View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IntechOpen

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

13

Predicting Virus Evolution

Tom Burr
Los Alamos National Laboratory
USA

1. Introduction

Viruses are an important cause of human disease, often because they are highly transmittable from human to human. A key tool from population genetics that can be applied to the study of viruses is coalescent theory. Coalescent theory predicts genealogical tree shapes as a function of how the studied organisms are evolving. Therefore, under its model assumptions, coalescent theory can be used to infer aspects of the demographic history of evolving organisms. For example, there are characteristics of tree shapes that imply whether the organism population has been constant, growing, or shrinking in size over time.

This chapter reviews some of the successes of coalescent theory in the context of inferring aspects of virus evolution, using human immunodeficiency (HIV) and influenza viruses as case studies. Next, the chapter describes limitations of coalescent theory, even as extended to allow some forms of selection, population subdivision, and viral recombination. The relatively new goal to predict influenza virus evolution (rather than infer past evolution) is used to emphasize modeling needs beyond standard or extended coalescent theory models. A new small-scale simulation that combines viral fitness with demographic population structures such as family and work groups is then described as an example extension to coalescent theory models.

Prediction goals include early detection of highly lethal new strains and improved vaccine designs that anticipate future evolutionary directions. Regardless which evolutionary model is used to predict virus evolution, because real virus evolution is complex beyond current understanding, there will be substantial model error. Model error, model parameter estimation error, and purely random effects can combine to make some forecast goals unattainable. In these cases the most appropriate prediction is similar to what is often said about stock markets: there will be change.

2. Background

Humans are susceptible to many viral pathogens, including the human immunodeficiency virus (HIV) and the influenza virus. Although it is a relatively new goal in population genetics, predicting virus evolution can help with vaccine design and with other mitigation strategies (Bush et al., 1999; Ferguson and Anderson, 2002; Plotkin et al., 2002; Rambaut et al., 2008). Using estimated phylogenetic (genealogical) tree shapes to infer aspects of evolution such as organism growth rates has received far more attention to date (Felsenstein et al., 1999; Innan and Stephan, 2000; Pybus et al., 2000; Stephens and Connelly, 2000; Ewing et al., 2004)

This chapter focuses on HIV and the influenza virus in the context of what might be predicted about virus evolution. Influenza is a highly transmittable disease that infects millions each year, resulting in many deaths. HIV is also transmittable through risky behaviors and it too results in many deaths each year.

In population genetics, coalescent theory (Kingman, 1982; Stephens and Donnelly, 2000; Burr et al., 2001; Ewing et al., 2004;) is a key tool that predicts genealogical tree shapes as a function of how the studied organisms (taxa) are evolving. Therefore, under its model assumptions, coalescent theory can be used to infer aspects of the demographic history of evolving organisms. For example, there are characteristics of tree shapes that imply whether the organism population has been constant, growing, or shrinking in size over time (Pybus et al., 2000).

This chapter will first review some of the successes of coalescent theory in the context of inferring aspects of virus evolution, using HIV (Rodrigo et al., 1999; Burr et al., 2001; Rambaut et al., 2001) and influenza viruses (Ferguson and Anderson, 2002; Plotkin et al., 2002, Burr et al., 2002) as case studies. Next, the chapter describes limitations of coalescent theory, even as extended to allow some forms of serial sampling, selection, population subdivision, and viral recombination (Excoffier and Foll, 2011). The relatively new goal to predict influenza virus evolution (rather than infer past evolution) is used to emphasize modeling needs beyond standard or extended coalescent theory models. A new small-scale simulation that combines viral fitness with demographic population structures such as family and work groups is then described as an example extension to coalescent theory models.

Most genetic data analyses rely on a forward model that specifies evolutionary forces and associated probabilities describing how offspring are generated. Evolutionary forces include drift, mutation, recombination, migration, and selection. Drift refers to random change over successive generations due to finite population sizes. In the absences of mutation and selection, the fraction of a population of size N having a given trait drifts randomly somewhat like the number of heads in a set of N coin tosses. Mutations are changes in the DNA sequence that occur for many reasons. Recombination ("reassortment" in the case of influenza) refers to sections of genome that are broken and then recombined, resulting in large genetic differences between offspring and parents, and complicating phylogenetic analyses because different genome sections can have different genealogies. Migration refers to exchange of genetic material among partly isolated subpopulations. There are many other evolutionary forces, too many to describe here.

For simulating DNA sequences from a population, the state of art invokes coalescent theory, which uses simplified models of the forward evolutionary process. These simplifications allow inverse analytical solutions and corresponding simulation software, but with questionable assumptions. This is done in order to avoid having to simulate directly from the forward model and track the evolutionary histories. Sample genealogies can instead be simulated by running time from the present toward the past and tracking probabilistically when lineages coalesce to share a common ancestor. An example genealogy of a sample taken at a single time from a population that is maintaining a constant population size is given in Figure 1. These coalescent-based simulated sample units are then used to infer how a population is evolving using features of the associated phylogenetic tree (Pybus et al., 2000). In addition, analytical approximations used in inference invoke the same model assumptions used in coalescent theory.

Agent-based models (Eubank et al., 2004) provide a richer framework than classical epidemiology models of disease spread, such as the susceptible-infected-recovered (SIR) model. Because agent-based models track individual rather than aggregate behavior, they are believed to more reliably predict, for example, the impact of candidate mitigation strategies such as vaccinations and isolation. In an analogous way, we describe predictions for virus evolution that probably will require a higher-fidelity modeling framework than coalescent theory and its extensions.

The following sections include HIV and influenza examples of using coalescent theory to infer aspects of prior evolution, limitations of coalescent theory to infer future HIV and influenza evolution, introduces the new small-scale simulation that combines viral fitness with demographic features, and discusses limitations of our current ability to predict viral evolution.

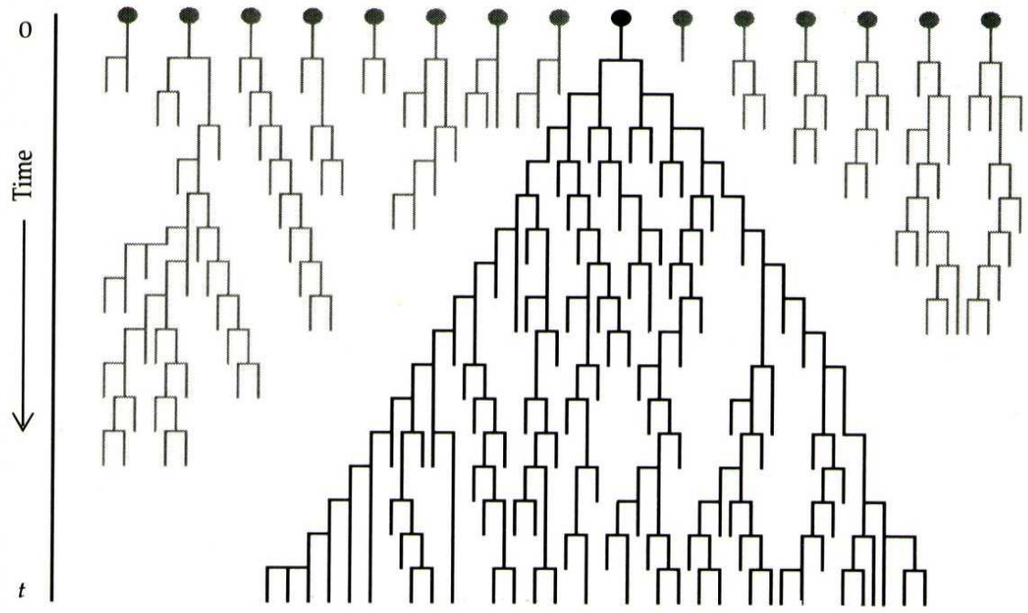


Fig. 1. The most recent common ancestor and sample genealogy from an evolving population. At time *t* a sample is collected, and at time 0 in the past, all individuals in the sample coalesce to share a common ancestor. For simplicity here, the population size is assumed to be constant over time, with each individual at time 0 represented by a dot (not all individuals are shown).

3. HIV and influenza examples

3.1 Example 1: HIV

Within host. A coalescent model within individuals has been applied (Rodrigo et al., 1999) to analyze HIV-1 viral load data from infected individuals after the administration of an HIV-1 inhibitor to estimate the HIV generation time in vivo. The estimate was 1-2 days, which agreed well with an estimate based on a different approach (Perelson et al, 1996), although it assumed nonrecombining DNA sequences from a population of constant effective size *N*. Consider samples from two sample times, separated by *d* days. The number of days per generation is estimated as $dn(n-c)/2Nc$ where *c* is the number of coalescent events that have occurred, *d* is the number of days between samples. The method assumed: (a) the

population size N is constant; (b) the estimated phylogeny is the same as the true genealogy of the sampled individuals, and (c) the exchangeability assumption that each individual virus has the same propensity to reproduce. Implicit in (b) is the further assumption that recombination, migration, and selection do not interfere with the ability to estimate the true phylogeny. Further, an approximate technique for accommodating serial samples is required, which has recently become available (Excoffier and Foll, 2011).

Between host. An example (Burr et al., 2001) involves whether the 8 to 10 approximately equidistant subtypes of HIV-1 (type M) could have arisen under available models of how HIV is evolving (Fig. 1). To examine this, coalescent theory was used to simulate DNA data from a very simplified forward model of how HIV is evolving at both the macro and micro levels (see Section 4.1). This provided a reference distribution against which to compare the data. If features of the observed data (such as the ratio of the between-subtype to within-subtype genetic distance) are in the tail of the coalescent-theory-based reference distribution of those same features, then the forward model used to simulate the data is not credible. Examples of phylogenetic trees estimated from coalescent-based simulated data are given in the right box of four subplots in Figure 2. Notice that subtypes are expected to arise in examples (b), (c), and (d), but not in (a). Subplot (a) is the classic "star phylogeny" that arises when the underlying population size is growing rapidly, forcing most coalescent events to occur early in the growth period, and all at nearly the same time.

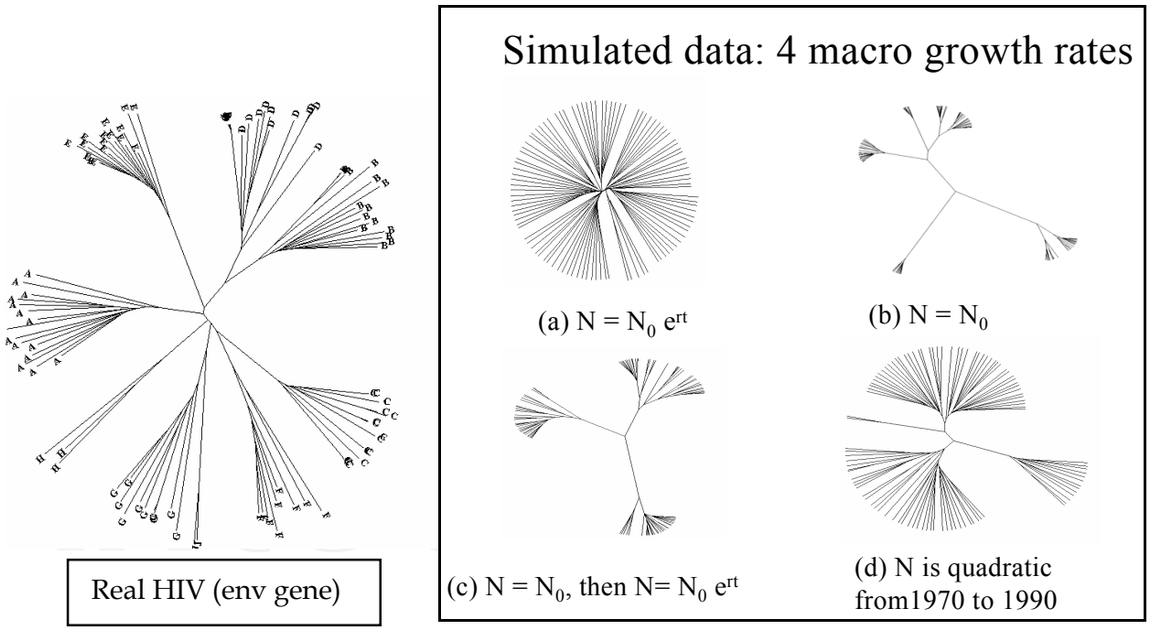


Fig. 2. **HIV, env region.** Consensus trees (of 100 bootstrap samples) using maximum likelihood for real (the left plot) HIV (env gene) sequences and for coalescent-based simulated (the four right plots) sequences under different assumptions about the time behavior of the number of infecteds N.

3.2 Example 2: Influenza

Figure 3 shows a principal coordinate (PC) plot (Venables and Ripley, 1999) of a 129-by-129 distance matrix based on the nucleoprotein (NP) region of 129 influenza viruses isolated

from humans, swine, and avian hosts. PCs provide a low dimensional way to represent a distance matrix. For this data, all pairs of distances can be quite accurately reproduced using only the first two PCs as in Figure 3. It is known that the NP region maintains a type of "species signature" such as depicted in Figure 3 (Burr et al., 1999, 2002; Chen et al., 2006). A key aspect of influenza evolution is the fact that avian and swine hosts occasionally act as "reassortment vessels" for human influenza, resulting in dramatically different strains that evade effective human immune response. As an aside, the term "reassortment" seems to be applied only to influenza, presumably because its genome consists of eight distinct segments. For our purposes, "reassortment" is the same as recombination, in which sections of the genome get recombined (Forrest and Webster, 2010).

Figure 4 shows a PC plot of a distance matrix based on the hamagglutinin (HA) region of influenza viruses. Figure 5 is a phylogenetic tree built using neighbor joining (Swofford et al., 2000) of the same HA sequences.

Figures 4 and 5 illustrate (Nelson and Holmes, 2007) that the HA region appears to display the effects of positive selection due to the cactus-like structure with most lineages dying out. This cactus shape is unlike the classic "star-like" shape HIV trees of type M as in Figure 2. Such a cactus shape can also arise without positive selection from a combination of serially sampled taxa and sequential random population bottlenecks (which can occur in influenza due to its strong seasonality). Therefore, the cactus shape by itself can indicate but does not prove that positive selection is in effect. More formally, the statistical notion of identifiability probably does not hold in this context. Model identifiability implies that as sample size increases toward infinity, model parameters can be uniquely estimated (see Section 6).

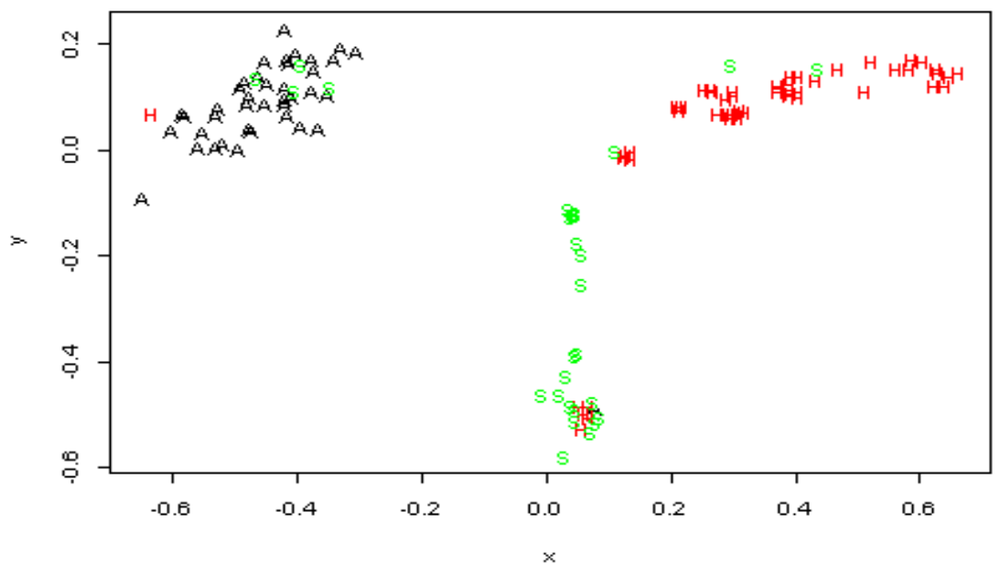


Fig. 3. Principal coordinate plot of the evolutionary distances among 129 influenza viruses extracted from human (H), avian (A), and swine (S) hosts. Distances are computed for the Nucleoprotein region of the virus, which exhibits species signatures. Among the 129, there are 14 "misidentified" taxa. However,the human (H) that is clustered with the avian group was known to have been infected by poultry. There are 44 avian, 57 humans, and 28 swine, all available from www.flu.lanl.gov.

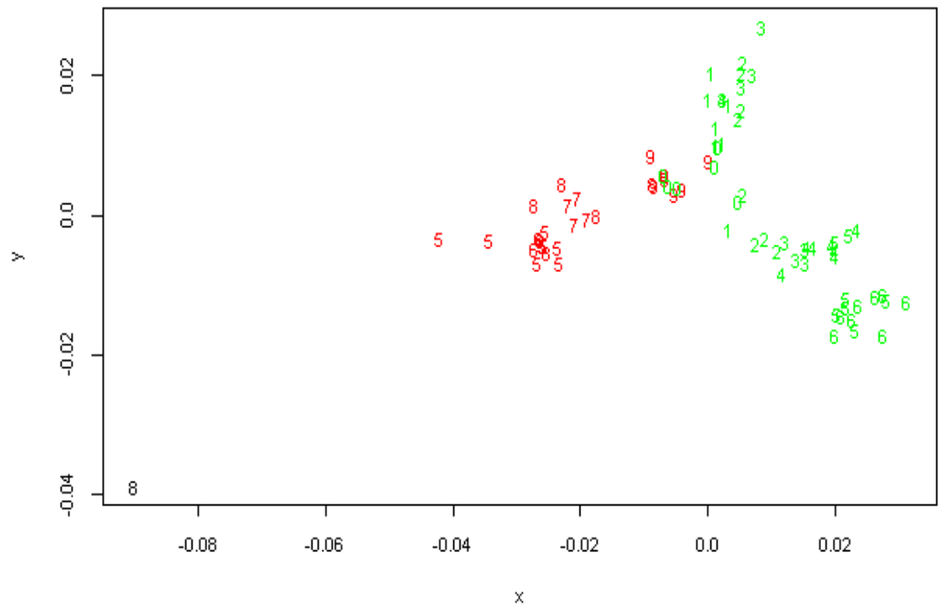


Fig. 4. Principal coordinate plot of the influenza viruses (HA region) found in humans. Digit = year, Black = 1960's, Red = 1980's, Green = 1990's. Genetic drift and strain extinctions are known to occur (cactus shape of typical tree).

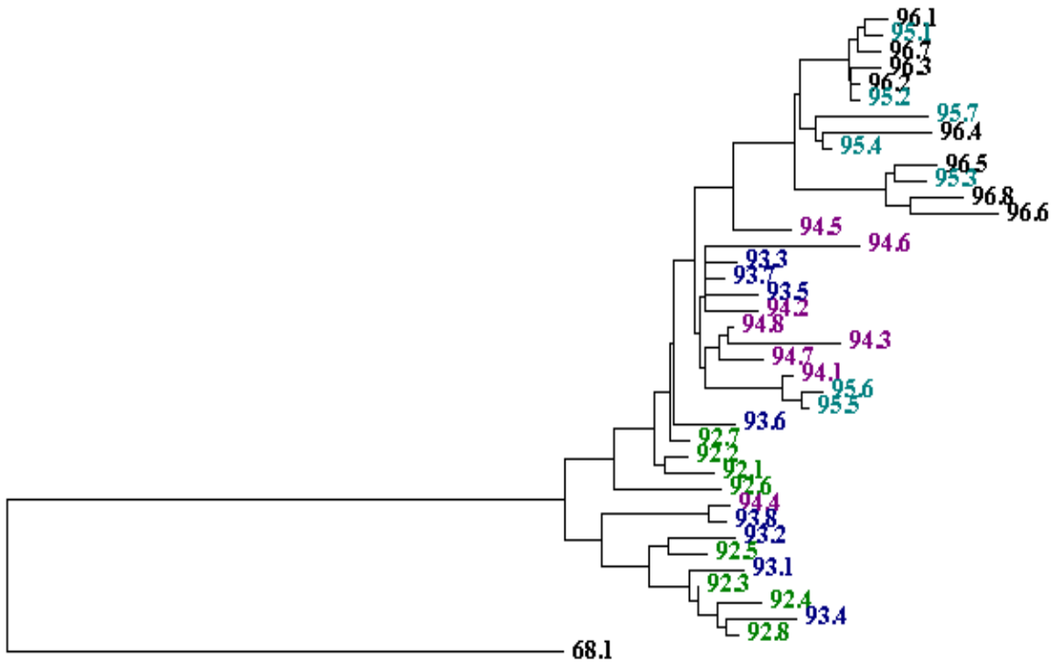


Fig. 5. Neighbor joining tree of the same HA sequences in humans that was used in Fig. 3.

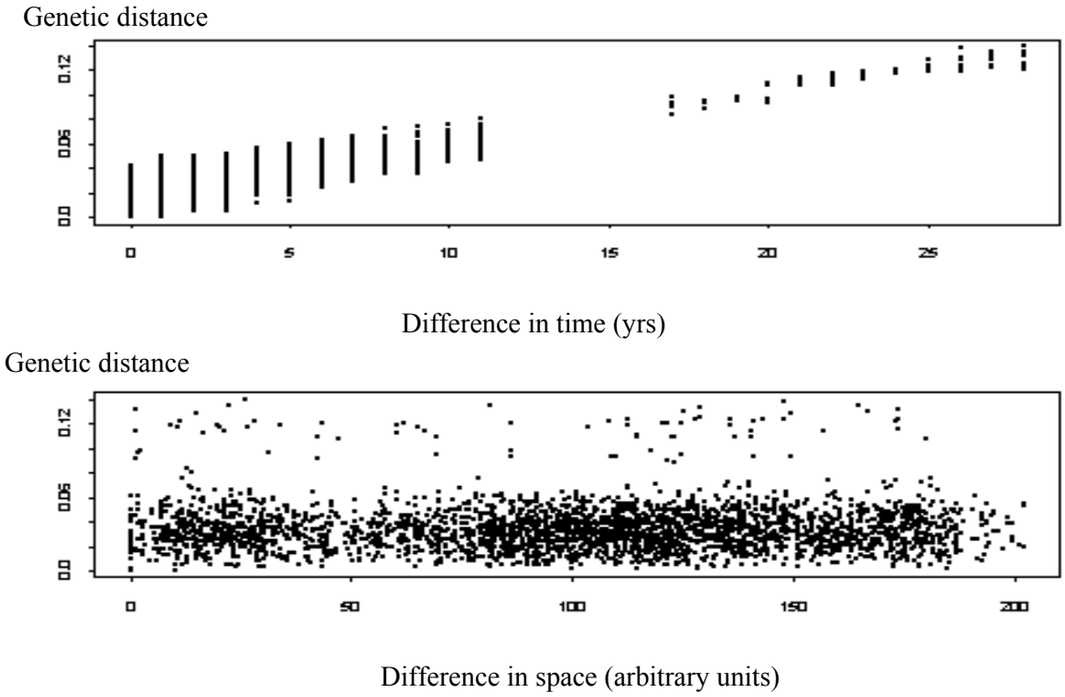


Fig. 6. Genetic distance versus time (top) and versus difference in space (bottom) for the HA region.

4. Limitations of coalescent theory for predicting HIV and influenza evolution

Coalescent theory leads to tremendous insights and powerful simulation and inference tools. However, limitations of coalescent techniques include (a)-(d) as follows:

a. Little is known concerning accuracy and robustness of coalescent theory's restrictive assumptions in many settings, although some forward models are known not to be well approximated by any coalescent model (depending on the relative time scales of various evolutionary effects such as drift, migration, and selection) (Sjodin et al., 2005);
b. Inference methods (Pybus et al., 2000; Stephens and Donnelly, 2000) invoke coalescent approximations to estimate the probability of candidate branching orders as part of the inference process. This leads to the undesirable situation of forcing a zero mismatch between the inference method's assumptions and the assumptions regarding how the population is evolving;
c. Coalescent theory is expanding along with associated software for implementation, but no current coalescent-based software includes all extensions to the original coalescent theory. However, one new option (Excoffier and Foll, 2011) for coalescent-based software includes many of the standard evolutionary features such as serial sampling, recombination, and geographic isolation;
d. Building trees supports inferences regarding, for example, whether a virus strain appears to be a natural branch from historical strains, or whether the strain seems to have made an unnatural leap indicating bioengineering. However, key coalescent assumptions that are violated by both HIV and influenza viruses are that all subtypes are equally transmissible and there is no recombination. Therefore, although to a limited extent and under restrictive assumptions, extensions to coalescent theory have

been made to accommodate recombination, selection, overlapping generations, and population subdivision, there are cases where the theory is either inadequate or the sensitivity of its conclusions to its assumptions is unknown. The corresponding inference quality using estimated trees is also unknown; the state of the art is therefore to quantify precision, but not accuracy.

The forward model is a key component of total uncertainty associated with population genetics inferences. The current approach is: specify an amenable-to-coalescent-theory forward model for how a population is evolving that includes for example, population size, structure, and selection effects; identify the coalescent effective population size N_e (Sjodin et al., 2005) in the nearest available coalescent model, which is often a complicated task. Then, use the closest coalescent model to simulate sample genealogies under restrictive assumptions about the population and the sampling process. The $N_{effective}$ notion arose from coalescent theory by mapping the actual population size N in a population that violates some coalescent assumptions (such as nonoverlapping generations) to a different size $N_{effective}$ such that in some aspects, the actual and model populations evolve probabilistically in approximately the same manner. Coalescent theory was originally applied to macroscopic populations such as plants and animals (for example, Innan and Stephan, 2000); it has also been applied to microscopic populations such as DNA sequences from virus populations (for example, Rodrigo et al, 1999).

Coalescent theory will continue to provide insight into evolutionary processes; however, it is currently unknown how robust associated inferences are with respect to model violations. For example, Innan and Stephan [5] assumed the wild plant Arabidopsis thaliana (useful for genetic studies because of its well known demographic history and genome) consists of many isolated colonies, each having negligible genetic variation within a colony. They applied a coalescent model (correcting for the growing population size of A. thalina) to simulate the probability distribution of Tajima's D statistic against which to compare the observed D in real samples, as a test for selection. Tajima's D statistic is based on the difference between two estimates of the amount of variation (one using the number of sites having genetic variation and the other based on pairwisedifferences between individuals). They concluded that there was evidence for selection (distinguishing the type of selection, such as balancing or purifying is a separate challenge). Because of the simplifications inherent in the coalescent approach, it is currently unknown how robust the evidence for selection is in this for A. thalina.

4.1 Example 1: HIV

Coalescent models of HIV reproduction within an individual might be adequate (Rodrigo et al., 1999); however, these would become prohibitively unwieldy if all HIV-infected humans were modeled. For example, all models must specify the macro components such as the reproductive rate in susceptible populations and/or subpopulations.

As mentioned in Section 3.1, an investigation into the development of the HIV subtypes led to an application of coalescent theory to model the population dynamics of HIV (Burr et al., 2001). Figure 2 illustrates the approach taken. Various features (such as the ratio of the between-subtype to within-subtype genetic distance) involving the subtypes of real HIV sequences (env gene) were compared to the same features in corresponding coalescent-based simulated data. However, it became apparent that it would be necessary to implement a model that made less restrictive assumptions than coalescent theory (Burr et al., 2001).

One possible new way to simulate sequences is to track each HIV case by geographic region including all known transmission routes such as sex, needles, blood transfusions, and mother-to-child, and track the genealogy of each case. One would then sample ~100 simulated sequences from around the world or in specified regions at a snapshot in time, or distributed in time, and distributed spatially in either case. With careful bookkeeping one could deduce the sample genealogy (which 2 samples coalesced first to their most recent common ancestor (MRCA), which samples coalesced next, etc.) back in time until all 100 sequences coalesced to the single MRCA. This would produce 99 coalescent times and sample identities, which define the genealogy of the sample. This genealogy could also be thought of as the true evolutionary tree for the sample to be compared to coalescent-based genealogies.

Related to the origin of the HIV subtypes is the goal to predict the stability of the subtypes because current vaccine design approaches rely on "mosaic" pseudo-HIV viruses that exploit the known characteristic or representative sequence of each subtype (Barouch et al., 2010). Although a few new subtypes have been defined since the original, the M clade trees with 8-10 subtypes have been remarkably stable over time (Korber and Myers, 1992; Burr et al., 2001). Both within-host and between-host modeling efforts should allow for multiple viral sequence types within hosts, because within-host variation in contemporaneous HIV sequences isolated from various regions of the genome exhibit substantial variation, easily up to 10% differences.

4.2 Example 2: Influenza

Imagine a particularly bad flu season. Not only does it appear that more people are infected than normal by early November, but there are anomalous deaths. Could this be a bio attack or perhaps a another human-to-human transmittable version of the swine-origin influenza A (H1N1) virus?

As Figure 6 suggests, there is empirical evidence that a time gap of three or more years is sufficient for a temporal signature. For example, strains isolated in 1993 should be genetically distinct from strains in 1996 or later (Burr et al., 1999). Therefore, we might be suspicious in 2012 if the strain looks like a 2008 strain. However, the empirical evidence assumes a constant population size because the genetic distance between two samples depends on the coalescent time since they evolved from the same ancestral sequence. And the time since the two samples shared a common ancestor depends on several factors, including how the population is structured and the size and growth rate of the population. If some of these factors change dramatically, then the three-year rule would become either shorter or longer. Currently, coalescent methods either hold these factors constant over time, or extensions to the approximations have not been implemented. Therefore, empirical reconstructions of phylogenetic trees such as those in Burr et al. (1999) are incomplete for assessing the robustness of candidate signatures. The corresponding inferences thus have unknown reliability.

The $N_{effective}$ concept is part of the success of coalescent theory, including in the influenza context Bedford et al (2010) estimate $N_{effective}$ for influenza A using a coalescent model that includes subdivision and migration. Bayesian Evolutionary Analysis Sampling Trees (BEAST, Pybus et al., 2007) was used to estimate μ assuming both N and μ are constant over time. In this example, it could be important that observed and reported influenza mutation rates need not be stable over time for several reasons, including the fact that $N_{effective}$ changes with time.

In influenza, genome reassortment, selection, the presence of multiple strains and multiple hosts, and host immunity all complicate matters. Given what is known about influenza evolution, what might we predict today about influenza evolution? Two related prediction goals to consider for influenza are: (1) in a given year, predict which new strains are most likely to be in the surviving lineage, and (2) predict the prevalent strains in the next year, so that vaccine design can be most effective.

Concerning prediction goal (1), Bush et al (1999) proposed a prediction method that involved whether influenza isolates on lineages having the most changes in positively selected codes were "more fit" that other isolates. At least 18 of the 329 AA codons in H3 HA1 are thought to experience positive selection, with mutations favoring new variants that can escape host immunity. An AA sequence was defined as "more fit" if it is more closely related to surviving lineages than another contemporary strain.

Concerning prediction goal (2) the world health organization (WHO) recommends three strains to target in the vaccine for each flu season. Plotkin et al. (2002) use non-hierarchical clustering over time to evaluate the number of HA1 sequences within each cluster over time. This leads to a sequence-based algorithm to choose vaccine strains and the recommended strains differed from the WHO recommendation in 9 of 16 years in the study period from 1985 to 2000. A limitation of the Plotkin et al (2002) study is the biased sampling used by WHO in which novel strains are deliberately overrepresented in the database.

The new small-scaled agent-based simulation described in Section 5 addresses both prediction goals 1 and 2.

5. New small-scale agent-based simulation for influenza

In choosing/developing an evolutionary model it is of course important to consider the modeling goals. What are the prediction goals? How should the dynamic host/pathogen system be modeled? Which hosts should be included? Human, swine, avian, other? Is it sufficient to use a detailed model of a region such as New York state and a less detailed model of the outside region?

The basic susceptible-infected-recovered (SIR) model in classical epidemiology mathematically describes average population behavior using differential equations to move from *S* to *I* to *R*. This SIR model has been extended in various ways including structures such as contact groups and stochastic effects such as varying contact rates (Burr and Chowell, 2008,2009). Figure 7 gives examples of different simulated outbreak shapes in a small population of 1000 individuals. The number of newly infected is plotted each day for the simulated data. The small population is either (a) an unstructured population with all individuals equally in contact with all other individuals; (b) a randomly generated network model in which individuals are only exposed to member of their own clique, but some individuals belong to multiple cliques. (c) A network model with cliques assigned to nodes in a lattice; (d) a more realistic spatial network in which cliques belong to small geographic regions. A clique is a small group of individuals for which the contact probability is relatively high and assumed to be the same between each member of the clique. Various signatures of non-homogeneity using the shape of typical outbreak curves were developed for such network models, which were then shown to be detectably different from outbreak curves from the basic SIR model with homogeneous individuals all equally mutually exposed (Burr and Chowell, 2008). One concept that arose in Burr and Chowell (2008) is that predictions of the total number of infected based on the basis SIR model were often quite

wrong in the structured population. This is an example of using failed predictions to determine that more fidelity is needed than the SIR model provides.

There has been progress toward merging SIR models with viral fitness. Minayev and Ferguson (2009a, 2009b) extended SIR models by including two key notions: cross immunity to similar strains in a host that has been previously infected by a similar strain, and transient strain-transcending immunity.

At least two related empirical studies have been published, both previously mentioned (Bush et al., 1999; Plotkin et al, 2002) In Bush et al. (1999), the number of AA changes in a lineage appeared to convey a selective advantage in the following sense. The lineage for which the most AA changes occurred were more likely to be represented in the surviving lineages. That is, mutation conveys selective advantage, which is believed to be the case simply because the human host has some immunity to prior strains. At this point, "strain" will be defined following Plotkin et al. (2002) as arising from cluster analyses bases on the Manhattan metric that counts the number of AA differences between pairs of sequences. With that definition of a strain, and using 2 AA changes as the threshold above which a new strain is defined, Plotkin et al. (2002) reported empirical assessment of the number of strains by calendar year in influenza samples.

5.1 Evidence of departures from standard models

Graves and Picard (1999) report evidence of violations of the classic SIR model for influenza. Signatures of departure from SIR (Burr et al., 2006) characterize departures from the "one season fits all" assumption (which assumes each flu season occurs like clockwork, peaking in the winter during the same weeks, etc) using a hierarchical model that captures year-to-year variation in baseline, and peak onset and duration (Burr et al., 2006). Burr and Chowell (2008) use a reference distribution of simulated outbreak curve shapes to assess whether a collection of simulated and real outbreak curves follow SIR-type models. On that basis, many real outbreaks do not follow SIR-type models.

5.2 Description of the new small-scale simulation

In the context of predicting viral evolution as considered here, the SIR-type model must be extended include population demographics and characteristics of the virus. Minayev and Ferguson (2009a,b) develop one approach to include viral characteristics. Our approach to be described in this section is similar, but is entirely stochastic and allows for demographic structure. With the present implementation, population sizes of approximately 10,000 can complete in reasonable (tens of minutes) run times, so the simulation is "small-scale."

Here is pseudo code to describe the new simulation. In some cases, parameter names such as "average.duration.of.infection" are used to clarify.

Pseudo-simulation code (Example `R` (R, 2004) code named `flu1()`)

1. Initialize the population matrix and the matrix of AA sequence.

pop.matrix is *N* rows (individuals) and 30 columns with:

Column 1 is current age.

Column 2 is infection status (0 = susceptible, 1 = infected, 2 = recovered and not susceptible).

Column 3 is the number of times the individual has been infected.

Column 4 is the family group. Column 5 is the work group. Column 6 is the "other group."

Column 7 is the time of first infection. Column 8 is an integer denoting the AA sequence of infection 1. Column 9 is the donor ID for infection 1.

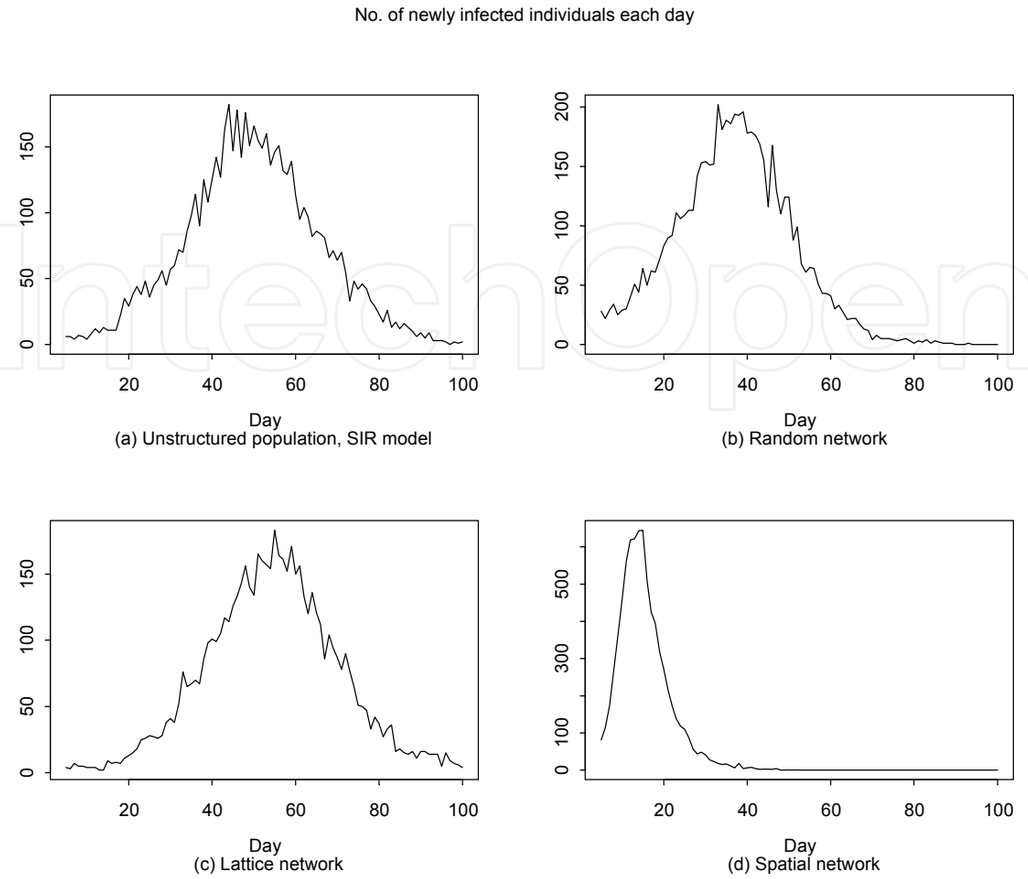


Fig. 7. Example time series of newly infected individuals in simulated SIR models of populations of 1000 individuals. (a) basic SIR model; (b) a randomly generated network model in which individuals are only exposed to member of their own clique, but some individuals belong to multiple cliques. (c) A network model with cliques assigned to nodes in a lattice; (d) a more realistic spatial network in which cliques belong to small geographic regions.

Columns 10-12 are the same as columns 7-9, but for the second infection by the individual. A maximum of 30 infections is allowed and then each new infection is recorded in the last 3 columns by writing over the previous data.
AA. seq.matrix begins with 2 rows (by default) and 329 AA sites (columns).
The default is to being with two random but distinct AA sequences.
At each time step (the default time step is 1 day), any of several events can occur:

2. There is a probability for each individual to change status from *S* (0) to *I* (1), or from *I* (1) to *R* (2), or from *R* back to *S*.

Any susceptible (*S*) individuals that an infect (*I*) individual contacts in the respective family, work, and other groups leads to a probability of infection determined by two parameters. First, there is a force of infection parameter for each of the three group types that characterizes how strongly individuals in the three group types interact. Second, the similarity of the infected individual's current strain to the closest strain of a given susceptible is computed and the cross-immunity function γ is calculated.
The value of the function γ in Minayev and Ferguson (2009b) alters the transmission probability accordingly. Cross immunity modeled by γ decreases to zero as a smooth

function of time (examples given below), with an average of 10 year total immunity from identical strains. And, values of γ.a and γ.b in γ can be altered to decease or increase the degree of cross-immunity as a function of the Manhattan distance between strains. The cross-immunity concept is that *S* individuals who have had the strain of the potential donor *I* are less susceptible to infection. As Plotkin et al (2002) describe, ideally a distance measure between two sequences should somehow reflect immuniological properties of the corresponding viral proteins. Although steps have been taken in that direction (Lapedes and Farber, 2001), more research is required before similar metrics can be defensibly applied in modeling contexts such as our new small-scale simulation.

Any newly infected host will have the donor strain, but the simulation allows for mutation to a new strain. There are 329 H3 HA1 (Bush et al., 1999) amino acid (AA) sites with one estimate of the effective mutation rate $N\mu$ being 0.0057 nucleotide substitutions per site per year. Of the 329 AA sites, at least 18 have exhibited positive selection effects (Plotkin et al., 2002). Here we will not consider estimation error in $N\mu$, so the simulation default value is $N\mu = 0.0057 \; x \; 3 = 0.171$ per AA site per year. A technical issue arises here because we use the actual popluation size N rather than the effective size $N_{effective}$. It would be more appropriate to use $N_{effective}$, but that value is currently unknown in the context of this model population. In future work, $N_{effective}$ could be defined and estimated on the basis of the number of observed distinct sequences during outbreak.

If a newly infected incurs any mutations, add the new strain to AA.seq.matrix, increasing the number of rows by one. Columns in pop.matrix identify which strains each host has had in the sequence matrix of all strains ever experienced in the model population

3. Any infected can recover.

The per-step recovery probability is 1/(average.duration .of. infection). The time to recovery is therefore a geometric random variable with average duration
average.duration.of.infection.

4. Any recovered can lose immunity.

The time *t* from recovery to immunity is random, with t ~ Normal(avg.time.to.immunity,σ). The default simulation values are
avg.time.to.immunity=10 years and σ = 1 year.

If at any time step (day) the number of infected is 0, then the infection would die out. Therefore, a reintroduction of new infected occurs at a random time with a user-specified average value with default value of 1 year, representing the typical time gap between outbreaks.

Figure 8 plots the percent currently infected at each time step for one 7-year realization of 10,000 individuals. A key output of `flu1` is the current strain of each infected individual at each time step. This allows us to consider strategies in Bush et al (1999) and in Plotkin et al. (2002) for prediction goals (1) and (2) described earlier in this section. Figure 9 plots four of the eight strains that emerged during the 7 simulated years. Following Plotkin et al. (2002), sequences were regarded as being the same strain if the number of AA differences among the 18 positively-selected AA sites is 2 or less. Equivalently, sequences were regarded as being distinct strains if the number of AA differences is 3 or more. Figures 8 and 9 used the values γ.a = 0.4, γ.b = 0.95.

Experimentation with `flu1` to generate multiple realizations of outbreaks having identical parameter values allows us to examine the role of chance in our models. Experimentation with `flu1` with different parameter values allows us to examine the effects of parameter changes. Small numerical experiments with `flu1` to date has lead to the following following conclusions:

1. The γ.a and γ.b in γ are as critical to the size of each outbreak as the overall transmission probabilities within the family, group, and other groups. For example, the function γ takes values 0.990, 0.891, 0.792, 0.693, 0.594, 0.495, 0.396, 0.297, 0.198, 0.099, 0.000,… for distances of 1, 2, …, 11,… respectively, for γ.a = 0.1, γ.b = 0.99, and takes values 0.8, 0.4, and 0.0, for distances of 1, 2, 3 …, respectively for γ.a = 0.5, γ.b = 0.8. The modeled transmission probability is multiplied by 1-γ so for γ.a = 0.1, γ.b = 0.99 there is very little chance of a susceptible individual acquiring influenza from a host having an AA sequence that differs in only 1 position among 18 positions from a strain the susceptible has had within the duration of immunity (10 years on average for example). This means that a single AA mutation can have dramatically different effects on transmission probability depending on γ.a and γ.b. Qualitatively, this is anticipated because if immunity to a new strain is very high in individuals with previous infection by a similar older strain, then the average outbreak size will be small if previous outbreaks due to the older strain were large.
2. The values of γ.a and γ.b in γ are also critical to the typical number of strains maintained in the population and to whether change occurrence of a large number of mutations in a newly infected will have strong selective advantage by avoiding the collective immune experiences of available human hosts. It is currently unknown whether the observed number of strains could adequately provide a model such as `flu1` with estimates of γ.a and γ.b (See section 6).
3. As expected, the group structures can lead to outbreak shapes that differ from classic SIR outbreak shapes (Burr and Chowell, 2008). This is evident in comparing the outbreak shapes in Figure 8 to the SIR-model outbreak in Figure 7a for example. The shapes in Figure 8 fall off very sharply, more like the spatial network in Figure 7d.
4. In population genetics, the composite parameter $\theta = 2N_{\text{effective}}\mu$ where μ is the mutation rate determines the rate of genetic changes and the expected amount of diversity in a random sample (see Section 6). The $N_{\text{effective}}$ concept for influenza sequences was addressed in Bedford et al. (2010), but as mentioned in Section 4, observed genetic diversity is interpreted in the context of idealized evolutionary models that are amenable to coalescent theory. More experiments with `flu1` are planned, and if possible, an approximate coalescent model as implemented in available software will be applied so that the adequacy of coalescent-based approximations can be evaluated in the context of simulated flu outbreaks. We caution that many genetic effects of influenza evolution are omitted from `flu1` and from any available coalescent-based simulation.

6. Model Identifiability and Inference

Model identifiability is a key statistical concept. A model is identifiable if its parameters can be accurately and precisely estimated as the sample size increases toward infinity. In population genetics, a key parameter that arises from coalescent theory considerations is the composite parameter $\theta = 2N_{\text{effective}}\mu$, which determines the rate of genetic changes. Many studies address methods to estimate θ but because $N_{\text{effective}}$ and the mutation rate μ enter θ as a product, they are confounded, leading to a lack of identifiability unless auxiliary data is used to separately estimate $N_{\text{effective}}$ or μ and a strict evolutionary clock is assumed, meaning that μ is constant over time and lineages.

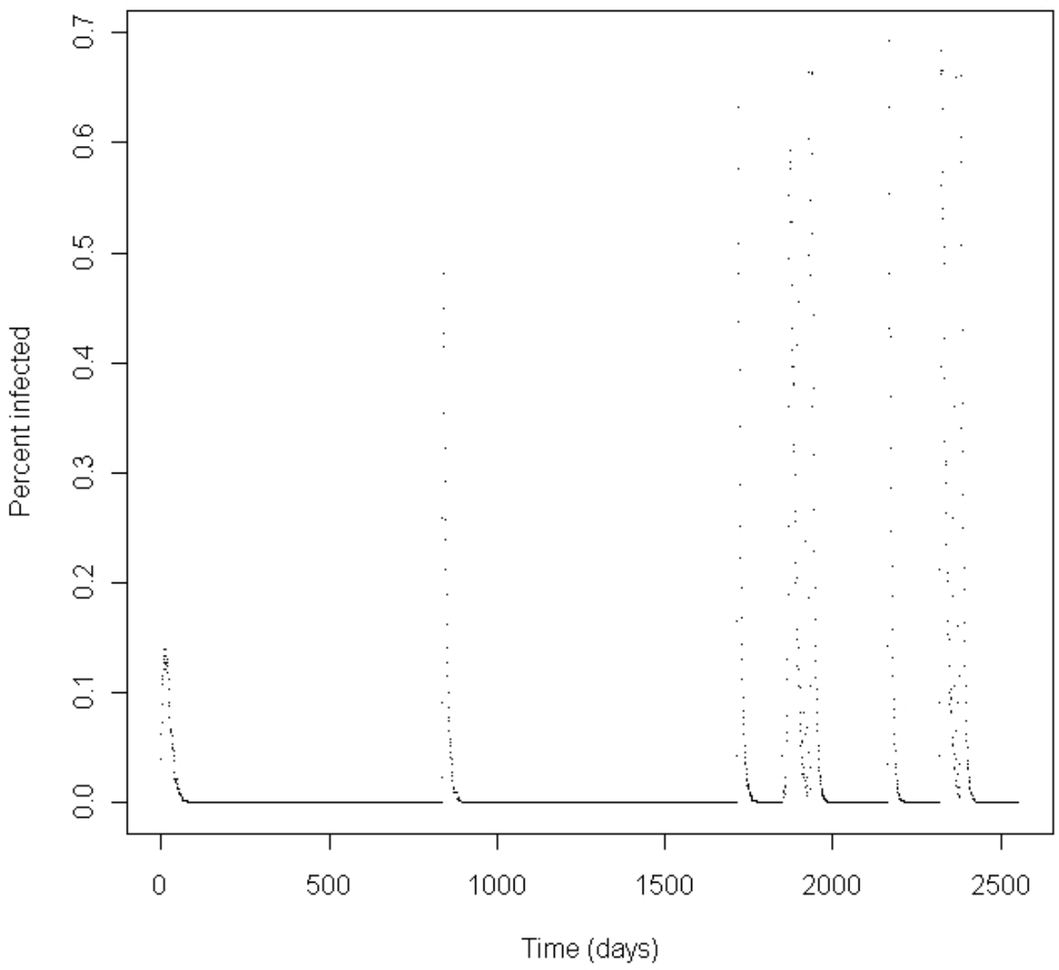


Fig. 8. Example simulated outbreaks from `flu1`. The number currently infected is plotted by day for years.

For a particular evolutionary model, inference is possible using Bayesian evolutionary analysis, for example, by using BEAST (Pybus et al., 2000) that relies on Markov Chain Monte Carlo, resulting in a posterior distribution on model parameters. The Bayesian approach allows one to repeatedly sample from the posterior probability of model parameters, and then generate hypothetical future genetic data for each set of model parameter values. This approach provides an envelope of possible future multivariate time series of genetic data from each sampled subject. It is computationally challenging even for a given model of evolution.

In Burr and Chowell (2008), in simulated data from models with demographic structure but no host immunity or viral strain information, predictions from SIR models with parameters estimated from the early portion of an outbreak were often badly wrong. Such bad prediction errors can indicate model violations, perhaps eventually leading to more appropriate models. To our knowledge, using prediction quality to assess model adequacy in this context is new. However it is possible that multiple wrong models provide adequate predictions, so model identifiability remains a research topic in this area.

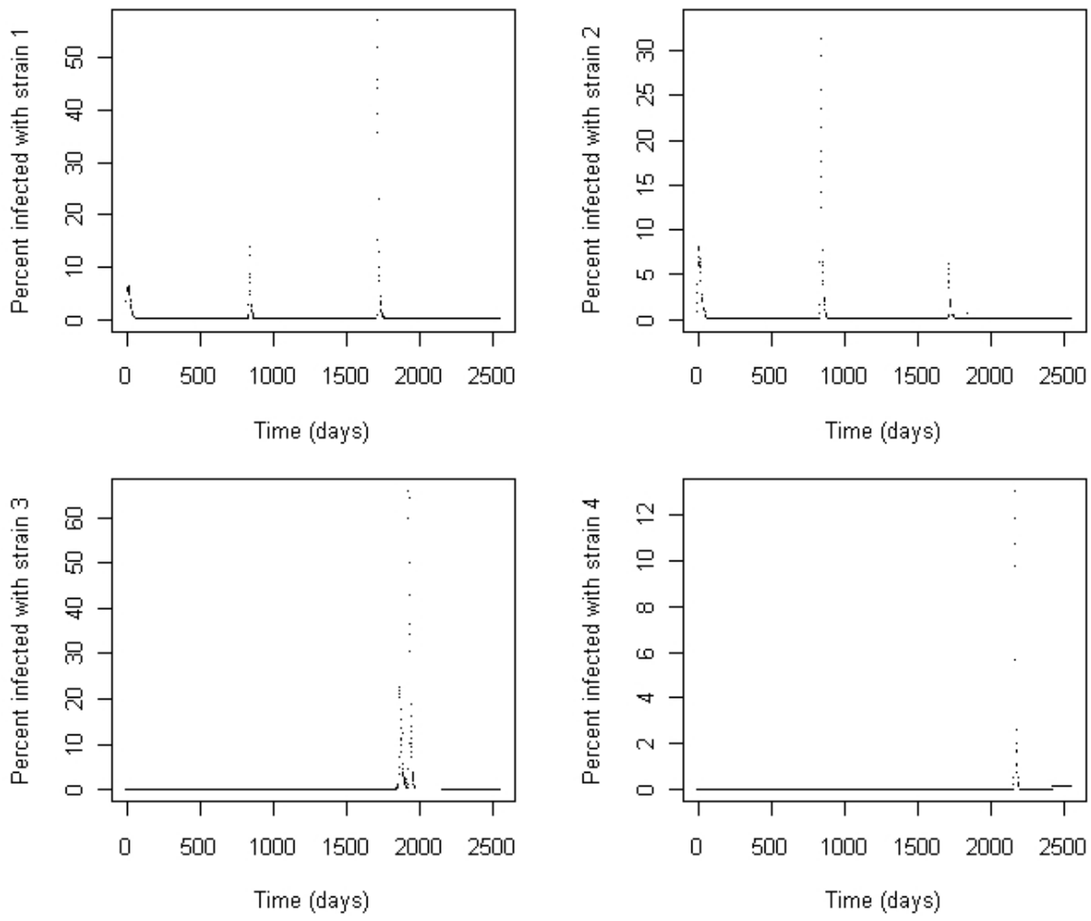


Fig. 9. Percent infected by day with strain 1, 2, 3 or 4 for the same simulated 7 years as in Figure 8.

7. Conclusions/summary

Coalescent theory and its success in some contexts at inferring aspects of past virus evolution were described. Then the argument was made that relatively new goals to predict aspects of virus evolution will require higher fidelity modeling that is anticipated to be available via coalescent theory or its extensions.

As a step toward high fidelity modeling, a small-scale agent based simulation was described and example results presented. For influenza, two prediction goals were considered: (1) in a given year, predict which new strains are most likely to be in the surviving lineage, and (2) predict the prevalent strains in the next year, so that vaccine design can be most effective. The new small-scale simulation code `flu1`in `R` can provide insight into the feasibility of meeting these goals, but it too makes restrictive modeling assumptions with unknown accuracy. A related concept was described that involves using prediction quality on simulated data that follows an assumed model to assess whether prediction performance on corresponding real data indicates model violations. Model violations that are evident from poor prediction quality can help prioritize future upgrades to the models.

8. References

Barouch, D. et al., (2010). Mosaic HIV-1 Vaccines Expand the Breadth and Depth of Cellular Immune Responses in Rhesus Monkeys, *Nature Medicine* 16, 319-323.

Bedford, T; Cobey, S; Peerli, P., & Pascual, M. (2010). Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2), *PloS Pathogens* 6(5), 1-9, 2010.

Burr, T.; Skourikhine A.; Bruno, W., & Macken, C. (1999). Confidence Measures for Evolutionary Trees: Applications to Molecular Epidemiology. *Proc. IEEE Inter. Conf. on Information, Intelligence and Systems, Genetics and Evolution Section* ; 107-114. Burr, T. (2000). Quasi-Equilibrium Theory for the Distribution of Rare Alleles in a Subdivided Population: Justification and Implications. *Theoretical Population Biology*, 57(3): 297-306.

Burr, T. ; Myers, G., & Hyman, J. (2001). The Origin of AIDS – Darwinian or Lamarkian? *Phil. Trans. R. Soc. Lond. B* 356:877-887.

Burr, T.; Gattiker, J., & LaBerge G. (2002). Genetic Subtyping using Cluster Analysis. *Special Interest Group on Knowledge Discovery and Data Mining Explorations* 3:33-42.

Burr, T.; Gattiker, J., & Gerrish,P. (2003). An Investigation of Error Sources and Their Impact in Estimating the Time to the Most Recent Ancestor of Spatially and Temporally Distributed HIV Sequences. *Statistics in Medicine* 22(9):1495-1516

Burr ,T.; Graves, T.; Klaman, R.; Michalek, S.; Picard ,R., & Hengartner, N. (2006). Accounting for Seasonal Patterns in Syndromic Surveillance Data for Outbreak Detection, *BioMedCentral, Medical Informatics and Decision Making*, 6:40.

Burr, T., & Chowell, G. (2008). Signatures of non-homogeneous mixing in disease outbreaks. *Mathematical and Computer Modelling* 48:122-140, 2008

Burr, T., & Chowell, G. (2009) The Reproduction Number R(t) in Structured and Non-structured Populations. *Mathematical Biosciences and Engineering* . 6(2) 239-259.

Bush, R.; Bender, C.; Subbarao, K; Cox, N; & Fitch, W. (1999). Predicting the Evolution of Human Influenza A, *Science* 286: 1921-1925.

Chen, G. et al, (2006). Genomic Signatures of Human versus Avian Influenza A Viruses, *Emerging Infectious Diseases* 12(9): 1353-1360.

Eubank, S.; Goclu, H.; Kumar, A.; Marathe, M.; Srinivasan, A.; Totoczkal, Z., & Wang, N. (2004). Modelling Disease Outbreaks in Realistic Urban Social Networks. *Nature*, 429:180-184.

Ewing, G.; Nicholls, G., & Rodrigo A. (2004). Using Temporally Spaced Sequences to Simultaneously Estimate Migration Rates, Mutation Rage and Population Sizes in Measurably Evolving Populations. *Genetics* 168:2407-2420.

Excoffier, L., & Foll, M. (2011). fastsimcoal: a Continuous-time Coalescent Simulator of Genomic Diversity under Arbitrarily Complex Evolutionary Scenarios, *Bioinformatics* Advance Access, March 2011.

Felsenstein, J.; Kuhner, M.; Yamato, J., & Beerli P. (1999). Likelihoods on Coalescents: a Monte Carlo Sampling Approach to Inferring Parameters from Population Samples of Molecular Data. pp. 163-185 in *Statistics in Molecular Biology and Genetics*, ed. Francoise Seillier-Moiseiwitsch. IMS Lecture Notes-Monograph Series, volume 33. Inst. of Math. Statistics and American Mathematical Society, Hayward, California.

Ferguson, N., & Anderson, R. (2002). Predicting Evolutionary Change in the Influenza A Virus, *Nature Medicine* 8(6): 562-563.

Forrest, H., & Webster, R. (2010). Perspectives on Influenza Evolution and the Role of Research, *Animal Health Research Reviews* 11(1): 3-18.

Grassley, N.; Harvey, P., & Holmes, E. (1999) Population Dynamics of HIV-1 Inferred from Gene

Sequences. *Genetics* 151: 427-438.

Graves, T., & Picard, R. (2003) Predicting the Evolution of P&I Mortality During A Flu Season, Los Alamos National Laboratory Unrestricted Release Report, LAUR02-4717.

Innan, H., & Stephan, W. (2000). The Coalescent in an Exponentially Growing Metapopulation and Its Application to Arabidopsis thaliana, *Genetics* 155, 2015-2109.

Kingman, J. (1982). On the Genealogy of Large Populations. *J. Appl. Probability* 19: 27-43.

Korber, B., & Myers, G. (1992) Signature Pattern Analysis: a Method for Assessing Viral Sequence Relatedness. *AIDS Res. Hum. Retro.*, 8, 1549–1560.

Lapedes, A., & Farber R. (2001). The Geometry of Shape Space: Application to Influenza, *Journal of Theoretical Biology* 212(1), 57-69.

Minayev, P., & Ferguson, N. (2009a). Improving the Realism of Deterministic Multi-strain Models: Implications for Modelling Influenza A, *J. R. Society Interface* 6, 509-518, 2009.

Minayev, P., & Ferguson, N. (2009b) Incorporating Demographic Stochasticity into Multi-strain Epidemic Models: Application to Influenza A, *J. R. Society Interface* 6, 989-996.

Nelson, M., & Holmes, E. (2007). The Evolution of Epidemic Influenza, *Nature Reviews Genetics* 8, 196-205.

Perelson, A. et al. (1996). HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time *Science* 271, 1582-1586.

Plotkin, J.; Dushoff, J., & Levin, S. (2002). Hemagglutinin Sequence Clusters and the Antigenic Evolution of Influenza A, *Proc. Nat. Acad. Sci, USA* 2002; 99: 6263-6268

Pybus, O.;Rambaut, A., & Harvey, P. (2000). An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics* 155: 1429-1437.

R: a Language and Environment for Statistical Computing, R Development Core Team, www.R-project.org.

Rambaut, A.; Pybus, O.; Nelson, M.; Viboud, C.; Taubenberger, J., & Holmes, E. (2008).The Genomic and Epidemiological Eynamics of Human Influenza A Virus. *Nature* 453:615-9

Rambaut, A.; Robertson, D.; Pybus, O.; Peeters, M., & Holmes ,E. (2001). Phylogeny and the Origin of HIV-1. *Nature*,410:1047-8

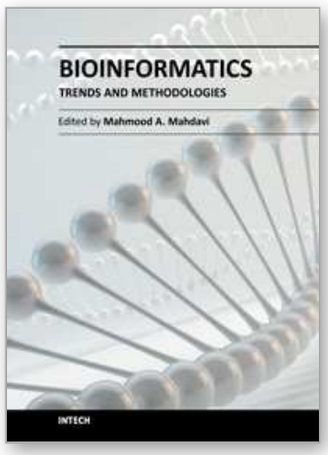
Rodrigo et al. (1999). Coalescent Estimates of HIV-1 Generation Time in Vivo. *Proc. Nat. Acad. Sci USA* 96:2187-2191.

Sjodin, P.; Kaj, K.; Krone, S.; Lascoux, M., & Nordborg, M. (2005). On the Meaning and Existence of an Effective Population Size. *Genetics* 169: 1061-1070.

Swofford, D.; Olsen, G.; Waddell, P., & Hillis, D., Phylogenetic Inference, Chapter 11 in Molecular Systematics, Edited by Hillis, D.; Moritz. C., & Mable, B, SInaer, Sunderland, Mass. 1996.

Stephens, M., & Donnelly P. (2000). Inference in Molecular Population Genetics. *J. Royal Statistical Soc* B 62(4); 605-655.

Venables, W., & Ripley, B. (1999) Modern Applied Statistics with Splus, Springer: New York.



Bioinformatics - Trends and Methodologies
Edited by Dr. Mahmood A. Mahdavi

ISBN 978-953-307-282-1
Hard cover, 722 pages
Publisher InTech
Published online 02, November, 2011
Published in print edition November, 2011

Bioinformatics - Trends and Methodologies is a collection of different views on most recent topics and basic concepts in bioinformatics. This book suits young researchers who seek basic fundamentals of bioinformatic skills such as data mining, data integration, sequence analysis and gene expression analysis as well as scientists who are interested in current research in computational biology and bioinformatics including next generation sequencing, transcriptional analysis and drug design. Because of the rapid development of new technologies in molecular biology, new bioinformatic techniques emerge accordingly to keep the pace of in silico development of life science. This book focuses partly on such new techniques and their applications in biomedical science. These techniques maybe useful in identification of some diseases and cellular disorders and narrow down the number of experiments required for medical diagnostic.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Tom Burr (2011). Predicting Virus Evolution, Bioinformatics - Trends and Methodologies, Dr. Mahmood A. Mahdavi (Ed.), ISBN: 978-953-307-282-1, InTech, Available from: http://www.intechopen.com/books/bioinformatics-trends-and-methodologies/predicting-virus-evolution

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen