Rannala B (2002). Molecular Phylogenies and Virulence Evolution. In: Adaptive Dynamics of Infectious Diseases: In Pursuit of Virulence Management, eds. Dieckmann U, Metz JAJ, Sabelis MW & Sigmund K, pp. 262–276. Cambridge University Press.

19

Molecular Phylogenies and Virulence Evolution

Bruce Rannala

19.1 Introduction

The effective management and prevention of outbreaks of virulent strains of microbes depends on information about when, where, and how such strains arise. In the case of newly emerging human pathogens, this might involve tracing the source of a zoonotic infection to an animal population – such was the case with a hantavirus outbreak in the US state of New Mexico (Nichol *et al.* 1993). In other cases, an existing, possibly benign, microbe infecting humans or livestock may suddenly give rise to a highly pathogenic (virulent) strain – such was the case with the 1918 "Spanish" flu pandemic. In this second case, information about the mechanism by which virulence arose can provide practical guidance to epidemiologists developing strategies to prevent, or forestall, future epidemics. As well, information about the time and location of origination of virulent strains can inform us about how quarantine measures ought to be applied in the future, or how effective such measures have been in particular instances in the past.

In recent years, molecular phylogenies have come to play an increasingly important role in epidemiological studies of microbial pathogens, as they provide information about the location, timing, and mechanisms by which virulent strains arise. In particular, sequences from disease-causing viruses and bacteria that infect humans and livestock have been studied extensively, with hundreds of phylogenies published in medical and veterinary journals in 2000 alone. (A search of PubMed for articles published in 2000 that contained both the words "virus" and "phylogeny" produced 699 articles.) There are at least two major reasons for this rapid growth in the use of phylogenies by epidemiologists. The first is that many viruses [especially ribonucleic acid (RNA) viruses] and bacteria experience mutations at a much higher rate than eukaryotes. This difference is compounded by generation times that are typically orders of magnitude shorter. The expected substitution rate per-site per-year, which for neutral genes is roughly the per-site mutation rate divided by the generation time (in years), is therefore much higher for viruses and bacteria than for eukaryotes, even if, as is the case for certain bacteria, their mutation rates are roughly equal. Synonymous substitution rates per-site per-year for nuclear genes in mammals, for example, are about 10^{-9}, whereas rates for RNA viruses such as influenza A and human immunodeficiency virus (HIV) are about 10^{-2} (Li 1997). The high rates of substitution found in viruses and bacteria allow phylogenies to be reconstructed for sequences that have diverged only recently. Phylogeny has therefore become relevant to the questions typically addressed by

epidemiologists such as the source of origin, and the rate of spread, of pathogenic strains of microbes. A second reason for the recent growth in the application of phylogenetics to microbial epidemiology is the development of new methods for isolating, amplifying, and sequencing nanogram quantities of deoxyribonucleic acid (DNA; and also RNA) obtained from blood or tissue samples, and, in particular, the polymerase chain reaction (PCR) method of amplifying DNA (Mullis 1986).

Advances in molecular genetic studies of viruses and bacteria have been paralleled by advances in the computational methods used to analyze nucleotide sequences and reconstruct phylogenetic trees of divergent strains or species. Explicit probabilistic models of nucleotide substitution have been developed (Jukes and Cantor 1969; Kimura 1980; Swofford *et al.* 1996) and used to derive quantitative statistical methods that allow phylogenies to be inferred for sequences under well-established criteria using likelihood (Felsenstein 1981) or Bayesian approaches (Rannala and Yang 1996). With these advances have come new opportunities to apply existing principles from the fields of evolutionary biology and population genetics to the effective management of virulent strains of microbes. This chapter focuses on two aspects of virulence management that have benefited from a phylogenetic perspective:

- Tracing when and where virulent strains arose;
- Identifying the genetic mechanisms by which they arose.

This information may often suggest practical forms of intervention to reduce the likelihood that virulent strains will emerge in the future. Examples are given that analyze sequences of virulent strains of influenza A from chickens and humans.

19.2 Phylogenetic Tools

To apply parametric statistical methods to estimate phylogeny using sequence data, a mathematical model is needed. The basic components of the model employed in a typical analysis are as follows:

- A set of potential phylogenetic trees, with branch lengths measured in units of the expected number of substitutions;
- A model of the process of nucleotide substitution that assigns a probability to any observed set of sequences given a phylogenetic tree (see Box 19.1).

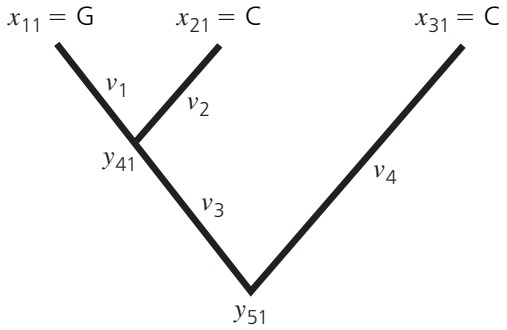
In this chapter, the substitution model proposed by Hasegawa *et al.* (1985) is used, which allows for biases in transitional substitutions (e.g., A to T) versus transversional substitutions (e.g., A to C). The method of Yang (1994) is implemented with this model to allow for gamma distributed rate variation among sites (HKY+Γ). Likelihood ratio tests (LRTs; see below) can be used to choose a substitution model that best fits the observed sequences without introducing superfluous parameters (Goldman 1993), thus reducing the arbitrary aspects of model choice in a phylogenetic analysis. In this chapter, maximum likelihood (ML) is used to estimate the phylogenetic tree and branch lengths. The researcher chooses as the best estimates

Box 19.1 Likelihood methods for phylogenetic inference

The starting point for estimating a phylogenetic tree from DNA sequence data by ML is a sample of a aligned sequences, each n nucleotides in length. The sequence data may be summarized as an $a \times n$ matrix $X = \{x_{jk}\}$, where x_{jk} is the nucleotide at the kth site of the jth sequence. A data matrix of only three sequences might have the form

$$X = \begin{pmatrix} \mathrm{G} & \mathrm{T} & \mathrm{T} & \ldots & \mathrm{C} \\ \mathrm{C} & \mathrm{T} & \mathrm{T} & \ldots & \mathrm{C} \\ \mathrm{C} & \mathrm{T} & \mathrm{T} & \ldots & \mathrm{A} \end{pmatrix},$$

where $x_{11} = \mathrm{G}$, $x_{21} = \mathrm{C}$, $x_{22} = \mathrm{T}$, etc. One of the three possible distinct rooted phylogenetic trees, to be denoted by τ_j, $j = 1, \ldots, 3$, for three sequences with branch lengths $V = \{v_1, v_2, v_3, v_4\}$ is



where the nucleotide observed at the first position of each sequence is shown at each tip of the tree (x_{11}, x_{21}, and x_{31}) as well as the ancestral nucleotides (y_{41} and y_{51}). The states of the ancestral nucleotides y_{41} and y_{51} are unobserved, and so the probability of the nucleotides at tips x_{11}, x_{21}, and x_{31} is calculated by summing over the probability obtained for each possible assignment of the ancestral nucleotides. The probability that the nucleotide observed at the root y_{51} is denoted by $\pi_{y_{51}}$, and is assumed to be that of the equilibrium distribution for the substitution model (this probability is usually estimated using the empirical frequencies of the nucleotides averaged over all sites in all sequences). The overall probability of the nucleotides observed at the first site in the example tree, which we denote as τ_1, is

$$\Pr(X_1|\tau_1, V, \Theta) =$$

$$\sum_{y_{51}, y_{41}} \pi_{y_{51}} p_{y_{51}y_{41}}(v_3|\Theta) p_{y_{51}\mathrm{C}}(v_4|\Theta) p_{y_{41}\mathrm{G}}(v_1|\Theta) p_{y_{41}\mathrm{C}}(v_2|\Theta),$$

where $X_1 = \{\mathrm{G}, \mathrm{C}, \mathrm{C}\}$, $p_{\mathrm{AC}}(v_j|\Theta)$ is the probability that nucleotide C is substituted for A over a branch of length v_j, and Θ is a vector of the parameters of the substitution model.

An example of a simple substitution model with a single parameter $\Theta = \mu$ is that of Jukes and Cantor (1969), which assumes that all possible nucleotide substitutions occur with an equal rate. A substitution model that often provides a

continued

Box 19.1 *continued*

good fit to real sequences, proposed by Hasegawa *et al.* (1985), has two parameters $\Theta = (\mu, \kappa)$, an overall substitution rate μ (proportional to the branch lengths), and a parameter κ, the bias in rates of transition versus transversion. Most methods of likelihood analysis assume that substitutions at different sites in a sequence are independent and identically distributed. The probability of observing the complete sample of sequences, X for a tree τ with branch length vector V, is then a product over the probabilities of the nucleotides observed at each successive site

$$\Pr(X|\tau, V, \Theta) = \prod_{j=1}^{n} \Pr(X_j|\tau, V, \Theta) .$$

The likelihood is defined as the probability of the observed sequences treated as a function of the model parameters τ, V, and Θ,

$$L(\tau, V, \Theta|X) = \Pr(X|\tau, V, \Theta) .$$

The likelihood function is maximized as a function of the parameters to obtain maximum likelihood estimates (MLEs) of the substitution model parameters, the branch lengths, and the phylogeny. The likelihood method proposed by Felsenstein (1981) estimates the branch lengths and parameters of the substitution model separately for each phylogenetic tree, and the MLE of phylogeny is chosen to be the tree with the highest relative likelihood. The logarithm of the likelihood is often used when probabilities are small.

of the phylogeny, branch lengths, and parameters of the substitution model those values that maximize the probability of the observed data (see Box 19.1).

Hypothesis tests using sequences

In a likelihood framework, one can also examine the support of the sequence data for different evolutionary hypotheses that may depend on the phylogeny, or the substitution model, by use of an LRT (reviewed by Huelsenbeck and Rannala 1997). The basic procedure is to calculate the relative probability of the observed sequence data under the null hypothesis versus the alternative hypothesis. Often the hypotheses in question are "nested," so that the null hypothesis is a special case of the alternative hypothesis. By considering the probability distribution of the LRT statistic under the null hypothesis, the significance of the value obtained for the sampled sequences can be determined. For nested hypotheses, the null distribution of the test statistic $-2 \ln \Lambda$, where Λ is the ratio of the likelihood under the null hypothesis to that under the alternative hypothesis, is approximately χ^2 with k degrees of freedom, where k is the difference in the number of free parameters under the null and alternative hypotheses. For non-nested hypotheses, the parametric bootstrap can be used to generate the null distribution of the likelihood ratio (see Goldman 1993).

In this chapter, LRTs are applied in some familiar ways, such as to test the fit of a molecular clock to sequence data, or of different models of substitution that incorporate effects such as transitional bias, or rate variation among sites (Goldman 1993; Huelsenbeck and Rannala 1997). The LRTs are also applied in some less familiar ways, such as to test whether virulent strains of a virus from a particular epidemic had a single recent (and perhaps local) origin, or instead were introduced multiple times, and to examine the agreement between phylogenies of viral sequences obtained using different genes to test whether recombination (exchanges of segments in individuals infected with multiple strains) is an important source of new virulent strains.

Molecular clock for virus sequences

The molecular clock hypothesis assumes that rates of substitution do not vary among phylogenetic lineages. If a molecular clock is imposed in a likelihood analysis, the branch lengths are constrained and the root of the tree chosen so that the sum of the branch lengths along any ancestor–descendent path from the root of the tree to any tip is the same. Many samples of viruses are temporally stratified – that is, they are made up of sequences isolated at different times (typically strains of viruses isolated in different years). As substitution rates for viruses and bacteria are so high, differences in sampling times can have an important effect on branch lengths. The likelihood-based molecular clock proposed by Felsenstein (1981), which assumes that the sequences are sampled simultaneously, is not expected to fit such data, even if substitution rates are constant among lineages.

Rate variation among lineages can be accommodated by performing an "unconstrained" likelihood analysis, in which the length of each branch in a tree (the product of the branch duration and the substitution rate) is treated as a separate parameter to be estimated jointly from the sequence data (Felsenstein 1981). Rambaut (1996) refers to this as the different rate (DR) model. Alternatively, if the times are specified at which the sequences are sampled, the ages of the tips of the tree can be set equal to the sampling times; this allows a joint estimation of the branch lengths under the constraint that the length of any path from the root of the phylogeny to any tip is proportional to the time at which the sequence at that tip was sampled minus the time at which the root ancestor existed. This can potentially distinguish a deviation from the molecular clock caused by differences in age among sequences from one caused by rate variation among lineages, and can allow the ages of common ancestors in the phylogeny, and of the root, to be estimated (Rambaut 1996). This model is referred to as the single rate, fossil sequence (SRFS) model (Rambaut 1996).

Sources of phylogenetic uncertainty

The most important factors that affect the accuracy of hypothesis tests involving phylogeny can be grouped into four broad categories:

- Errors through the finite length of sequence sampled;

- Errors resulting from an inaccurate model of either the substitution process or the evolutionary process;
- Errors in sequence alignment;
- Errors from population sampling.

Sampling errors of the first type are adequately accounted for when using likelihood methods. It is more difficult to guard against errors of the second and third types. Improved models of nucleotide substitution can be used to account more adequately for many of the nonuniform substitution patterns commonly observed among sampled sequences, including transition versus transversion biases among nucleotides and rate variation among sites or genomic regions; LRTs can be used to decide when a more complicated substitution model provides a significant improvement in the fit of the sequence data. However, certain complications of the substitution process, such as intragenic recombination and nonindependence of substitutions among sites, cannot be easily accommodated using presently available methods; if such factors are important, overly simple models could potentially lead to incorrect conclusions (see the review by Huelsenbeck and Rannala 1997). Alignment errors are neglected in most studies, although, in some cases, they may be an important source of phylogenetic uncertainty (Goldman 1998).

The fourth source of uncertainty arises because phylogenies are typically constructed for a very small sample of sequences that represents only a fraction of the total population. This is particularly important for viruses and bacteria, which may have very large (and subdivided) populations even over a limited geographical scale. Each population sample has a phylogeny and branch lengths associated with it the form of which may vary substantially among samples. As a result, the outcomes of hypothesis tests involving phylogeny also typically vary from sample to sample, which introduces additional uncertainty into the analysis. In the strictest sense, the statistical tests discussed in this chapter, only apply to the samples of sequences and the models being considered and should not be too readily extended to the population of sequences in a geographical region or an epidemic as a whole.

19.3 Case Studies

Below we consider three case studies that show how the framework outlined above can shed light on the evolution of three related viral diseases.

Influenza A

The influenza A virus is of great medical and economic importance. In the 20th century alone, four successive pandemics of human influenza resulted in the deaths of between 20 million and 40 million persons. In addition, virulent strains of avian influenza A frequently arise that, in extreme cases, may kill millions of birds; the global costs of losses for poultry producers related to influenza A can reach millions of dollars annually. Influenza A is a negative-strand RNA virus with eight segments carrying a total of 10 protein-coding genes that make up the influenza A genome (see Voyles 1993). Two of these genes code for proteins that are expressed in the viral envelope, which is derived primarily from the host cell membrane.

These two proteins, hemagglutinin (HA) and neuraminidase (NA), are of critical importance as antigenic determinants of the host immune response. HA is a trimer that appears to be involved in host cell recognition, and NA is a tetramer that is possibly involved in mediating the release of newly formed viruses.

Influenza A was first isolated from humans in the 1930s, and it was recognized early on that different genetic variants, or epitopes, of the virus at the HA and NA loci elicited different immune responses. The variants were originally classified according to whether exposure to one produced antibodies that were cross-reactive to the other. Numerous epitopes were identified that were not cross-reactive, and such serologically novel strains were sequentially numbered according to their HA and NA types. Examples are H1N1 (1918 "Spanish" flu) and H2N2 (1957 "Asian" flu; see Levine 1992). It was soon recognized that antigenic shifts could occur through mutational changes to new subtypes (N1 to N2, H1 to H2, etc.), a process known as antigenic drift (see e.g., Both *et al.* 1983), as well as by the exchange of viral segments between strains in individuals infected with multiple strains (H1N1/H2N2 giving rise to H1N2, H2N1, etc.; see e.g., Li *et al.* 1992). Additionally, virulent strains of avian influenza are known to arise by mutations in the HA gene that increase its cleavability. This appears to be a critical step in facilitating the spread of the viral infection from the respiratory tract in progression to a more severe systemic infection (Bosch *et al.* 1981; Kawaoka *et al.* 1987).

Phylogenies have been used to study the evolution of virulent strains of influenza A in several different ways. The most common applications are in studies of the geographical or zoonotic origins of certain virulent strains (see e.g., Rohm *et al.* 1995), and to the study of the mechanisms by which virulent forms have arisen in particular cases – whether by reassortment of segments among strains in swine that were multiply infected with viruses from ducks and humans (see e.g., Yasuda *et al.* 1991), for example, or instead by point mutation (see e.g., Horimoto *et al.* 1995). Phylogenies have also been used in attempts to study the propensity of different lineages to give rise to new genetic forms that are novel antigens and potentially capable of producing a pandemic (Fitch *et al.* 1997). In this chapter, I consider some simple ways that phylogenetic trees can be used to study the question of where and when virulent strains arose, as well as the question of how (i.e., whether by recombination or point mutation).

Mexican chicken flu

The most recent major North American epidemic of highly virulent H5N2 avian influenza occurred in 1983–1984 among turkeys and chickens in Pennsylvania (Bean *et al.* 1985). The indirect costs of this epidemic to the poultry industry have been estimated at over a quarter of a billion dollars (Horimoto *et al.* 1995). In 1993, an outbreak of type H5N2 avian influenza occurred among Mexican chickens. Most isolates of the virus produced only mild respiratory symptoms and, for economic reasons, infected chickens were not eliminated, nor were infected poultry farms quarantined. As a result, the virus was able to spread unchecked and several highly pathogenic isolates ultimately appeared (Horimoto *et al.* 1995). At

least two pathogenic strains of H5N2 from Mexican chickens isolated in 1994 and 1995 (labeled CP607 and CQ19, respectively, see below) appear to have arisen by an insertion in the HA connecting peptide, which rendered it highly cleavable (Horimoto *et al.* 1995).

Horimoto *et al.* (1995) examined HA gene segments for three H5N2 isolates from Mexican chickens. One of the isolates, CQ19, was highly pathogenic and contained an insertion coding for two additional amino acids at the HA cleavage site. A second, CP607, was mildly pathogenic and also contained the insertion. A third, CM1374, was nonpathogenic and did not contain the insertion. Horimoto *et al.* (1995) compared sequences encoding the HA1 subunit for these three Mexican strains as well as 10 additional strains from other regions of North America (the USA and Canada), Europe, and Africa, using a maximum parsimony method of phylogenetic analysis. An important epidemiological question these authors attempted to address is whether the virulent Mexican flu strains arose locally; in that case, they would share a most recent common ancestor (MRCA) unless some of their ancestral strains were reintroduced into Canada or the USA. If the strains arose in the USA or Canada, with a subsequent introduction into Mexico, they would not share an MRCA unless none of the intervening ancestral strains from the USA and Canada were sampled. In this section, I reanalyze a subset of the sequences originally examined by Horimoto *et al.* (1995) using an LRT to examine support for the hypothesis that the strains of H5N2 isolated during the recent Mexican chicken flu epidemic did not arise from strains in the USA and Canada. Additionally, I estimate the times at which the virulent strains arose.

Sequences for 11 of the isolates of influenza A examined by Horimoto *et al.* (1995) were obtained from Genbank and aligned using ClustalW (Higgins *et al.* 1991). The isolates are as follows: chicken/Mexico/26654-1374/94 (CM1374); chicken/Puebla/8623-607/94 (CP607); chicken/Queretaro/14588-19/95 (CQ19); A/chicken/Pennsylvania/13609/93 (CP13609); chicken/Florida/25717/93 (CFLA 93); A/ruddy turnstone/Delaware/244/91 (RD244); chicken/Pennsylvania/1/83 (CP1); chicken/Pennsylvania/1370/83 (CP1370); turkey/Ontario/7732/66 (TO66); tern/South Africa/61 (TS61); chicken/Scotland/59 (CS59).

An ML tree was constructed using the program PAUP* (Swofford 1998) and applying the HKY+Γ substitution model with no constraints on the branch lengths (molecular clock not enforced). The likelihoods obtained for these sequences using several different substitution models are shown in Table 19.1. The HKY+Γ model provided a significantly better fit to the sequences than the other models considered. Parameters of the model were estimated from the data. The shape parameter of the gamma distribution describing the among-site rate variation was $\alpha = 0.501$, indicating considerable rate variation. This may arise because, for influenza A viruses, antigenic sites may experience positive selection (and consequently increased substitution rates) by comparison with nonantigenic sites (Ina and Gojobori 1994). The transition–transversion bias was $\kappa = 8.085$. The ML tree has a log-likelihood of -5739.47, and places the Mexican isolates as forming a monophyletic group and sharing a MRCA with a subclade of isolates from birds

Table 19.1 Likelihood ratio tests of the fit of several models of substitution and the molecular clock to HA sequences of avian influenza A strain H5N2. The log-likelihoods under the null and alternative models are denoted by $\ln L_0$ and $\ln L_1$, respectively, and Λ is the ratio of the likelihoods under the null versus the alternative model. HKY denotes the Hasegawa *et al.* (1985) model of nucleotide substitution with $\kappa = 1$ (model 0: no bias in rates of transition versus transversion) and $\hat{\kappa} = 3.45$ (model 1: the value of κ estimated by ML). HKY+Γ denotes the HKY model with among-site rate variation following a gamma distribution with shape parameter α (Yang 1994). Note that $\alpha \to \infty$ (model 0) implies no rate variation among sites and $\hat{\alpha}$ is the MLE of the rate variation (shape) parameter (model 1). SRFS denotes the model of Rambaut (1996) (model 0), and DR is the Felsenstein model (1981), which allows different rates among lineages (model 1). ** denotes significance at the 0.001 level.

Model of DNA substitution	$\ln L_0$	$\ln L_1$	$-2\ln\Lambda$
Test of equal transition–transversion rate			
HKY($\kappa = 1$) vs HKY($\hat{\kappa} = 3.45$)	–6079.78	–5801.36	556.84**
Test of equal rates among sites			
HKY+Γ($\alpha = \infty$) vs HKY+Γ($\hat{\alpha} = 0.576$)	–5801.36	–5739.63	123.46**
Test of molecular clock			
SRFS vs DR	–4290.37	–4278.45	23.84**

in the eastern USA. Three other strains CP1, CP1370, and TO66, from chickens and turkeys in the USA and Canada, form a separate (monophyletic) group. The tree was rooted using two sequences from South Africa (TS61) and Scotland (CS59). The topology of the ML tree for the nine North American strains is identical to that of the tree shown in Figure 19.1, although that tree is a partially constrained (SRFS) tree with branch lengths shown proportional to time.

To test the hypothesis that the Mexican strains did not arise from a recent source in the USA or Canada, an LRT was performed. Under the null (constrained) hypothesis, the Mexican isolates do not share an MRCA (other than the root ancestor) with any subset of the strains from the USA and Canada, while under the alternative hypothesis, they may have any ancestry. The best tree under the constrained (null) hypothesis has a log-likelihood of -5746.06. The test statistic is $T = -2\ln\Lambda$, where Λ is the ratio of the likelihood under the null (numerator) versus alternative (denominator) hypotheses. For the HA1 sequences examined, $T = 13.18$. As the hypotheses are not nested, a parametric bootstrap method was used to evaluate the significance of T. A total of 100 simulated data sets were generated using the PAML program (Yang 1997), with the same substitution model as was used to analyze the original data (where MLEs of the parameters of the substitution model were substituted for the true parameter values). The model tree for the simulations was the phylogeny obtained under the constrained (null) hypothesis. The original aligned sequences were of variable length with several insertions and deletions inferred from a ClustalW alignment, and a program was written to remove missing data and insertions and deletions from the simulated sequences to make them identical to the original sequences. T was calculated for each simulated

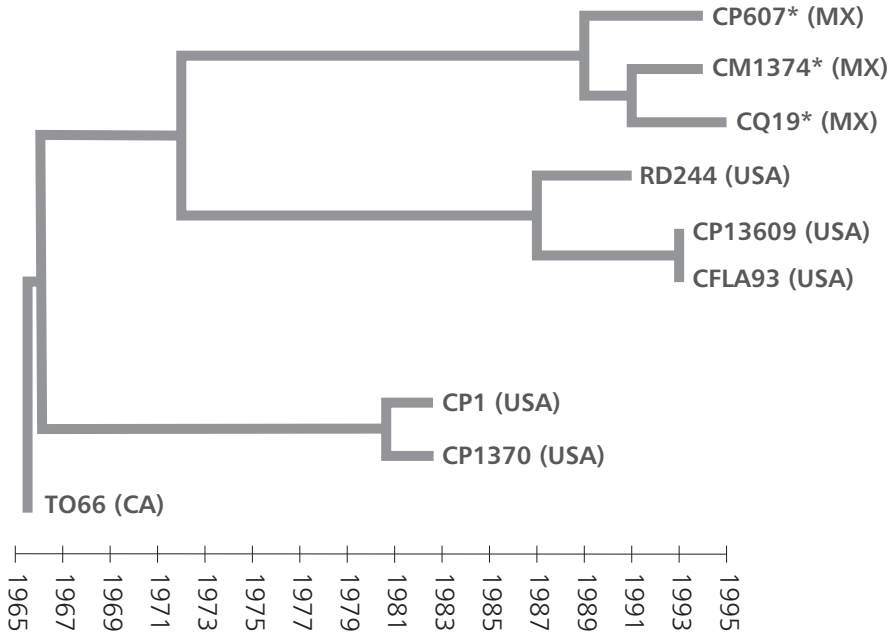


Figure 19.1 Maximum likelihood estimate of phylogeny of nine North American strains of H5N2 avian influenza. The strain abbreviations are given in the text. An asterisk indicates that a strain is pathogenic. The geographical sources for the strains are indicated by the abbreviations: MX = Mexico, USA = United States of America, CA = Canada. The divergence years were estimated using a partially constrained molecular clock (i.e., the SRFS model) and the branch lengths are calibrated in units of years.

dataset, and the proportion of times that the value of T obtained for the original dataset was exceeded by a value of T obtained for a simulated dataset was taken to be the significance of the test (i.e., a value of T at least as large as that observed for the original data would be observed under the null hypothesis with probability p, the significance). Since none of the simulated values of T exceeded the observed value, the null hypothesis can be rejected with $p \leq 0.01$. The Mexican strains do not appear to form a separate monophyletic group from the remaining North American strains. This agrees with the suggestion of Horimoto *et al.* (1995) that H5N2 influenza might have been introduced into Mexican chickens by their contact with migratory waterfowl from the USA; the RD244 strain shares an ancestor with the Mexican strains and is from a US shorebird.

One can also attempt to estimate the most recent time at which the Mexican strains of H5N2 might have been introduced from US birds. The program SPATULA (Rambaut and Grassly 1996) was used to calculate the likelihood of the tree in Figure 19.1, using only the nine North American strains and constraining the tips of the tree to be equal to the times at which the viral strains were sampled. This allowed the ages of the ancestors in the phylogenetic tree to be estimated. The likelihood under a clock hypothesis using the HKY+Γ substitution model, and allowing for the fact that the sequences have been sampled at different times using the SRFS model of Rambaut (1996), is $L_0 = -4290.37$. Relaxing the clock assumption by allowing each branch in the phylogeny to have a different

substitution rate (Felsenstein 1981, the DR model), again using the HKY+Γ substitution model, gives $L_1 = -4278.45$. There are nine degrees of freedom (df) under the (null) SRFS model ($s - 1$ internal node times, where s is the number of sequences, and one overall rate parameter must be estimated; Rambaut 1996) and 15 df under the (alternative) DR model ($2s - 3$ branch-specific rates must be estimated; Felsenstein 1981). The LRT statistic for a test of the molecular clock is $T = 23.84$. Because we assume the phylogeny is the same under both hypotheses, the models are nested and the distribution of the test statistic is approximately χ^2 with 6 df (the difference in df between the null and alternative hypotheses). The SRFS molecular clock can be rejected in this case with $p \leq 0.01$. The SRFS tree with branch lengths scaled in units of years is shown in Figure 19.1. The estimated substitution rate per year is 5.18×10^{-4}. Ignoring possible rate variation among lineages indicated by a LRT, a rough estimate of the time at which the H5N2 flu strain might have been introduced into Mexico is about 1972. The results of this analysis suggest that, although nonvirulent H5N2 influenza was probably introduced to Mexico from the US, the virulent strains of H5N2 that subsequently appeared during the Mexican chicken flu epidemic likely arose locally. The practical implication of this result is that, to reduce the threat of a major outbreak of highly virulent H5N2 influenza, poultry producers should attempt to contain local outbreaks of even mildly pathogenic strains.

1918 Spanish flu and 1997 Hong Kong flu

The so-called "Spanish" influenza pandemic of 1918 resulted in the deaths of over 20 million people, with mortality rates over 25 times higher (about 2.5%) than for a typical influenza strain (about 0.1%). The reason for this virulence is not well understood. One suggestion is that the strain was not an unusual one, but global malnourishment and urban overcrowding following World War I created an immunologically suppressed population and conditions suitable for influenza transmission. The population of the USA was largely unaffected by the war in Europe, however, and yet still suffered high mortality during the 1918 influenza pandemic. Another suggestion is that the 1918 Spanish influenza pandemic was caused by a new highly virulent strain that arose by recombination between human (or swine) and avian strains. The earliest samples of human influenza date from the 1930s, and so genetic analysis has not, until recently, been available to study the origin of the 1918 influenza. Early analyses of antibody titers from survivors of the 1918 influenza did suggest, however, that the strain was probably an H1N1 subtype (see Taubenberger *et al.* 1997). Recently, Taubenberger *et al.* (1997) isolated viral RNA from paraffin-embedded tissue samples from a patient who died of 1918 influenza. They successfully amplified and sequenced fragments of several genes including HA and NA. The strain was designated A/South Carolina/1/18 (SP18).

In 1997, a highly pathogenic strain of chicken influenza, H5N1, emerged as a source of virulent influenza infections in humans exposed to infected chickens. At least 12 confirmed cases of human infection with the strain have since been documented, six of which were fatal. Subbarao *et al.* (1998) first isolated this virus

from a 3-year-old boy, who subsequently died. The isolate, designated A/Hong Kong/156/97 (HK97), was sequenced for segments of several genes, including HA and NA, to investigate the genetic properties of the strain and, in particular, whether it arose by recombination between human (or swine) and avian strains. We analyze the HA and NA sequences for this strain, for the SP18 strain, and for several additional reference strains from humans, swine, and birds, to examine the evidence that either strain HK97 or SP18 arose by recombination between animal or human and avian strains. We also examine whether the SP18 strain shares a recent ancestry with the classic H1N1 strains isolated in the 1930s, as has been suggested.

Reference isolates that had been sequenced for both the HA and NA genes were chosen for the analysis. In the absence of recombination, the phylogeny of the isolates obtained by an analysis of each gene should agree; recombination can generate disagreements between the two gene trees. An LRT was used to quantitatively assess the evidence for recombination (different underlying gene trees) taking into account phylogenetic uncertainty (Huelsenbeck and Bull 1996). Eight strains were analyzed in total, including the HK97 and SP18 strains. The additional six strains are A/swine/Ehime/1/80 (SwEhm80), a swine influenza isolated in 1980 (H1N1); A/duck/Alberta/60/76 (DkAlb76), a North American duck influenza isolated in 1976; A/WSN/33 (WSN33), a mouse-adapted human influenza isolated in 1933 (H1N1); A/Puerto Rico/8/34 (PR34), a human influenza isolated in 1934 (H1N1); A/Yamagata/32/89 (FLA89), a Japanese swine influenza isolated in 1989 (H1N1); and A/WI/4754/94 (AWI94), a swine influenza with documented transmission to humans isolated in 1994 (H1N1).

Sequences were aligned using ClustalW (Higgins *et al.* 1991). The HA gene for SP18 was partially sequenced as three nonoverlapping fragments of variable length (Taubenberger *et al.* 1997), and these were combined to construct a composite HA sequence, with the unsequenced regions between fragments represented as missing data. An ML phylogenetic analysis was performed using PAUP* (Swofford 1998) for each gene separately and for a combined dataset of both genes. The ML tree for HA is shown in Figure 19.2.

The log-likelihood obtained in an unconstrained analysis (no molecular clock imposed; DR model) was –8315.52, with the transition–transversion bias estimated to be $\hat{\kappa} = 4.501$ and the shape parameter of the gamma distribution estimated to be $\hat{\alpha} = 0.573$.

The ML gene tree for NA is shown in Figure 19.3. The log-likelihood obtained for an unconstrained analysis was –6293.18, with the transition–transversion bias estimated to be $\hat{\kappa} = 6.208$, and the shape parameter of the gamma distribution estimated to be $\hat{\alpha} = 0.384$.

Both gene trees group together the human influenza sequences PR34, WSN33, and SP18. However, the HA gene groups the Japanese swine sequence FLA89 with the human sequences, whereas the NA gene does not. This suggests that the HA gene sequenced for FLA98 might have been introduced into this strain by recombination with a human strain. The HK97 strain diverges before the human

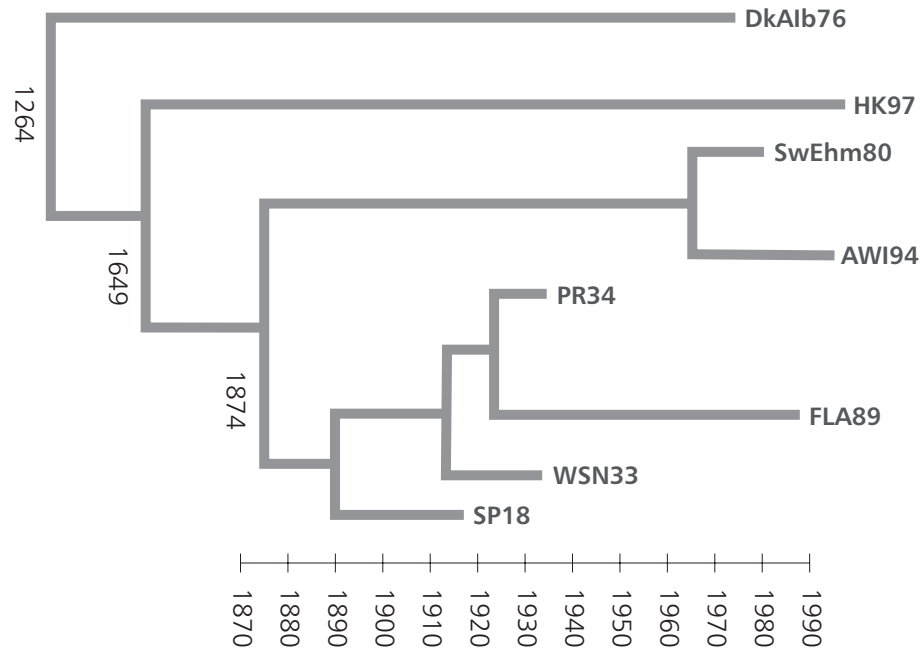


Figure 19.2 Maximum likelihood estimate of phylogeny of eight strains of influenza A isolated from humans, swine, and birds based on an analysis of the HA gene. The strain abbreviations are given in the text. The divergence years prior to 1870, estimated using a partially constrained molecular clock, are shown at the left of the branch. The branch lengths (after 1870) are calibrated in units of years (scale at bottom).

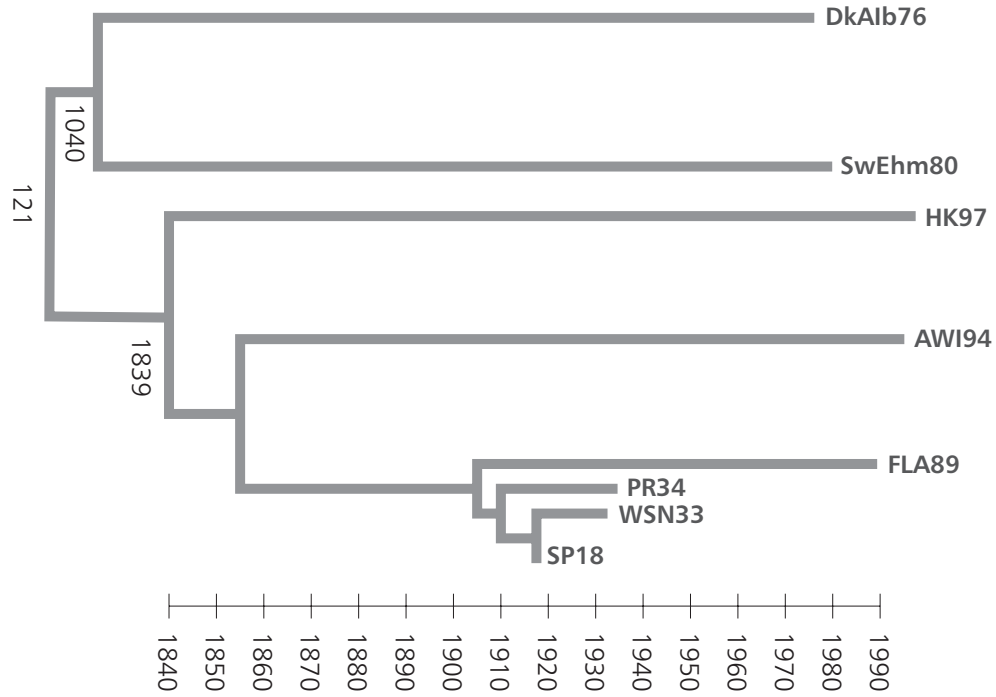


Figure 19.3 Maximum likelihood estimate of phylogeny of eight strains of influenza A isolated from humans, swine, and birds based on an analysis of the NA gene. The strain abbreviations are given in the text. The divergence years prior to 1840, estimated using a partially constrained molecular clock, are shown at the left of the branch. The branch lengths (after 1840) are calibrated in units of years (scale at bottom).

strains in both gene trees and before the swine strains as well, except in the NA gene tree, which places the SwEhm80 strain with the DkAlb76 strain; this could be either evidence for recombination of SwEhm80 with a duck strain or could be an error in the phylogeny, perhaps because of long branch attraction, as these sequences are very divergent. The HA gene tree of Figure 19.2 suggests that the human influenza strains PR34 and WSN33 share a recent ancestry with SP18, which could have arisen by recombination with an avian strain. The NA gene tree of Figure 19.3, on the other hand, suggests that the SP18 NA ancestor arose from an ancestral strain that is descended from the ancestor of PR34, and therefore is not of direct avian origin.

An LRT was used to test the hypothesis that recombination (exchange of segments) among strains, involving either the HA or the NA genes, has occurred at some point in their shared ancestry (Huelsenbeck and Bull 1996). Under the null hypothesis, the two gene trees are identical and the log-likelihood is –15281.29. Under the alternative hypothesis, each gene may have a different tree and the log-likelihood is the sum of the log-likelihoods obtained in the unconstrained analyses of the two genes, which is –14608.70. The LRT test statistic is then $T = 1345.18$, which is significant at the $p \leq 0.01$ level (based on 100 simulated datasets). The LRT, which takes into account phylogenetic uncertainty, therefore provides strong evidence for past recombination between strains.

The program SPATULA (Rambaut and Grassly 1996) was used to estimate the times at which different strains diverged under the SRFS model. For the HA gene, the log-likelihood under this model was –8330.06, with the rate of substitution estimated to be 1.48×10^{-3}. An LRT of the DR model versus the SRFS model gives $T = 56.64$, which is significant at the $p \leq 0.01$ level. The ML HA gene tree of Figure 19.2 has branch lengths scaled in units of years, and the estimated years at which different lineages diverged are indicated. This tree suggests that if SP18 arose by recombination with an avian lineage, this occurred quite recently (about 1890). The HK97 strain, on the other hand, appears to have diverged from the human and swine influenza strains roughly 200 years earlier. The ML NA gene tree of Figure 19.3 has a log-likelihood under the SRFS model of –6305.32, with the rate of substitution estimated to be 1.08×10^{-3}. An LRT of the DR versus SRFS model for this gene gives $T = 24.28$, which is significant at the $p \leq 0.01$ level. The human influenza strains appear to have diverged from the swine strains (apart from SwEhm80) in about 1910, and the HK97 strain appears to have diverged roughly 100 years earlier.

19.4 Discussion

In this chapter, several examples are given to illustrate how phylogenetic methods may be used to study the evolution of virulence. In the first example, a chicken influenza outbreak in Mexico, it is shown that a phylogenetic analysis strongly suggests that the virulent strains appearing during that epidemic originated locally; this is probably because a mildly pathogenic strain of H5N2 avian influenza was

allowed to spread unchecked through the chicken population. This result suggests that measures should be taken to contain even mildly pathogenic outbreaks of chicken influenza when they arise to prevent the eventual evolution of more virulent forms.

In the example of two highly virulent influenza strains affecting humans, the 1918 Spanish influenza SP18 and the recent Hong Kong chicken influenza HK97, capable of infecting humans who came into direct contact with infected chickens, it is shown that, although recombination appears to be involved in creating new pandemics, it is not necessarily the cause of virulence, or infectivity, in these two strains. The SP18 strain appears closely related to other less-virulent human influenza strains at both the HA and NA loci, and a novel recombination event with an avian strain does not appear to be an explanation for its virulence (Taubenberger *et al.* 1997). It appears possible that the HA gene in all the human influenza strains arose by recombination with an avian strain, but this does not explain why SP18 is so much more virulent than the others. The HK97 strain, on the other hand, appears to be a typical avian influenza with genes at both HA and NA very distantly related to those of both human and swine strains. This suggests that HK97 is unlikely to become a pandemic strain in humans without first undergoing further genetic changes. There is still a significant risk that a recombination event between HK97 and a human influenza strain in an individual who is multiply infected could produce a highly virulent pandemic strain, however, and that risk alone makes a rapid response aimed at eliminating the HK97 strain from both chickens and humans of critical importance (Subbarao *et al.* 1998).

References

References in the book in which this chapter is published are integrated in a single list, which appears on pp. 465–514. For the purpose of this reprint, references cited in the chapter have been assembled below.

Bean WJ, Kawaoka Y, Wood JM, Pearson JE & Webster RG (1985). Characterization of virulent and avirulent A/Chicken/Pennsylvania/83 influenza A viruses: Potential role of defective interfering RNAs in nature. *Journal of Virology* **54**:151–160

Bosch FX, Garten W, Klenk HD & Rott R (1981). Proteolytic cleavage of influenza virus hemagglutinins: Primary structure of the connecting peptide between HA1 and HA2 determines proteolytic cleavability and pathogenicity of avian influenza viruses. *Virology* **113**:725–735

Both GW, Sleigh MJ, Cox NJ & Kendal AP (1983). Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: Multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. *Journal of Virology* **48**:52–60

Felsenstein J (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**:368–376

Fitch WM, Bush RM, Bender CA & Cox NJ (1997). Long-term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the USA* **94**:7712–7718

Goldman N (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**:182–198

Goldman N (1998). Effects of sequence alignment procedures on estimates of phylogeny. *Bioessays* **20**:287–290

Hasegawa M, Kishino H & Yano T (1985). Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**:160–174

Higgins DG, Bleasby AJ & Fuchs R (1991). CLUSTAL V: Improved software for multiple sequence alignment. *Computer Applications in the Biosciences* **8**:189–191

Horimoto T, Rivera E, Pearson JE, Senne D, Krauss S, Kawaoka Y & Webster RG (1995). Origin and molecular changes associated with emergence of a highly pathogenic H5N2 influenza virus in Mexico. *Virology* **213**:223–230

Huelsenbeck JP & Bull JJ (1996). A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology* **45**:92–98

Huelsenbeck JP & Rannala B (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276**:227–232

Ina Y & Gojobori T (1994). Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. *Proceedings of the National Academy of Sciences of the USA* **91**:8388–8392

Jukes TH & Cantor CR (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*, ed. Munro HN, pp. 21–123. New York, NY, USA: Academic Press

Kawaoka Y, Nestorowicz A, Alexander DJ & Webster RG (1987). Molecular analysis of the hemagglutinin genes of H5 influenza viruses: Origin of a virulent turkey strain. *Virology* **158**:218–227

Kimura M (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**:111–120

Levine AJ (1992). *Viruses*. New York, NY, USA: WH Freeman and Company

Li W (1997). *Molecular Evolution*. Sunderland, MA, USA: Sinauer Associates Inc.

Li XS, Chao CY, Gao HM, Zhang YQ, Ishida M, Kanegae Y, Endo A, Nerome R, Omoe K & Nerome K (1992). Origin and evolutionary characteristics of antigenic reassortant influenza A (H1N2) viruses isolated from man in China. *Journal of General Virology* **73**:1329–1337

Mullis K (1986). Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology* **51**:263–273

Nichol ST, Spiropoulou CF, Morzunov S, Rollin PE, Kziazek TG, Feldmann H, Sanchez A, Childs J, Zaki S & Peters CJ (1993). Genetic identification of hantavirus associated with an outbreak of acute respiratory illness. *Science* **262**:914–917

Rambaut A (1996). *The Use of Temporally Sampled DNA Sequences in Phylogenetic Analysis*. PhD Thesis. Oxford, UK: Oxford University

Rambaut A & Grassly NC (1996). *SPATULA Version 1.0*. Oxford, UK: Oxford University Press

Rannala B & Yang Z (1996). Probability distribution of evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* **43**:304–311

Rohm C, Horimoto T, Kawaoka Y, Suss J & Webster RG (1995). Do hemagglutinin genes of highly pathogenic avian influenza viruses constitute unique phylogenetic lineages? *Virology* **209**:664–670

Subbarao K, Klimov A, Katz J, Regnery H, Lim W, Hall H, Perdue M, Swayne D, Bender C, Huang J, Hemphill M, Rowe T, Shaw M, Xu X, Fukuda K & Cox N (1998). Characterization of an influenza A (H5N1) virus isolated from a child with a fatal respiratory illness. *Science* **279**:393–396

Swofford DL (1998). *PAUP* 4d63*. Sunderland, MA, USA: Sinauer Associates Inc.

Swofford DL, Olsen GJ, Waddell PJ & Hillis DM (1996). Phylogenetic inference. In *Molecular Systematics*, eds. Hillis DM, Moritz C & Mable BK. Sunderland, MA, USA: Sinauer Associates Inc.

Taubenberger JK, Reid AH, Krafft AE, Bijwaard KE & Fanning TG (1997). Initial genetic characterization of the 1918 "Spanish" influenza virus. *Science* **275**:1793–1796

Voyles BA (1993). *The Biology of Viruses*. New York, NY, USA: Mosby.

Yang Z (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**:306–314

Yang Z (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**:555–556

Yasuda J, Shortridge KF, Shimizu Y & Kida H (1991). Molecular evidence for a role of domestic ducks in the introduction of avian H3 influenza viruses to pigs in southern China, where the A/Hong Kong/68 (H3N2) strain emerged. *Journal of General Virology* **72**:2007–2010