{0}, \overrightarrow{0}) \quad \in \quad \mathbb{R}^{2n+1}$.
This can be used to calculate the basic reproductive rate $\Re_0^i$. Begin by considering the $\dot{I}_i$ equation:

$$\dot{I}_i = I_i \left[ \beta_i \frac{S}{N} - \gamma_i - \mu \right]$$

Linearizing about the DFE (N,0,0), this reduces to:

$$\frac{\partial \dot{I}}{\partial I_i} = \beta_i - (\gamma_i + \mu)$$

$$\text{Therefore, } \Re_0^i \equiv \frac{\beta_i}{\gamma_i + \mu}$$

From this expression for $\Re_0^i$, it is easy to see that the population of hosts infected with strain $i$ will increase from any non-zero initial size as long as $\beta_i > \gamma_i + \mu$, and it will decline if $\beta_i < \gamma_i + \mu$. Note that, since $\beta_i$ and $\gamma_i$ are strain-specific parameters, each strain has its own $\Re_0^i$.

## 5.3 Competitive-Exclusion Equilibria

If selection by the immune system does occur, then it is expected that one strain will dominate over any other competing strain. This is known as competitive-exclusion. Consider the case of two influenza strains $i$ and $j$:

Suppose $I_i > 0$ and $I_j = 0$ where $j \neq i$.

$$\begin{aligned}
\text{Then } 0 &= I_i^\infty \left[\beta_i \frac{S^\infty}{N} - \gamma_i - \mu\right] \\
\Leftrightarrow S^\infty &= \frac{N}{\beta_i}(\gamma_i + \mu) \\
\Rightarrow \frac{S^\infty}{N} &= \frac{1}{R_{i0}} \\
\Longrightarrow 0 &= \mu N - \left(\frac{\beta_i I_i^\infty}{N} + \mu\right) S^\infty \\
\left(\frac{\beta_i I_i^\infty}{N} + \mu\right) S^\infty &= \mu N \\
\left(\frac{\beta_i I_i^\infty}{N} + \mu\right) &= \frac{\mu N}{S^\infty} \\
\left(\frac{\beta_i I_i^\infty}{N}\right) &= -\mu + \frac{\mu N}{S^\infty}
\end{aligned}$$

Which implies

$$I_i^\infty = \frac{\mu N}{\beta_i}\left[-1 + \frac{N}{S^\infty}\right] = \frac{\mu N}{\beta_i}\left[-1 + \Re_0^i\right]$$

Therefore the competitive-exclusion equilibrium for the two-strain ($n = 2$) population-level model occurs at:

$$\begin{aligned}
S^\infty &= N \\
I_i^\infty &= \frac{\mu N}{\beta_i}\left[-1 + \frac{N}{S^\infty}\right] = \frac{\mu N}{\beta_i}\left[-1 + \Re_0^i\right] \\
I_j^\infty &= 0 \quad j \neq i \\
R_i^\infty &= \frac{\gamma_i}{\mu} I_i^\infty = \frac{\gamma_i}{N}\beta_i\left[-1 + \Re_0^i\right] \\
R_j^\infty &= \frac{\gamma_i}{\mu} I_j^\infty = 0
\end{aligned}$$

Since the competitive-exclusion equilibrium exists, one strain can dominate during an outbreak at the population level.

## 5.4 Coexistence Equilibrium

A coexistence equilibrium occurs when multiple strains can circulate through the population alongside one another. In order to determine the existence of a coexistence equilibrium

for two strains, it is necessary to set $\dot{I}_i = 0$ and $\dot{I}_j = 0$, as such:

$$\text{Suppose } i \neq j$$

$$0 = I_i \left[ \beta_i \frac{S}{N} - \gamma_i - \mu \right]$$

$$0 = I_j \left[ \beta_j \frac{S}{N} - \gamma_j - \mu \right]$$

Solving for S in each equation, an endemic equilibrium exists if and only if the following conditions on S are true:

$$S = \frac{N(\gamma_i + \mu)}{\beta_i}$$

and

$$S = \frac{N(\gamma_j + \mu)}{\beta_j}$$

However, both of these requirements are satisfied if and only if $\Re_0^i = \Re_0^j = \frac{\beta_i}{\gamma_i + \mu}$. This implies that the two strains reproduce at the same rate. For example, this can occur if $\beta_i = \beta_j \ \forall \ i, j$ and $\gamma_i = \gamma_j \ \forall \ i, j$, meaning that strains $i$ and $j$ have identical transmission and recovery rates. This would mean that the strains are exactly alike (that is, $I_1^\infty = I_2^\infty = I_3^\infty = \cdots = I_i^\infty$), which is the case when the strains differ by synonymous mutations. When the strains differ more significantly and the basic reproduction rates of the strains are different, then the conditions on S are not satisfied, and coexistence between the two strains is again not possible. This indicates that viral fitness manifests itself at the population level as the strains' ability to remain viable outside of the body long enough to be transmitted and the length of time which the strain can infect and be spread to other hosts before the immune system of the infected host eliminates it. That is, competitive exclusion will occur if $\Re_1 > 0$ and $\Re_2 > 0$ but $\Re_1 \neq \Re_2$. Since the strains have different basic reproductive rates, then the fitness of the strains would also differ. Therefore, the strain with lower basic reproductive rate would be less fit. Such fitness differences are often targeted by natural selection, and this is also the case with immune system selection. The less-fit strain would be selected against by the immune system, so the more-fit strain will exclude it due to its competitive advantage.

From the determination of these equilibria, the phase portrait shown in Figure 21 can be determined for the two-strain population-level model.

## 5.5 Single-Outbreak ($\mu = 0$)

Influenza outbreaks usually occur during colder seasons. Therefore it is necessary to observe the epidemiological dynamics during a single epidemic season using a single outbreak S-I-R model. A single outbreak S-I-R model considers a small enough time scale that the

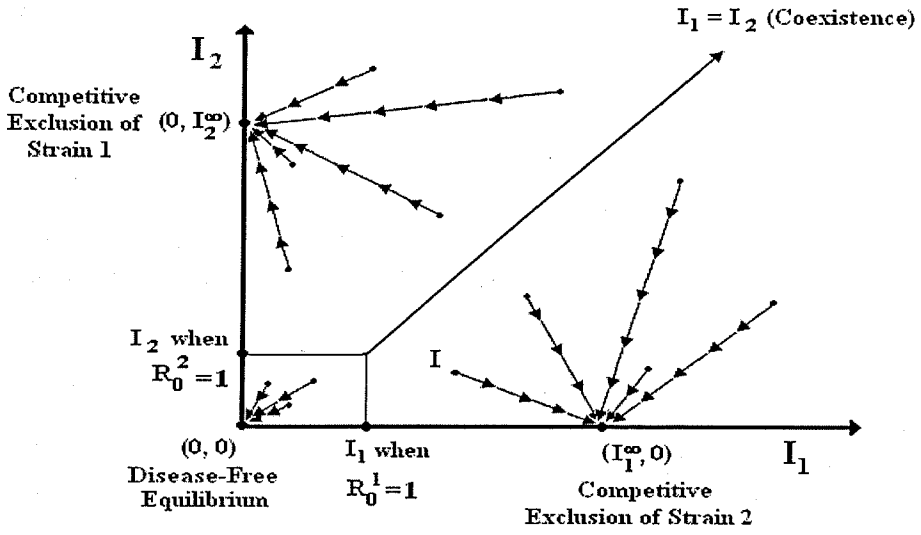



Figure 22: Population phase portrait for two strains

birth and death rates (both equaling $\mu$) are negligible, which reduces the S-I-R model in Eqn. (4) to the one shown in Eqn. (6).

$$\begin{cases} \dot{S} &= -\left[\sum_{k=1} \beta_k \frac{SI_k}{N}\right] \\ \dot{I}_i &= \beta_i \frac{SI_i}{N} - \gamma_i I_i \\ \dot{R}_i &= \gamma_i I_i \end{cases} \tag{6}$$

For a particular strain $i$, the basic reproductive rate $\Re_0$ during a particular outbreak can be calculated as such:

$$\begin{aligned} & \beta_i \frac{S(t)}{N} - \gamma_i > 0 \\ \Longleftrightarrow & \frac{\beta_i}{\gamma_i} \frac{S(t_0)}{N} > 1 \\ & \Re_0^i \equiv \frac{\beta_i}{\gamma_i} \frac{S(t_0)}{N} \end{aligned}$$

In this case, $\Re_0^i$ tells the outcome of strain $i$ during the outbreak. If the $\Re_0^i < 1$, then strain $i$ will die out and the outbreak. However, if $\Re_0^i > 1$, then strain $i$ will begin an

epidemic. In addition, the $\Re_0^i$ allows for the following asymptotic behavior to be inferred:

$$(1) \lim_{t\to\infty} S(t) = S^\infty$$
$$(2) \lim_{t\to\infty} I_i(t) = 0 \ \forall i$$
$$(3) \lim_{t\to\infty} R_i(t) = R_i^\infty$$

Hence, $R_i^\infty$ can be referred to as the prevalence of the $i^{th}$ strain during the outbreak. By finding this prevalence in stochastic simulations of influenza outbreaks, the dominant strain during the outbreaks can be determined.

## 5.6 Single-Outbreak Stochastic (Poisson) Model

The population level model is considered during a single outbreak without demographic rates. The parameters includes two $1 \times n$ arrays $\overrightarrow{\beta}$ and $\overrightarrow{\gamma}$. Each $\beta_i$ of $\overrightarrow{\beta}$ is a function of three arrays $\overrightarrow{d}, \overrightarrow{\omega} \& \overrightarrow{b}$. Each $d_i$ corresponds to the genetic distance to a parental strain, $\omega_i$ is a weighting constant that accounts for antigenic distance, and $b_i$ is a strain-specific viral reproduction rate. Adding stochastic effects to the deterministic model in Eqn. (6) yields Eqn (7) [1], [9].

$$\text{Prob}\{\triangle S(t) = k, \triangle I_i(t) = l_i, \triangle R_i(t) = m_i, 1 \le i \le n | S_i(t), I_i(t), R_i(t)\} =$$

$$\begin{cases} \frac{\beta_i S(t) I(t)}{N}\triangle t + o(\triangle t), \\ (k, l_1, \ldots, l_n, m_1, \ldots, m_n) = (-1, \overbrace{0, \ldots, 0}^{i-1}, 1, \overbrace{0, \ldots, 0}^{2n-i}) \\ \gamma_i I_i(t)\triangle t + o(\triangle t), \\ (k, l_1, \ldots, l_n, m_1, \ldots, m_n) = (\overbrace{0, \ldots, 0}^{i}, -1, \overbrace{0, \ldots, 0}^{n-1}, 1, \overbrace{0, \ldots, 0}^{n-i}) \\ 1 - \left[\sum_{j=1}^n I_j(t)\left(\frac{\beta_j S(t)}{N} + \gamma_j\right)\right]\triangle t + o(\triangle t), \\ (k, l_1, \ldots, l_n, m_1, \ldots, m_n) = (\overbrace{0, \ldots, 0}^{2n+1}) \\ o(\triangle t), \text{otherwise} \end{cases} \tag{7}$$

The distances from Section 4.4 are used in this simulation. Simulations were run for each of the four sets of New York phylogenetic data from 2000 through 2004 and three sets of global phylogenetic data from 2001 through 2005 by inputting the genetic distances as $\overrightarrow{d}$. The New York and global distances were run separately by year. All simulations chose $b_i$ from a uniform distribution, and $\overrightarrow{\gamma}$ was set equal to $\overrightarrow{1}$. With each data set, the simulation was run twice. The first simulation set $\overrightarrow{\omega} = \overrightarrow{1}$, which would mean that antigenic

distance is solely determined by genetic distance, while the second simulation includes different weighting constants for each strain. In the second simulation, the strain which had the smallest genetic distance received the highest weighting constant (to represent a mutation that causes a key antigenic change), while all other strains received weighting constants of 0.1 (to represent minor mutations).

When comparing the two simulations for New York from the years 2000 to 2004, it was found that the weighting constants $\omega$ do in fact make a significant difference for $\Re_0$, which are displayed in Table 5.

| Year | Different weights | Same weights |
|---|---|---|
| 2000 | 15.243 | 72.766 |
| 2001 | 25.379 | 2.8567 |
| 2002 | 7.342 | 2.9787 |
| 2003 | 6.305 | 2.8656 |
| 2004 | 10.267 | 13.344 |

Table 6: $\Re_0$ values for New York due to the variation of the weighting constant $\omega$

The weighting constant also has an impact on the final size. Table 6 shows that having the same weight values for all strains resulted in the final size being much larger than when the genetic weights were varied by strain.

| Year | Different weights | Same weights |
|---|---|---|
| 2000 | 1.7012 | 74.009 |
| 2001 | 2.7880 | 130.30 |
| 2002 | 0.9539 | 29.748 |
| 2003 | 0.7053 | 21.433 |
| 2004 | 19.662 | 53.253 |

Table 7: Means of final size resulting from the variation of the weighting constant $\omega$

These results are consistent with the results from the global data, which indicate that by varying the weights there is a significant difference in the outcome of the outbreak. This lends credence to the antigenic distance hypothesis, and that the antigenic distance of a strain from an ancestral strain greatly impacts the epidemic spread of the strain under the constraints of the model.

# 6 Concluding Remarks

Cellular-level dynamics of selection acting upon different strains of the influenza viruses by the immune system certainly affects the spread of the strains through a susceptible host population. The complex interactions involved cross scales and indicate the importance of understanding the interplay between evolution and epidemiology that has often been overlooked.

This research implemented phylogenetic distances created from differences in amino acid sequences of the primary antigenic protein for influenza (a surface molecule known as hemagglutinin). Influenza seasons were observed globally and locally and revealed that phylogenetic trees can lend some understanding to the dominance of a specific strain during an outbreak.

Genetic distances much be properly weighted to account for the location of mutations along the antigen sequence. The locations of mutations yield different antigenic distances of strains from their common ancestor, which in turn determine the fitness of the new strains. When weighted to account for antigenic distances, the genetic distances obtained from phylogenetic trees can be used to effectively estimate of the dominant strain for the following year.

The analyses conducted with statistics and with stochastic simulations produced sets of strains with large genetic distances from unknown ancestral strains that correctly contained the dominant strain the next year. However, the data available was not structured to allow for the fact that flu seasons transverse years, beginning in fall of the first year and ending just before spring the next year. Nevertheless, the analysis was able to account for regional differences, which is important due to the fact that people in different regions of the world have susceptibility to different strains. This is referred to as an immunological history. The data at global, regional, and local levels all followed a gamma distribution, which can account for high levels of susceptibles each year followed by high levels of recovered.

It was observed that, at the cellular-level, coexistence of very similar strains was possible during the infection of the host when the strains do not mutate. At the population-level, coexistence can only occur between strains with equal $\Re_0$, which indicates strains with the same fitness. Very different strains cannot coexist due to the different fitness values of the viral strains, since the immune system selects against the lowest fitness value. At first, this might seem to indicate that all influenza strains should eventually evolve similar properties. This is not the case, again due to the immune system, which easily eliminates strains that are similar to previously-encountered strains.

It seems that the heterogeneity of hosts' immune systems in fact greatly contribute to major epidemics. Individuals in different regions have different immunological histories, as was indicated by the fact that phylogenetic trees indicate different dominant strains in different regions. A strain might seem to be mild when present in the region in which it develops, since the hosts in that region have encountered it or a very similar strain in the past. However, when that seemingly-mild strain is carried by an infected individual to an

area that has not encountered a virus like it before, then symptoms associated with that would be more severe. Its fitness would suddenly be raised, and it would have a much larger impact in the new region. Eventually, though, the individuals in that region would build a general (herd) immunity toward that strain. However, the virus would mutate and continue to evolve and evade the immune system of the hosts. It could also potentially undergo an antigenic shift under certain conditions [12],, such as close proximity of the host to another host type that can also be infected with the flu (such as ducks or pigs)[20]. This would allow the virus strains from the two hosts to recombine into a brand new strain with a completely different hemagglutinin type. Such antigenic shifts have accounted for most of the pandemics mentioned in the introduction [4],[24].

It is apparent that the population-level model affects the cellular-level selection, though this was not examined with the model presented. The movements of hosts between regions as well as contact with other influenza host species is quite important to the evolution of the virus. In addition, at the cellular level, genetic distances alone are not the best predictor of strain dominance. Instead, the antigenic distances proposed by Smith *et al.*[21] could more accurately relate a circulating strain to its ancestral strain. This is due to the fact that not all mutations actually change the antigenicity of the virus. Future research should include closer links between the cellular-level model and the population-level model. In addition, future research could also study the antigenic distances, perhaps by determining exactly how the $\omega_i$ weights used in this model relate the genetic distances from the phylogenetic trees to these antigenic distances.

## 7 Acknowledgements

This research has been partially supported by grants from the National Security Agency, the National Science Foundation, the Theoretical Division of Los Alamos National Laboratory (LANL), the Sloan Foundation, and the Office of the Provost of Arizona State University. The authors are solely responsible for the views and opinions expressed in this paper; it does not necessarily reflect the ideas and/or opinions of the funding agencies, Arizona State University, or LANL. The authors wish to thank the MTBI staff, especially Carlos Castillo-Chavez, Linda Gao, Bao Sung, Christopher Kribs Zaleta, and Priscilla Greenwood. They also thank Alan Perelson, Catherine Macken, Ruy Ribeiro, and Luis Bettencourt of the Los Alamos National Laboratory, as well as Jeff Long of the University of Michigan and John Jungck of Beloit College.

## References

[1] L. J. S. Allen, *An Introduction to Stochastic Processes with Applications to Biology* (Prentice Hall, New Jersey 2003)

[2] P. M. Colman, Protein Sci **4** 1687-1696 (1994).

[3] N. Cox, in *Options for the Control of Influenza II*, edited by C. Hannoun, A. P. Kendal, H. D. Klenk, and F. L. Ruben (Elsevier Science, Amsterdam 1993), pp. 193-201.

[4] Department of Health and Human Services Center for Disease Control and Prevention. *Influenza (Flu) Activity.* Accessible online: http://www.cdc.gov/flu/weekly/fluactivity.htm (2005).

[5] O. Diekmann and J. A. P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation.* (Wiley, New York 2000).

[6] P. van den Driessche and J. Watmough, J of Math Biosci **180**, 1, 29-48, (2002).

[7] D. J. D. Earn, J. Dushoff, and S. A. Levin, Trends Ecol Evol **17**, 7, 334-340 (2002).

[8] A. P. Galvani, Trends Ecol Evol **18**, 3, 132-139 (2003).

[9] P. Greenwood, lecture series.

[10] A. Gronoff, in *The Encylopedia of Virology*, edited by Robert G. Webster, p. 824-829.

[11] M. Holder and P. O. Lewis, Nat Rev Genet **4**, 4, 275-284, (2003).

[12] C.A. Janeway, P. Travers, M. Walport, and M. J. Shlomchik, *Immunobiology: The Immune System in Health and Disease.* (Garland Science, New York 2005), 6th Edition, pp. 351-356, pp. 410-432.

[13] S. Kumar, K. Tamura, and M. Nei, Breifings in Bioinformatics **5**, 150-163.

[14] M. S. Lee and J. S-E. Chen. Emerg Infect Dis **10**, 8, 1385-1390 (2004).

[15] C. Macken, H. Lu, J. Goodman, and L. Boykin, in *Options for the Control of Influenza IV*, edited by A. D. M. E. Osterhaus, N. Cox, and A.W. Hampson (Elsevier Science, Amsterdam 2001), pp. 103-106.

[16] M. A. Nowak and C.R. Bangham, Science **272**, 5258, 74-79 (1996).

[17] National Center for Biotechnology Information (NCBI). *Influenza Virus Sequence Database.* Accessible online from: http://www.ncbi.nlm.nih.gov/genomes/influenza/list.cgi (2005).

[18] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho, Science **271**, 5255, 1582-1586 (1996).

[19] A. S. Perelson (unpublished).

[20] C. Scholtissek, U. Schultz, S. Ludwig, and W. M. Fitch, in *Options for the Control of Influenza II*, edited by C. Hannoun, A. P. Kendal, H. D. Klenk, and F. L. Ruben (Elsevier Science, Amsterdam 1993), pp. 193-201.

[21] D. J. Smith, S. Forrest, D. H. Ackley, and A. S. Perelson. P Natl Acad Sci USA **96**, 24, 14001-14006 (1999).

[22] W-Y. Tan and H. Wu, Math Biosci **147**, 2, 173-205 (1998).

[23] P. H. Thrall and J. J. Burdon, Plant Pathol. **49**, 6, 767-773 (2000).

[24] World Health Organization. *FluNet Global Influenza Surveillance Database.* Accessible online: http://www.who.int/csr/disease/influenza/influenzanetwork/en/index.html (2005).

[25] Y. Xia, J. Gog, and B. Grenfell, Appl Stat **54**, 3, 659-672 (2005).

# 8 Appendix

## 8.1 Cellular Level Calculations

### 8.1.1 Calculation of $\Re_0$

$$J = \begin{bmatrix} -d - \sum_{k=1}^{n} \beta_k V_k^* & 0 \ldots 0 & -\beta_1 X^* \ldots - \beta_n X^* & 0 \ldots 0 \\ \begin{matrix} \beta_1 V_1^* \\ \vdots \\ \beta_n V_n^* \end{matrix} & \clubsuit & \begin{pmatrix} \beta_1 X^* & 0 & \ldots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \beta_1 X^* & 0 \\ 0 & 0 & \ldots & \beta_1 X^* \end{pmatrix} & \begin{pmatrix} -pY_1^* & 0 & \ldots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & -pY_{n-1}^* & 0 \\ 0 & 0 & \ldots & -pY_n^* \end{pmatrix} \\ \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & \begin{pmatrix} m_{11}k_1 & m_{21}k_2 & \ldots & m_{n1}k_n \\ m_{12}k_1 & m_{22}k_2 & \ldots & m_{n2}k_n \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n}k_1 & m_{2n}k_2 & \ldots & m_{nn}k_n \end{pmatrix} & \begin{pmatrix} -\mu & 0 & \ldots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & -\mu & 0 \\ 0 & 0 & \ldots & -\mu \end{pmatrix} & \bigcirc \\ \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & \bigstar & \bigcirc & \circledast \end{bmatrix}$$

where:

$$\clubsuit = \begin{pmatrix} -(a+pZ_1^*) & 0 & \ldots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & -(a+pZ_{n-1}^*) & 0 \\ 0 & 0 & \ldots & -(a+pZ_n^*) \end{pmatrix}$$

$$\bigstar = \begin{pmatrix} c_1 Z_1^* & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & c_{n-1} Z_{n-1}^* & 0 \\ 0 & 0 & \dots & c_n Z_n^* \end{pmatrix}$$

$$\circledast = \begin{pmatrix} c_1 Y_1^* - b & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & c_{n-1} Y_{n-1}^* - b & 0 \\ 0 & 0 & \dots & c_n Y_n^* - b \end{pmatrix}$$

The Jacobian for the Disease Free Equilibrium )DFE) can be written as follows:

$$J_{\text{DFE}} = \begin{bmatrix} -d & 0 \dots 0 & -\beta_1 \frac{\Lambda}{d} \dots - \beta_n \frac{\Lambda}{d} & 0 \dots 0 \\ 0 \\ \vdots \\ 0 & \begin{pmatrix} -a & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & -a & 0 \\ 0 & 0 & \dots & -a) \end{pmatrix} & \begin{pmatrix} \beta_1 \frac{\Lambda}{d} & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \beta_{n-1} \frac{\Lambda}{d} & 0 \\ 0 & 0 & \dots & \beta_n \frac{\Lambda}{d} \end{pmatrix} & \bigcirc \\ 0 \\ \vdots \\ 0 & \begin{pmatrix} m_{11}k_1 & m_{21}k_2 & \dots & m_{n1}k_n \\ m_{12}k_1 & m_{22}k_2 & \dots & m_{n2}k_n \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n}k_1 & m_{2n}k_2 & \dots & m_{nn}k_n \end{pmatrix} & \begin{pmatrix} -\mu & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & -\mu & 0 \\ 0 & 0 & \dots & -\mu \end{pmatrix} & \bigcirc \\ 0 \\ \vdots \\ 0 & \bigcirc & \bigcirc & \begin{pmatrix} -b & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & -b & 0 \\ 0 & 0 & \dots & -b \end{pmatrix} \end{bmatrix}$$

$$\lambda_1 = -d < 0$$

$$\lambda_{2n+2}, \dots, \lambda_{3n+1} = -b < 0$$

The basic reproduction number, $R_0$, is calculated by using the second generator approach as described on Diekmann and Heesterbeek, and van den Driessche and Watmough [6])

$$M = \begin{bmatrix} \bigcirc & \begin{pmatrix} \beta_1\frac{\Lambda}{d} & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \beta_{n-1}\frac{\Lambda}{d} & 0 \\ 0 & 0 & \dots & \beta_n\frac{\Lambda}{d} \end{pmatrix} \\ \begin{pmatrix} m_{11}k_1 & m_{21}k_2 & \dots & m_{n1}k_n \\ m_{12}k_1 & m_{22}k_2 & \dots & m_{n2}k_n \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n}k_1 & m_{2n}k_2 & \dots & m_{nn}k_n \end{pmatrix} & \bigcirc \end{bmatrix}$$

$$D = \begin{bmatrix} \begin{pmatrix} a & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & a & 0 \\ 0 & 0 & \dots & a \end{pmatrix} & \bigcirc \\ \bigcirc & \begin{pmatrix} \mu & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \mu & 0 \\ 0 & 0 & \dots & \mu \end{pmatrix} \end{bmatrix}$$

$$MD^{-1} = \begin{bmatrix} \bigcirc & \begin{pmatrix} \frac{\beta_1}{\mu}\frac{\Lambda}{d} & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & \frac{\beta_{n-1}}{\mu}\frac{\Lambda}{d} & 0 \\ 0 & 0 & \dots & \frac{\beta_n}{\mu}\frac{\Lambda}{d} \end{pmatrix} \\ \begin{pmatrix} m_{11}\frac{k_1}{a} & m_{21}k_2 & \dots & m_{n1}\frac{k_n}{a} \\ m_{12}\frac{k_1}{a} & m_{22}\frac{k_2}{a} & \dots & m_{n2}\frac{k_n}{a} \\ \vdots & \vdots & \ddots & \vdots \\ m_{1n}\frac{k_1}{a} & m_{2n}\frac{k_2}{a} & \dots & m_{nn}\frac{k_n}{a} \end{pmatrix} & \bigcirc \end{bmatrix}$$

### 8.1.2 Calculation of the Immune-Free Equilibrium

The equilibrium can be evaluated using Equation (8).

$$\frac{\det A_1}{\det A}V_1^\infty + \frac{\det A_2}{\det A}V_2^\infty + \cdots + \frac{\det A_n}{\det A}V_n^\infty = 1, \tag{8}$$

where

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{n1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1n} & \alpha_{2n} & \cdots & \alpha_{nn} \end{bmatrix},$$

$$\alpha_{ij} = \frac{\Lambda}{\mu a} m_{ij} k_i \beta_i, \quad i \neq j,$$

$$\alpha_{ii} = \frac{\Lambda}{\mu a} m_{ii} k_i \beta_i - d,$$

and $A_j$ is the matrix obtained by replacing the $j^{th}$ row of $A$ with the row vector $\vec{\beta}^T = [\beta_1 \ \beta_2 \ \cdots \ \beta_n]$.

The equation for the $V_i^\infty$ comes from the $n$ equilibrium conditions obtained by setting $\dot{V}_i = 0 \ \forall \, i$ and substituting the above expressions for $X^\infty$, $Y_i^\infty$, and $Z_i^\infty$. These can be rewritten as a single vector equation:

$$\left(\vec{V}^\infty \vec{\beta}^T - A + dI\right) \vec{V}^\infty = \vec{0}$$

($I$ is the identity matrix). Solutions to this require either that $\vec{V}^\infty = \vec{0}$ or that the matrix $\left(\vec{V}^\infty \vec{\beta}^T - A + dI\right)$ be singular. Since we are interested in the endemic equilibrium, we discard the all-zero solution for $\vec{V}^\infty$ and set the determinant of the matrix to zero. Some algebra yields the linear equation (8) given above.

Note that (8) has infinitely many solutions (a set of dimension $n-1$ in general), unless all the coefficients $\det A_i / \det A$ are negative, i.e., $\det A_i$ all have the same sign as each other ($i = 1, 2, ..., n$) and the opposite sign as $\det A$, in which case there are no solutions. This creates infinitely many (non-isolated) partially immune-free equilibria.

## 8.2 Stochastic (Poisson Process) Model

The cellular level model can be characterized with stochastic effects by Equation (9) [1],[9]

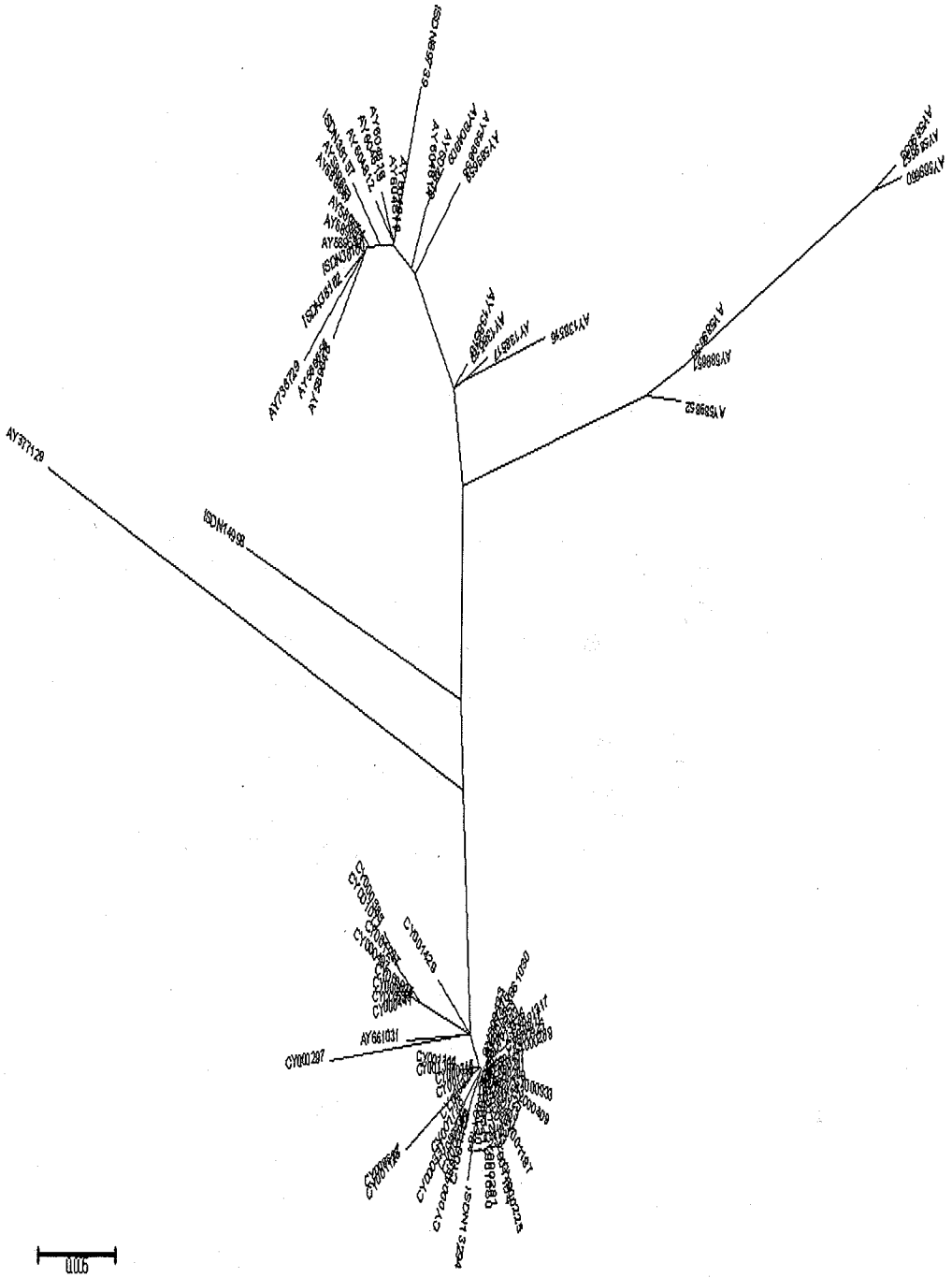
$$
\begin{aligned}
&\text{Prob}\{\Delta N(t)=j, \Delta S(t)=l, I_i(t)=m_i, R_i(t)=p_i, 1\le i\le n \mid N(t), S(t), I_i(t), R_i(t)\} \\
&= \begin{cases}
\Lambda\Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (1, \overbrace{0, \dots 0}^{3n}) \\
dX(t)\Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (-1, \overbrace{0, \dots 0}^{3n}) \\
X(t)\beta_i V_i(t)\Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (-1, \overbrace{0, \dots, 0}^{n+i-1}, 1, \overbrace{0, \dots 0}^{2n-i}) \\
Y_i(t)[a + pZ_i(t)]\Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (\overbrace{0, \dots, 0}^{n+i}, -1, \overbrace{0, \dots 0}^{2n-i}) \\
X(t)\sum_{j=1}^{n} k_j m_{ji} Y_j \Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (\overbrace{0, \dots, 0}^{i}, 1, \overbrace{0, \dots 0}^{3n-i}) \\
\mu V_i(t)\Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (\overbrace{0, \dots, 0}^{i}, -1, \overbrace{0, \dots 0}^{3n-i}) \\
c_i Z_i(t) Y_i(t)\Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (\overbrace{0, \dots, 0}^{2n+i}, 1, \overbrace{0, \dots 0}^{n-i}) \\
bZ_i(t)\Delta t + o(\Delta t), \\
\quad (e, f_1, f_2 \dots f_n, g_1, g_2, \dots g_n, h_1, h_2, \dots h_n) = (\overbrace{0, \dots, 0}^{2n+i}, -1, \overbrace{0, \dots 0}^{n-i}) \\
1 - \Big[\Lambda + X(t)(d + \sum_{r=1}^{n} \beta_r V_r(t)) + \sum_{r,j=1}^{n} k_j m_{jr} Y_j(t) \dots \\
+ \sum_{j=1}^{n} [\mu V_j(t) + (c_j Y_j(t) - b) Z_j(t) + (a + pZ_j(t)) Y_j(t)]\Big] \Delta t + o(\Delta t), \\
\quad (e, f_1, \dots, f_n, g_1, \dots g_n, h_1, \dots, h_n) = \overbrace{0, \dots, 0}^{3n+1} \\
o(\Delta t) \quad \text{otherwise}
\end{cases}
\end{aligned}
\tag{9}
$$

The cellular level has $3n+1$ classes including: One susceptible class, $X$, which corresponds to the healthy cell population; $n$ infected classes, $Y_i$, which correspond to cells infected by each viral strain $i$; $n$ infective classes, $V_i$, which correspond to the viral load of strain $i$; and $n$ CTL classes, $Z_i$, which corresponds to the population of Cytotoxic T Lymphocytes acting against each cells infected with viral strain $i$. There are also $n^2+3n+6$ parameters: $\Lambda$ is the constant rate of birth of healthy cells; $d$ is the proportional rate of natural death of healthy cells; $\vec{\beta}$ which is a $1 \times n$ array of infection rates of each free

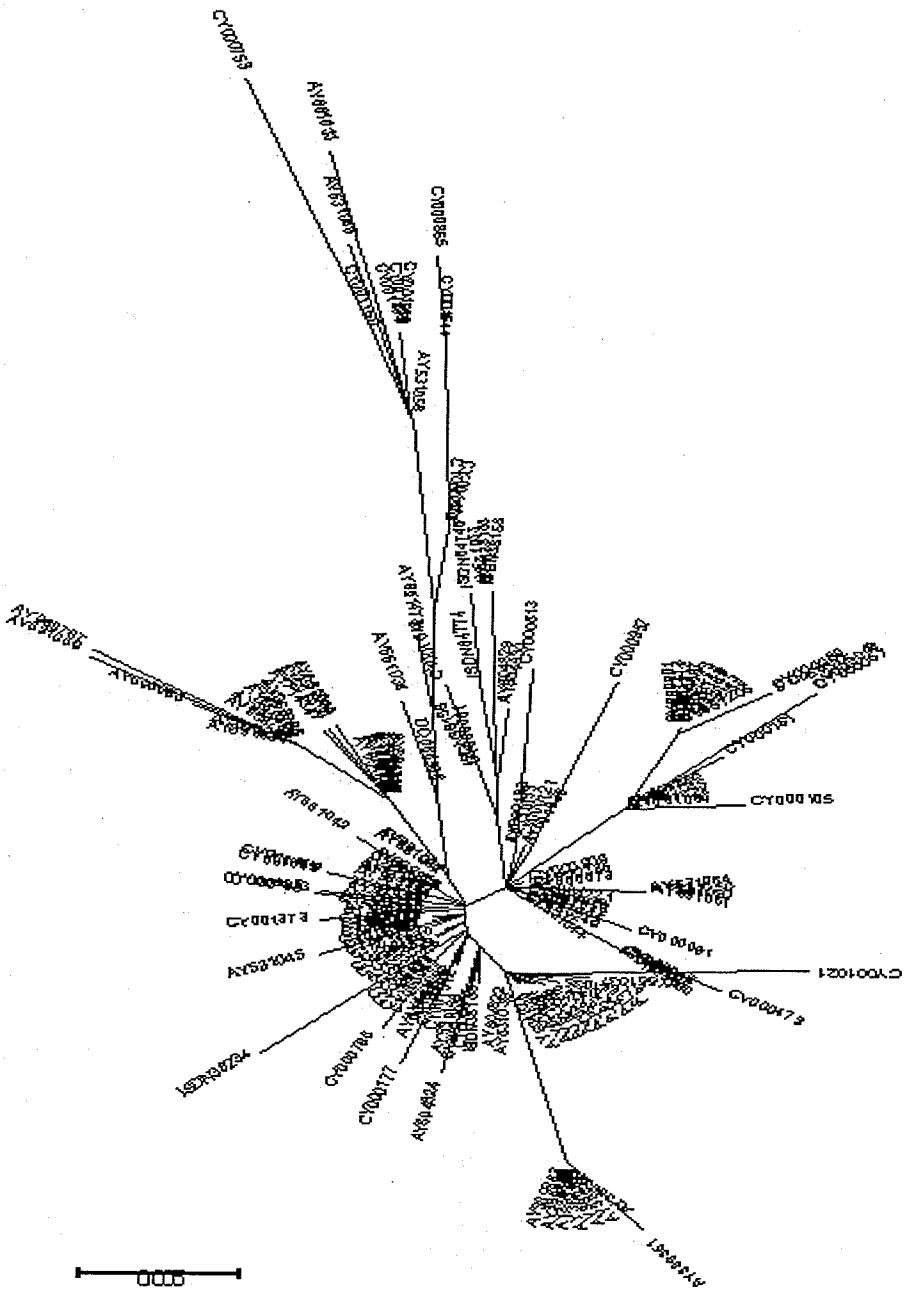
virus; $\overrightarrow{k}$, which is a $1 \times n$ array of replication rates of each strain within an infected cell; a matrix $m$ of mutation rates, where $m_{ji}$ is the rate of infection from strain $j$ to strain $i$; and a clearance rate $\mu$ of free virus;

## 8.3 Phylogenetic Trees

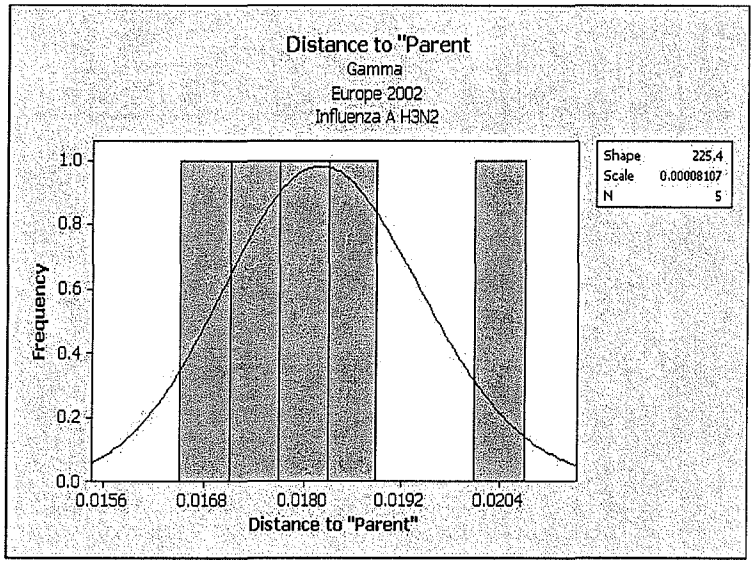
### 8.3.1 Global Phylogenetic Tree 2002



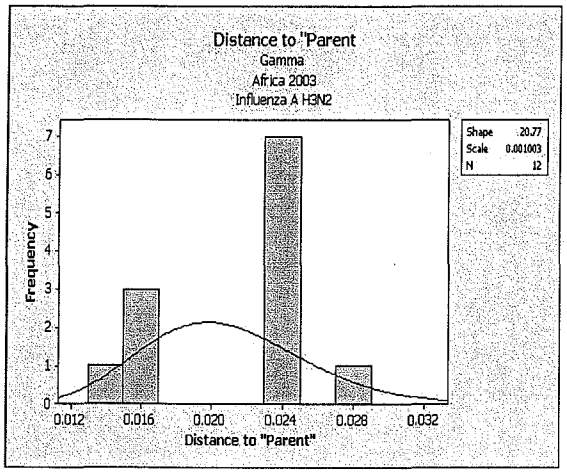
### 8.3.2 Global Phylogenetic Trees 2003

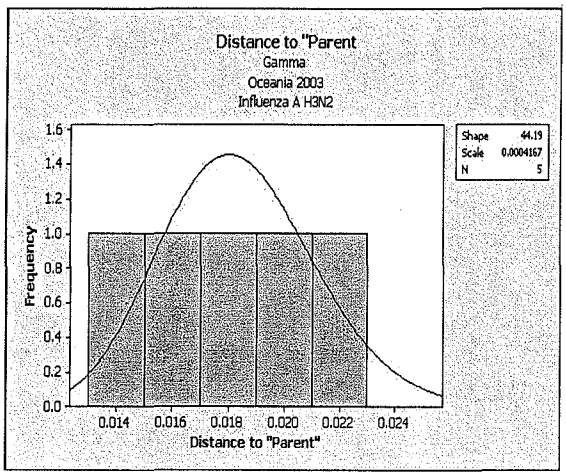


## 8.4 Histogram of 2002 Regional Data




## 8.5 Histogram of 2003 Regional Data



## 8.6 Histogram of 2004 Regional Data