Inference of Viral Evolutionary Rates from Molecular Sequences

Alexei Drummond[1,2], Oliver G. Pybus[1] and Andrew Rambaut[1]

[1]*Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS,UK*
[2]*Department of Statistics, University of Oxford, South Parks Road, Oxford, OX1 3TG, UK*

Abstract 332
1. Introduction 332
   1.1. Ancestral Diversity and Evolutionary Non independence 334
2. General Linear Regression and Other Distance Based Methods 337
   2.1. Root to tip Linear Regression 337
   2.2. Pairwise Distance Linear Regression 338
   2.3. Generalised Least squares on a Tree 339
   2.4. Hypothesis Testing and Estimation of Errors 339
   2.5. Examples of Linear Regression Methods 341
3. Maximum Likelihood Estimation 343
   3.1. Hypothesis Testing and the Likelihood Ratio Test 344
   3.2. Rate Variation Through Time 344
   3.3. Examples of Maximum Likelihood Methods 345
   3.4. Shortcomings of Current Maximum Likelihood Implementations 346
4. Bayesian Inference of Evolutionary Rates 347
   4.1. Estimation of Errors Using MCMC 348
   4.2. Examples of MCMC Estimation Methods 349
5. Discussion 350
   5.1. Estimation of Divergence Times 351
   5.2. The Neutral Theory of Molecular Evolution and the Molecular Clock 352
   5.3. Estimating Generation Length 353
   5.4. Conclusion 354
Acknowledgements 354
References 355

ADVANCES IN PARASITOLOGY VOL 54
0065-308X $35.00

ABSTRACT

The processes of mutation and nucleotide substitution contribute to the observed variability in virulence, transmission and persistence of viral pathogens. Since most viruses evolve many times faster than their human hosts, we are in the unusual position of being able to measure these processes directly by comparing viral genes that have been isolated and sequenced at different points in time. The analysis of such data requires the use of specific statistical methods that take into account the shared ancestry of the sequences and the randomness inherent in the process of nucleotide substitution. In this paper we describe the various statistical methods for estimating evolutionary rates, which can be classified into three general approaches: linear regression, maximum likelihood, and Bayesian inference. We discuss the advantages and shortcomings of each approach and illustrate their use through the analysis of two example viruses; human immunodeficiency virus type 1 and dengue virus serotype 4. Reliable estimates of viral substitution rates have many important applications in population genetics and phylogenetics, including dating evolutionary events and divergence times, estimating demographic parameters such as population size and generation time, and investigating the effect of natural selection on molecular evolution.

1. INTRODUCTION

As a general rule, parasites have faster rates of mutation than their hosts. Parasites tend to be smaller in size, with shorter generation times, and therefore undergo more rounds of reproduction per unit time. This difference in mutation rates is particularly clear for viruses and their human hosts, not only because viral generation times are often very short, but also because replication of their genetic material is commonly many times more error-prone than in humans. That said, viruses do vary widely in mutation rate as a result of differences in their life cycles and mode of replication (e.g., Holland *et al.*, 1982; Smith and Inglis, 1987; Jenkins *et al.*, 2002).

The most significant consequence of the high mutation rate of viruses is their ability to quickly adapt to their environment, as illustrated by the rapid evolution of human immunodeficiency virus (HIV) strains that are resistant to anti-viral drugs or are capable of evading the hosts' immune response (e.g., Nijhuis *et al.*, 1997; Goulder *et al.*, 2001). In other

circumstances, mutation may allow a virus to productively infect new cell types or new host species. As the rate of mutation contributes to the adaptive potential of a virus it is obviously important to accurately measure this value. In addition, the mutation rate is a key parameter in population genetic and phylogenetic analyses of viral populations and is therefore necessary to understand both the pattern of viral genetic diversity observed today and the timescale of past evolutionary and epidemiological events.

In this article we describe the various methods by which mutation rates can be estimated from molecular sequence data, and discuss the advantages and disadvantages of each. We illustrate these methods by applying them to two human viruses that cause worldwide morbidity and mortality; human immunodeficiency virus type 1 (HIV-1) and dengue virus serotype 4 (DEN-4). Both are RNA viruses whose large genetic diversity and high mutation rates directly contribute to their virulence and pathogenicity. Our choice of data sets also illustrates the range of evolutionary timescales across which the methods we describe can be applied, as the HIV-1 data are taken from a study of viral evolution within an individual infected patient (Shankarappa *et al.*, 1999), whereas the DEN-4 data have been sampled from infected individuals across several decades in many different countries (Lanciotti *et al.*, 1997). Both these data sets are characterized by the fact that the sequences were sampled at different points in time (commonly referred to as temporally spaced, or serially sampled sequences).

Although we only consider viruses here, the methods described are equally applicable to any population from which gene sequences sampled at different points in time show a statistically significant number of genetic differences. Populations from which estimates of mutation rates can be readily obtained are characterised by some combination of the following properties: (i) a high mutation rate, (ii) long periods of time between samples, as is the case for "ancient DNA", and (iii) long stretches of sampled sequence data.

At this point we must introduce the distinction between mutation rates and substitution rates, although the two terms are sometimes confused in the literature. The former is the rate at which mutational errors are incorporated into a genome during replication, and can be expressed as the number of mutations per nucleotide site per replication event. This rate is largely determined by the particular viral or host polymerase used and the presence or absence of post-replicative repair systems. RNA viruses and small DNA viruses tend to lack such repair systems and thus have higher mutation rates. Molecular biology techniques can be used to estimate the mutation rate of viruses *in vitro* and *in vivo* (e.g., Mansky and Temin, 1995).

In contrast, the substitution rate of a virus depends on many factors and is a property of the viral population as a whole. It is the rate at which new mutations spread and become fixed in the population as a result of natural selection or random genetic drift, and is expressed as the number of substitutions per nucleotide site per unit time (days, years or generations). The substitution rate depends on the complex interaction between the effective size of the population and the distribution of mutational selection coefficients, that is, the relative proportion of mutations that are advantageous, neutral or disadvantageous. Some of these interactions can be unravelled by comparing the substitution rates separately at synonymous and non-synonymous nucleotide sites – mutations at synonymous sites do not change the encoded amino acid and can therefore be considered as having little or no selective effect (for example, Kimura, 1977). One of the most important theoretical results in molecular evolution is that if all mutations are selectively neutral then the substitution rate is equal to the mutation rate (Kimura and Ohta, 1971; Kimura, 1987). Although useful, the argument that synonymous sites are neutral should be applied carefully as it assumes the absence of several factors, some of which are common in viruses, namely (i) fitness differences arising from the use of alternative codons, (ii) secondary RNA or DNA structure in coding and non-coding regions, and (iii) overlapping reading frames. In general, if selection is acting at the nucleotide level as well as the encoded protein absolute statements about selection at the protein level can be difficult, although relative statements can often still be made.

All the procedures outlined below estimate the substitution rate, not the mutation rate. Although the substitution rate depends on several parameters and may sometimes be difficult to interpret, it is important precisely because it does contain information about many fundamental evolutionary processes. For example, if two genes in the same viral genome have unequal substitution rates then we might conclude that different selection pressures have been acting on them, since the underlying mutation rate is unlikely to differ. Furthermore, comparison of substitution rates among different virus species and strains may shed light on the varied roles that mutation and genetic diversity play in maintenance of viral infection and transmission.

1.1. Ancestral Diversity and Evolutionary Non-independence

Intuitively, one might expect to be able to estimate substitution rates using a simple argument along the lines of "distance equals rate multiplied by

time''. If two gene sequences differ at d nucleotide sites and were sampled at different points in time, t_1 and t_2, then this rationale suggests that the substitution rate μ equals $d/(t_2-t_1)$ (Figure 1a). However, this will only be true if one sequence is a direct ancestor of the other. In practice, as illustrated in Figure 1b, this method will overestimate μ when the time of most recent common ancestor (t_{root}) of the two sequences exists prior to t_1. This is known as the problem of ''ancestral diversity'' and occurs because the population at time t_1 contains some genetic variation. In most cases t_{root} will be unknown so the time over which genetic distance d has accumulated will also be unknown. Ancestral diversity can be taken into account by adding a third outgroup sequence. The difference between the genetic distances (d_1 and d_2) of the two sampled sequences to this outgroup thus reflects the difference between their sampling times (Figure 1c; Li *et al.*, 1988). This argument holds even if individual nucleotide sites have undergone multiple substitutions, provided that an accurate probabilistic model of nucleotide substitution is used to estimate d (see Swofford *et al.*, 1996). In common with many other methods this approach assumes that the substitution rate remains constant through time, an assumption known as the molecular clock.

Extending this methodology to multiple sequences appears straightforward; for example, if y_i is the genetic distance from sequence i to the most recent common ancestor of the sampled sequences (measured off a phylogenetic tree) and t_i is the sampling time of sequence i, then the

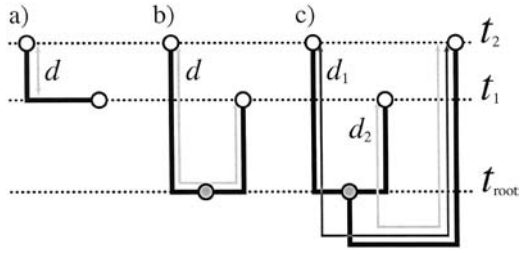

Figure 1 The problem of ancestral diversity. Gene sequences (open circles) have been sampled at two time points, t_1 (earlier) and t_2 (later). The vertical dimension represents genetic distance. (a) There is zero genetic diversity at the earlier time point (t_1) so the genetic distance d between the sequences reflects the difference in sampling times. (b) Due to genetic diversity at the earlier time point, the common ancestor of the sampled sequences (t_{root}) exists prior to t_1. Hence the genetic distance d is erroneously large. (c) The problem of ancestral diversity can be avoided by using an outgroup. The difference between the distances d_1 and d_2 correctly reflects the difference in sampling times.

gradient of a linear regression of y_i against t_i should provide an estimate of the substitution rate μ (Figure 2). We call this method the root-to-tip linear regression method and it has often been used to estimate substitution rate, but unfortunately it has serious shortcomings. It assumes that each y_i is statistically independent, whereas the sampled sequences are linked by a common evolutionary history and are thus not independent. The non-independence arises from the internal branches of the phylogeny that describes this shared history, as these branches contribute to multiple pairs of y_i and t_i values. This problem of non-independence arises in many other evolutionary problems (e.g., Harvey and Pagel, 1991) and can be solved by developing methods that explicitly incorporate the phylogenetic structure implicit in sampled sequence data. The use of models that do not incorporate this information will produce unpredictable biases in inference and hypothesis-testing procedures.

While many computational methods have been developed to estimate the phylogenetic structure of sequence data under the assumption of a molecular clock, few allow for temporally spaced sequences. Methods such as UPGMA (Sokal and Michener, 1958), likelihood ratio tests of the molecular clock (Felsenstein, 1981) and coalescent methods in population genetics (Kingman, 1982; Hudson, 1990; Kuhner *et al*., 1995) all assume that there are no significant differences between the sampling times of the individual sequences. Rather than being a potential problem, we show here that the unique structure of temporally spaced sequence data is an asset that can be exploited to a number of novel ends, including the accurate estimation of substitution rates.

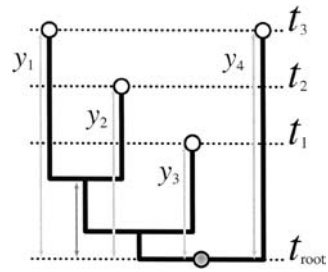

Figure 2 Root to tip distances measured on a phylogeny. Four gene sequences (open circles) have been sampled at three different time points (t_1, t_2, t_3). The y values represent the genetic distances from each sequence to the root (filled circle). Many of the y values are non independent because of shared ancestry of the sequences. For example, the dark arrow denotes the shared part of distances y_1 and y_2.

The remainder of this article is organised with each section focusing on a different approach for estimating molecular evolutionary rates. We start with the simplest methods and progress to the more sophisticated. In each section we demonstrate the relevant methods on HIV-1 and DEN-4 data sets, discussing the advantages and shortcomings of each.

2. GENERAL LINEAR REGRESSION AND OTHER DISTANCE-BASED METHODS

Some of the first estimates of substitution rates using temporally spaced sequences were obtained by comparing sequences of human influenza A strains (Krystal *et al*., 1983; Martinez *et al*., 1983; Hayashida *et al*., 1985). All of this early work involved direct comparison of the genetic distance between two sequences with the interval separating their isolation times (Figure 1a). As explained in the previous section, this method is only accurate if the genetic diversity of the population at the time of sampling is negligible, such that sequences isolated at different times differ only by substitutions accumulated during the time interval. If this condition is not met then this method has an upward bias and can provide an upper limit for the estimate of substitution rate. Consequently, the outgroup method described in Figure 1c was introduced in the context of estimating the rate of evolution of HIV-1 (Li *et al*., 1988).

From the mid 1980s to the present, a series of distance-based regression methods were employed by various researchers to remove the ''ancestral diversity'' bias generated by the substantial population polymorphism that exists in most viral populations (for example, Buonagurio *et al*., 1986; Saitou and Nei, 1986; Gojobori *et al*., 1990; Fitch *et al*., 1991; Leitner and Albert, 1999; Pagel, 1999; Shankarappa *et al*., 1999; Drummond and Rodrigo, 2000; Korber *et al*., 2000). With some exceptions, these methods were largely introduced in an informal manner, and often not linked to past research. In the next sections we will describe the three major classes of regression methods that this body of research fall into.

2.1. Root-to-tip Linear Regression

The ''root-to-tip'' linear regression method, described briefly above, has been a common choice for the estimation of substitution rate. This method proceeds by first estimating a rooted phylogeny of the sequences under analysis and then performing a linear regression between the time of

sampling of each tip and the genetic distance (sum of reconstructed branch lengths) from the root to each tip. The linear model is thus:

$$E[d_{\text{root},i}] = \mu(t_i - t_{\text{root}}) = \mu t_i - \mu t_{\text{root}}$$

where t_i is the time of tip i, μ is the unknown substitution rate and t_{root} is the unknown time of the root (in population genealogies this is equal to the time of the most recent ancestor). Under this model, the gradient of a linear regression of $d_{\text{root},i}$ against t_i provides an estimate of the substitution rate and the y-intercept is equal to $-\mu t_{\text{root}}$. By definition, the x-intercept is equal to t_{root}. Shankarappa *et al.* (1999) used this linear regression technique to study the long-term intra-host rate of HIV-1 evolution in nine infected patients, and Korber *et al.* (2000) used it to date the origin of HIV-1 group M.

2.2. Pairwise Distance Linear Regression

The second regression method was introduced by Leitner and Albert (1999) and formally described and extended upon in Drummond and Rodrigo (2000). This method relies on a result from the population genetics literature; that the expected distance between two random sequences sampled at the same time in a haploid population is equal to $\Theta = 2N_e\mu_g$, where N_e is the effective population size and μ_g is the mutation rate per site per generation. Extending this model to pairs of sequences i and j with times t_i and t_j, results in the following linear model:

$$E[d_{i,j}] = \mu|t_i - t_j| + \Theta$$

Following this model, the gradient of a linear regression of $d_{i,j}$ against $\triangle t_{ij} = |t_i - t_j|$ is an estimate of the substitution rate and the y-intercept is an estimate of the population genetic parameter Θ. Because μ_g is measured in mutations per generation, while the slope of the regression is typically in units of substitutions per year or per day, it is only possible to interpret the x-intercept of this regression as the product of $2N_e$ and generation length in the time units used. Unlike the root-to-tip method, this method does not require an estimate of the tree, and has been shown to be an unbiased estimator (Drummond and Rodrigo, 2000). However, because it does not use information about the correlation of the sequences due to shared ancestry the power of this method is significantly reduced, typically leading to very large confidence intervals. Leitner and

Albert (1999) used a parametric bootstrapping technique to demonstrate that the distribution of pairwise distances in the HIV-1 transmission history they analysed was not over-dispersed. This suggests that a simple Poisson process could not be rejected as an adequate description of molecular evolution of HIV-1 in the transmission history they studied. However, this result could also be due to the low power of this method.

2.3. Generalised Least-squares on a Tree

An interesting approach to substitution rate estimation arises from literature concerning the use of the comparative method in evolutionary biology (Harvey and Pagel, 1991; Pagel, 1999). The comparative method is a general framework for estimating the covariance of phenotypic traits among species, which correctly accounts for the non-independence arising from shared ancestry. These methods can be regarded as a class of generalised least squares (GLS) approaches. Within this framework the correlation between a continuous trait and the branch lengths of the tree itself can also be examined. If we regard time itself as a continuous trait, then we can consider the correlation of time with the branch lengths of the tree. If the branch lengths of the tree are in units of substitutions per site then we can obtain an estimate of the rate of substitution per site per unit time (see Pagel, 1999). Because this method explicitly considers the correlation structure of the tree, the non-independence of observed sequences is correctly accounted for. However, the interpretation of this method is difficult, as it assumes that time is a random variable, when clearly the substitutions themselves are the source of stochasticity, not time. Nevertheless, this method represents an interesting intermediate between the distance methods outlined above and the fully probabilistic methodologies outlined in later sections.

2.4. Hypothesis Testing and Estimation of Errors

Point estimates are meaningless without a measure of the error associated with the estimate. This is particularly important when the aim of estimation is hypothesis testing. Unfortunately we cannot make use of standard linear regression tests and statistics because these assume that the data are independent, whereas the sequences are not independent because they share evolutionary history. For example, the use of confidence intervals about a

regression line (e.g., Tanaka *et al.*, 2002) can produce an underestimate of the true error about substitution rate.

However, the linear regression procedures are amenable to general statistical techniques for error estimation such as bootstrapping (parametric and non-parametric; Efron and Tibshirani, 1993) and jackknifing (Wu, 1986). These methods involve the construction of pseudo-replicate data sets to estimate the stochastic error in the original data. In the case of bootstrapping, this can be done by randomly sampling the original data with replacement (non-parametric bootstrapping), or by simulating data using an assumed or inferred model of evolution (parametric bootstrapping). However, care must be taken because of the non-independence of the genetic distance data used. For example, Korber *et al.* (2000) used the root-to-tip regression method to estimate the evolutionary rate and age of the root (t_{root}) of HIV-1 (group M) viruses. Korber *et al.* attempted to estimate confidence intervals around their point estimate of t_{root} by re-sampling with replacement (non-parametric bootstrapping) the linear regression data points. Their analysis gave a tight confidence interval around their estimate of t_{root} and thus enabled them to reject a recent hypothesis for the origin of HIV-1. However, in this setting the bootstrap procedure they used will underestimate the true confidence intervals as it does not take into account the correlation of bootstrap replicates due to shared ancestry (i.e., it treats each root-to-tip distance as an independent piece of information about substitution rate). A statistically rigorous approach would have been to bootstrap the nucleotide sites in the sequence alignment, rather than the root-to-tip distances, as each site in an alignment represents an independent realization of the substitution process. In the next section we verify this theoretical result and show that bootstrapping the nucleotide sites rather than the regression points produces significantly larger and more realistic confidence intervals for a set of DEN-4 virus sequences.

Non-parametric bootstrapping of sequence data could also be used to estimate the confidence intervals of the GLS method. However, non-parametric bootstrapping of sequences is not sufficient for error estimation in the pairwise distance regression. The pairwise distance regression has two sources of statistical error: (i) error due to the substitution process, and (ii) error due to the coalescent process. Thus, a full parametric bootstrap must be employed to correctly assess the error associated with the pairwise distance method (Drummond and Rodrigo, 2000).

In all three cases the statistical error due to phylogenetic reconstruction is difficult to accommodate. However, these problems can be circumvented by the use of full likelihood or Bayesian methods, which we describe in Sections 3 and 4, below.

2.5. Examples of Linear Regression Methods

In this section we compare the root-to-tip and pairwise distance linear regression methods on two example datasets (DEN-4 and HIV-1) in order to illustrate their relative performance. For both methods, a matrix of pairwise genetic distances was calculated using an empirically derived F84 model of substitution (described in Swofford *et al.*, 1996). For the root-to-tip regression method, this matrix was used to estimate a neighbour-joining tree (Saitou and Nei, 1987), and the root of the tree was picked so as to maximise the R^2 value of the regression. In a real-life application, some form of model selection process should be used to choose the substitution model that best describes the data.

Figures 3 and 4 display the results of both root-to-tip and pairwise distance regression analyses on the DEN-4 and HIV-1 datasets, respectively. The root-to-tip estimate of substitution rate for the DEN-4 dataset was 8.14×10^{-4} substitutions per site per year with a confidence interval of $[4.69 \times 10^{-4}, 14.1 \times 10^{-4}]$ estimated from 5000 bootstrap replicates of the sequence data. The corresponding estimate of the date of the root is 1928 with a confidence interval of [1901, 1946]. The estimated rate compares with 9.10×10^{-4} $[0, 33.7 \times 10^{-4}]$ for the pairwise distance method. These confidence intervals are very large and include a rate of zero. The two

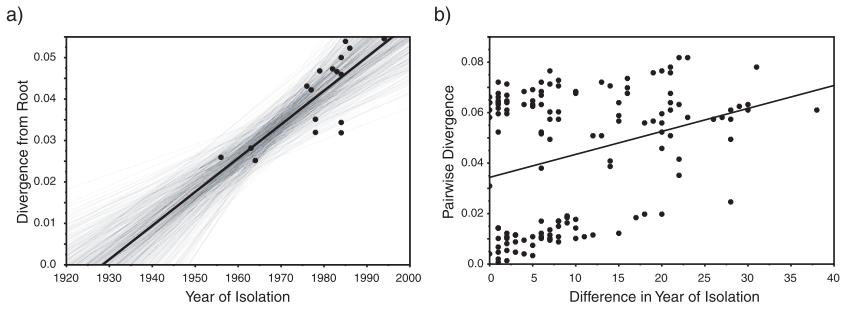

Figure 3 Application of the linear regression methods to 17 Dengue virus serotype 4 (DEN 4) sequences, isolated from different patients across five decades. (a) Root to tip linear regression. The gradient of the regression (bold line) is an estimate of the substitution rate (μ) and the x axis intercept is an estimate of the time of the most recent common ancestor (t_{root}). The pale lines represent 500 bootstrap replicates of the original sequence alignment and provide an estimate of the statistical error in these estimates. (b) Pairwise distance linear regression. The gradient of the regression (bold line) is an estimate of the substitution rate (μ) and the y axis intercept is an estimate of Θ (see text for details).

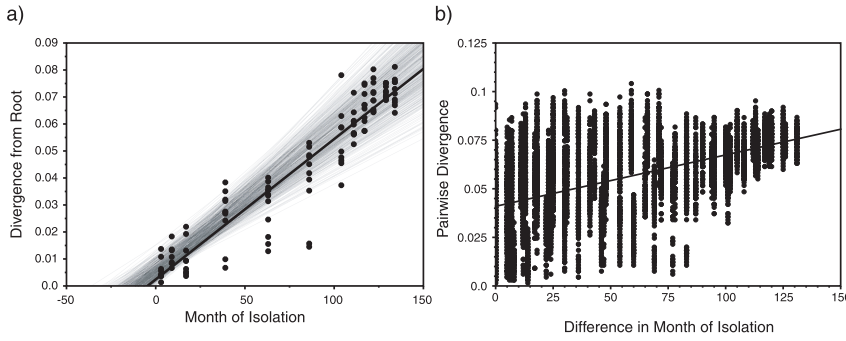

Figure 4 Application of the linear regression methods to 117 HIV 1 sequences, isolated over 134 months from a single infected patient. (a) Root to tip linear regression. (b) Pairwise distance linear regression. See Figure 3 legend for more details.

methods are congruent, and as we will see in later sections they agree well with the results of more sophisticated methods, but suffer from a lack of power, and less flexibility in model specification.

Note that if we had followed Korber *et al*. (2000) and bootstrapped the root-to-tip distances instead of the sequence data we would have calculated a confidence interval for substitution rate of only $[5.83 \times 10^{-4}, 11.1 \times 10^{-4}]$, and a confidence interval for t_{root} of [1912, 1942]. These intervals are 56% and 67% of the intervals produced by the correct bootstrapping method. Discrepancies of this magnitude can easily result in incorrect conclusions when testing specific hypotheses.

Figure 4 shows the results of the distance-based analyses of the HIV-1 dataset. The pairwise regression method gave an estimate rate of 3.17×10^{-3} $[0.26 \times 10^{-3}, 8.61 \times 10^{-3}]$ substitutions per site per year, whereas the root-to-tip regression method gave an estimate of 6.24×10^{-3} $[4.64 \times 10^{-3}, 7.89 \times 10^{-3}]$, with confidence intervals that exclude the pairwise estimate. A possible reason for this discrepancy is model misspecification in one or both of the methods. However because of the very wide confidence intervals on the pairwise method, the discrepancy between these results may not be very important.

In general, linear regression procedures are fast and useful for visualising new data sets. They can assist in model selection, by suggesting whether a uniform or variable model of substitution rate through time is necessary to explain the temporal structure in a given data set (for example, Shankarappa *et al*., 1999). However, they make several limiting assumptions and we do not recommend that they provide the final result of an analysis of

temporally spaced viral data. Newer methods, such as maximum likelihood and Bayesian statistical inference, can utilise more information from the sequences and can potentially allow much more complex models of molecular evolution and demography to be investigated. We will describe these methods in the next two sections.

3. MAXIMUM LIKELIHOOD ESTIMATION

Since the initial attempts to estimate rates using distance-based methods, a number of researchers have developed and tested ML methods that accommodate the time structure of temporally spaced sequences (Rambaut, 2000; Drummond *et al*., 2001; Seo *et al*., 2002b). Each tip of the tree has a known time (the isolation date of the sequence). The times of the internal nodes of the tree are initially unknown and are given arbitrary starting times consistent with their order in the tree. An additional parameter, the substitution rate, is then used to scale these times into units of expected number of substitutions per site. Given the tree and the expected number of substitutions per site for each branch of that tree, the likelihood of the model can be calculated (Felsenstein, 1981). The vector of internal node times, $(t_0, t_1, \ldots, t_n)$ along with the substitution rate (μ) and any parameters of the substitution model (such as the transition–transversion ratio) are then put into a standard multi-dimensional optimisation procedure to find the values that provide the maximum likelihood. This model has been labelled as the 'single rate dated tips' (SRDT) model (Figure 5; Rambaut, 2000).

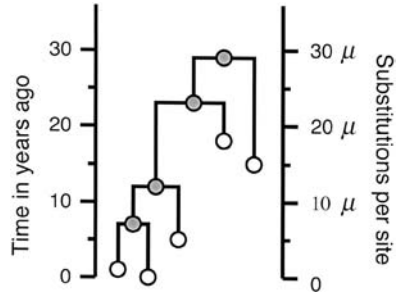

Figure 5 Maximum likelihood estimation of substitution rate using the single rate dated tips (SRDT) model. Ancestral divergence times (filled circles) are unknown and free to vary. The isolation times of the sampled sequences (open circles) are known and fixed. The substitution rate (μ) is used to convert the isolation times into genetic distances (measured in units of substitutions per site). The ancestral divergence times and μ are then found by maximum likelihood.

These methods are more sensitive and accurate than distance-based methods, as can be seen by the reduced confidence intervals reported in Section 3.3.

3.1. Hypothesis Testing and the Likelihood Ratio Test

One of the strengths of the ML inference framework is that it provides powerful tools for hypothesis testing and model comparison through tests such as the likelihood ratio test (LRT). This test uses the difference in log likelihood between two hypotheses to assess whether one provides a significantly better fit to the data than the other. The LRT requires that the hypotheses are nested, that is, one or more parameters of the more general hypothesis are constrained to particular values in order to obtain the more specific hypothesis. For example, the substitution rate parameter could be constrained to a particular *a priori* value (perhaps a previously inferred value estimated from different data). The likelihood ratio would then be used to test whether this substitution rate was a significantly worse fit to the data than the maximum likelihood estimate of rate (the more general hypothesis). For such cases, an approximate null distribution of the likelihood ratio statistic has been described (Wilks, 1938) or it can be generated using simulation (e.g., Huelsenbeck and Rannala, 1997). One of the first descriptions of such a test in phylogenetics was the test of the molecular clock by Felsenstein (1981) but this test assumes that all sequences were sampled contemporaneously. The application of the likelihood ratio test to temporally sampled data is described in more detail by Rambaut (2000) and Drummond *et al*. (2001).

3.2. Rate Variation Through Time

There are a number of reasons why evolutionary rates may vary through time across an entire viral population. For example it has been suggested that HIV-1 viruses in a host exhibit a slowdown in substitution rate at the end of the asymptomatic period (Shankarappa *et al*., 1999). These patterns of concerted population-wide changes in rate through time have also been observed due to external changes in environment such as application of anti-retroviral drugs (Drummond *et al*., 2001). It is relatively straightforward to design an LRT that will test the hypothesis of concerted rate variation through time, and this has already been described for the case of stepwise changes in substitution rate (Drummond *et al*., 2001). Models of this variety may be described as multiple rates dated tips (MRDT) models.

Further advances in this direction will assist in the rigorous and detailed dissection of the molecular evolutionary process and its variation through time.

3.3. Examples of Maximum Likelihood Methods

Under a maximum likelihood framework we have greater flexibility in model selection. Using the program TipDate (Rambaut, 2000) we estimated the substitution rate of the DEN-4 dataset using the HKY model of substitution with a different rate at each codon position. The input tree topology was estimated under the different rates (DR) model and the same substitution model in PAUP* (Swofford, 1998). The maximum likelihood estimate of the mean substitution rate was 7.91×10^{-4} [6.07×10^{-4}, 9.86×10^{-4}] substitutions per site per year and the estimated age of the root is 1922 [1900, 1936]. Figure 6 shows the maximum likelihood trees for DEN-4 under the SRDT model of evolution.

Using the program PAML (Yang, 1997), which also allows estimation of rate under the SRDT model but is better able to handle large data sets, we

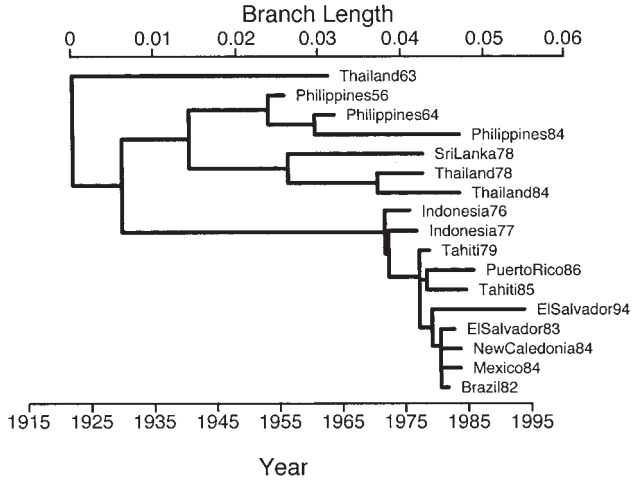

Figure 6 Application of the maximum likelihood approach to the DEN 4 data set. The phylogeny shown was estimated under the SRDT model, so that each sequence is positioned correctly with respect to its sampling date. The tips are labelled with the year of sampling and the top scale gives the genetic distance from the root. At the bottom is the timescale in years estimated using maximum likelihood.

estimated the rate of evolution of the HIV-1 sequences. Again, the input tree topology was estimated under the different rates (DR) model and under the HKY model of substitution in the program PAUP*. The maximum likelihood estimate of the mean substitution rate was 4.24×10^{-3} $[3.26 \times 10^{-3},\ 5.36 \times 10^{-3}]$ substitutions per site per year. The associated confidence intervals do not contain either the pairwise or the root-to-tip regression point estimates. This inconsistency in error estimates between different methods arises from the implicit assumptions of each method, such as the assumption of perfect knowledge of the tree topology, which we discuss in the next section.

3.4. Shortcomings of Current Maximum Likelihood Implementations

One limitation of current ML implementations, such as PAML and TipDate, is that only a single tree is considered. This can be a problem for two reasons: Firstly, there is usually considerable uncertainty in our estimation of the true tree, so that it becomes important to reflect this uncertainty in the confidence interval associated with the estimated evolutionary rate. Secondly, the maximum likelihood tree topology under the single rate dated tips (SRDT) model can be different from the maximum likelihood tree topology under the DR model (Drummond *et al.*, 2001), so using an ML topology from PAUP* can bias substitution rate estimation using TipDate. As with the root-to-tip regression analysis, the use of a single tree introduces the potential for bias.

One could attempt to simultaneously find the maximum likelihood tree and the evolutionary rate using heuristic optimisation. This would involve progressively making changes to the tree accepting those that improve the likelihood (the hill-climbing approach). Such techniques are used to estimate the maximum likelihood tree in phylogenetics packages such as PAUP* and PHYLIP, although not for the case of temporally sampled sequences. Whilst such methods are feasible for small numbers of sequences, the number of possible trees increases explosively as the number of sequences increases (Schröder, 1870). This, in addition to the complex nature of the constraints of the SRDT model, would make hill-climbing extremely susceptible to producing sub-optimal solutions. On the other hand, if we were not directly interested in the ancestral tree itself, it would be preferable to have a method that took into account the shared ancestry of the data without basing inference on a single estimation of ancestral genealogy. Markov chain Monte Carlo (MCMC) methods provide exactly this opportunity.

4. BAYESIAN INFERENCE OF EVOLUTIONARY RATES

Markov chain Monte Carlo (MCMC) integration is often used in statistical inference to summarise high-dimensional probability densities where analytical solutions are difficult or impossible to calculate. MCMC works by sampling the probability density function of interest, so as to provide a representative sample of parameter values of the chosen model, given the data. To estimate substitution rates, the chosen model generally includes the tree topology, the times of ancestral nodes in the tree, the substitution rate, and substitution parameters such as the transition/transversion ratio (Drummond *et al*., 2002).

In phylogenetics and population genetics we often want to estimate parameters, such as substitution rate, despite not knowing the true ancestral history of the sequences. A good solution to this problem would be to estimate the substitution rate from each of a large set of different ancestral histories and then combine these individual estimates such that each rate is weighted proportional to the likelihood of the corresponding tree. Trees that make the data highly probable contribute most to the overall estimate. By making the tree a nuisance parameter of the model, it becomes possible to sample all plausible trees in an MCMC analysis in order to find the range of plausible substitution rates. Unlike ML, which typically employs some kind of hill-climbing procedure, MCMC is a stochastic algorithm and is thus able to avoid getting stuck in local sub-optimal solutions because it samples the whole distribution of interest. At each step in the algorithm, MCMC proceeds by proposing a new set of parameter values (of which the tree topology is one) and then either accepting or rejecting the newly proposed state based on the Metropolitan-Hastings criterion (Metropolis *et al*., 1953; Hastings, 1970). In essence, if the proposed state is better than the previous state, it is accepted. However, if the proposed state is, say, 10 times worse than the current state, it is accepted with a probability of $p = 1/10 = 0.1$. If the proposed state is rejected then the MCMC retains the current state and the process is repeated. Using this acceptance criterion, the proportion of times the MCMC algorithm visits a particular tree is an estimate of its relative probability given the data.

Sample-based inference using MCMC readily lends itself to Bayesian inference, in which prior information can be incorporated into the analysis. Probability theory tells us that *Posterior Probability* $\propto$ *Likelihood* $\times$ *Prior Probability*. One natural approach to assigning prior probabilities to genealogies that represent large populations is the coalescent process (Kingman, 1982). In addition, parameters such as t_{root} and effective population size (N_e) can be given prior distributions that reflect information

from independent sources, or are simply used in an exploratory manner to investigate different *a priori* assumptions and hypotheses. For example, in the case of a set of viruses sampled from a single infected host, a potential prior distribution on t_{root} is the age of the host. This prior represents the assumption that the initial infection was from a single viral particle or a small homogeneous population with no double-infection.

The historical population processes that shape the genetic diversity of a population can be illuminated by genealogical methods such as the coalescent (Kingman, 1982). The coalescent is the most appropriate framework for studying the evolutionary genetics of a large population from which a sample of sequences is drawn, and provides a number of opportunities for inference in viral populations (e.g., Pybus *et al*., 2000, 2001). A description of the coalescent for serially sampled sequences has recently been given (Rodrigo and Felsenstein, 1999). This formulation of the coalescent has been used to develop methods that estimate population sizes and substitution rates from serially sampled sequences whilst taking into account the uncertainty of the tree topology using MCMC (Drummond *et al*., 2002). Others have used the coalescent to describe a pseudo-maximum likelihood method of estimating population size or substitution rates when the tree is known (Seo *et al*., 2002a). However, it should be noted that currently implemented coalescent methods make a number of limiting assumptions, specifically, no population subdivision, no recombination within the genome region under investigation, and no selection.

Theoretical developments in the future will enable these assumptions to be relaxed, but for the time being our understanding of the molecular biology and life cycle of the virus concerned should be used to carefully interpret the results of each analysis. For example, strong natural selection acts on many viruses, but often acts unequally at different levels: within an infected individual, HIV is constantly adapting in response to the host's cellular and humoral immune responses and selection is obviously strong. However, successful transmission to a new host almost always leads to the establishment of a persistent infection, so the number of "offspring" infections generated by one infected host is primarily determined by that host's behaviour, rather than by particular mutations in the viral genome. Thus the reproductive success of an HIV infection at the epidemiological level has a low heritability, and consequently selection at this level will be weak and slow-acting.

4.1. Estimation of Errors Using MCMC

Highest posterior density (HPD) intervals and central posterior density (CPD) intervals are Bayesian analogues of the confidence interval. HPD and

CPD intervals of a parameter of interest, such as substitution rate, can be obtained empirically from the frequency distribution of the parameter's values sampled by the MCMC algorithm. This is valid because after the MCMC algorithm has had an appropriate time to converge (referred to as the burn-in period) it will begin to sample values of a parameter at a frequency proportional to their (posterior) probability density. The resulting frequency distribution of a parameter of interest is thus an empirical estimate of the marginal posterior probability density of the parameter. These marginal densities can be used to reject specific *a priori* hypotheses; for example, the substitution rates of two genes are the same.

4.2. Examples of MCMC Estimation Methods

MCMC was used to estimate the substitution rate of DEN-4 without assuming exact knowledge of the tree topology. In addition, a coalescent prior on node heights was introduced to investigate its effect on rate estimation. Figure 7 shows the marginal probability distributions for each of the three codon positions as well as the mean substitution rate across all nucleotide positions. The posterior estimate of mean substitution rate in DEN-4 was 8.29×10^{-4} [6.33×10^{-4}, 10.4×10^{-4}] substitutions per site per year. Notice that this HPD interval is slightly larger than the ML confidence interval (Figure 9). This reflects the increased uncertainty in the rate due to the uncertainty in the exact genealogical relationships of the sequences. By assuming a single tree topology the ML analysis gave artificially tight confidence intervals. Interestingly, the confidence intervals for the age of the root are smaller in the MCMC analysis. This probably reflects the effect of the coalescent prior on the tree topology, as assumptions about population processes will tend to reduce the variance in estimates of node times.

Figure 8 shows the resulting probability densities of substitution rate for two independent MCMC analyses of HIV-1 dataset. The only difference between the models used was that the first analysis assumed a constant population size whereas the second allowed an exponentially expanding population (with the growth rate included as a parameter of the model). The two distributions are remarkably similar demonstrating a robustness of the estimate of rate to the exact choice of prior on the distribution of internal node ages. The posterior estimates and HPD intervals of substitution rate for the constant size and exponentially expanding population models were 6.19×10^{-3} [5.32×10^{-3}, 7.07×10^{-3}] and 6.11×10^{-3} [5.33×10^{-3}, 6.88×10^{-3}] substitutions per site per year, respectively. With these examples, we have tried to demonstrate that properties such as (i) low variance, (ii) flexibility of modelling and (iii) accurate assessment of

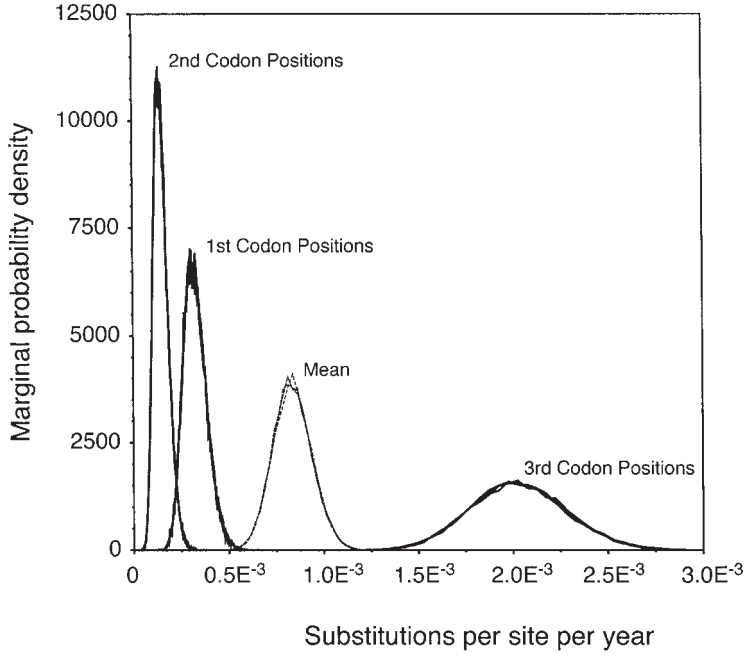

Figure 7 Application of the Bayesian inference approach to the DEN 4 data set. The figure shows the estimated posterior distributions of substitution rate for each codon position. In addition, the estimated posterior distribution of the mean rate is shown. Interestingly, the mean rate distribution does not overlap with any of the codon position distributions. The figure overlays results from four separate runs of the MCMC algorithm on the same data each with different random starting topologies. The similarity of the four distributions indicates that the algorithm has converged to the correct posterior distribution and thus has sampled the parameter space of the model adequately.

statistical errors, make MCMC an attractive and practical option for the estimation of evolutionary parameters such as substitution rate.

5. DISCUSSION

Temporally spaced data from rapidly evolving viruses provide an opportunity to ask questions about population dynamics and molecular evolution that are not possible with slow-evolving organisms or contemporaneous sequence data. Although we have concentrated on the estimation of molecular evolutionary rates there are a number of closely

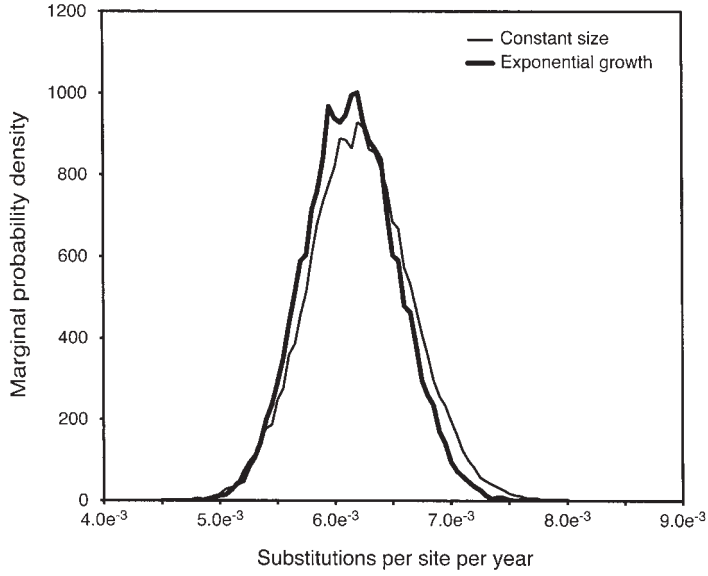

Figure 8 Application of the Bayesian inference approach to the HIV 1 data set. The figure shows results from two separate runs of the MCMC algorithm, the first assuming a constant population size for the coalescent model (thin line) and the second allowing an exponentially expanding population (bold line).

related problems that can be tackled using the methods above. We outline a few of them below.

5.1. Estimation of Divergence Times

Temporally spaced sequence data allows for the independent estimation of divergence times in viral phylogenies and genealogies. Traditionally, in the wider field of phylogenetic inference, independent calibration information has been used to determine the divergence time of an anchor node and then, assuming a molecular clock, used to estimate the ages of other divergences in the tree (for example, Shields and Wilson, 1987). However internal-node calibration methods suffer difficulties when there are few calibration points and when the substitution rates over long timescales are used to calibrate divergences over short timescales. It is also generally unlikely that internal-node calibrations will be available for viral sequence data, although a few examples do exist (e.g., Leitner and Albert, 1999; Pybus *et al.*, 2001; Van Dooren *et al.*, 2001).

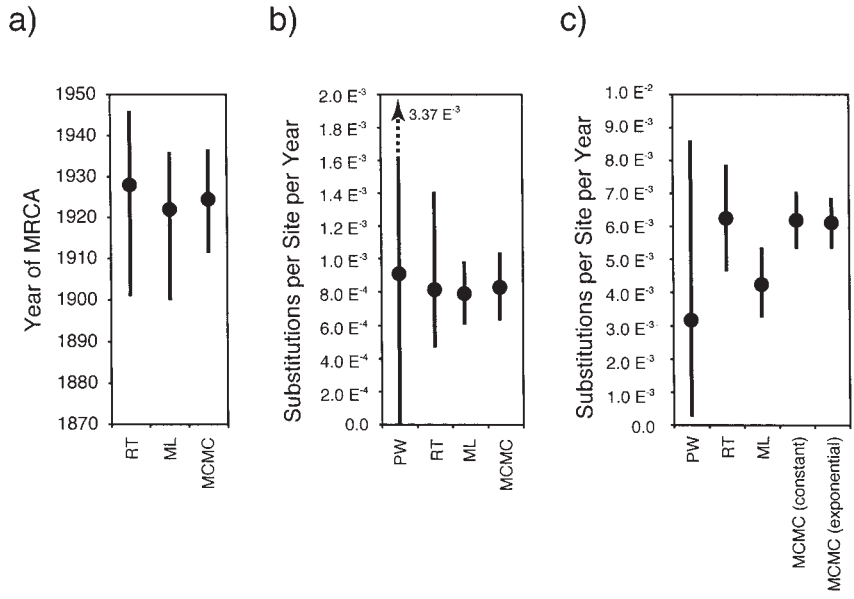

Figure 9 A comparison of the parameter estimates produced by the different methods discussed in this paper; root to tip linear regression (RT), pairwise linear regression (PW), maximum likelihood (ML) and Bayesian Markov chain Monte Carlo inference (MCMC). (a) Comparison of the results obtained for the time of the most recent common ancestor for the DEN 4 data set. (b) Comparison of the results obtained for the rate of substitution for the DEN 4 data set. (c) Comparison of the results obtained for the rate of substitution for the HIV 1 data set. For this data set, the MCMC estimates under both the constant population size and the exponential growth models are shown.

5.2. The Neutral Theory of Molecular Evolution and the Molecular Clock

The clock-like nature of many rapidly evolving viruses has been used to support both the molecular clock hypothesis (e.g., Leitner and Albert, 1999) and Kimura's neutral theory of evolution (e.g., Gojobori *et al.*, 1990). Although there is now fairly strong evidence of positive selection in HIV-1 (Nielsen and Yang, 1998), it still appears to be a relatively minor contribution to the molecular evolution of the HIV-1 genome as a whole. In fact, recent preliminary evidence of a negative correlation between population size and mutation rate suggests that negative selection imposed by functional constraints is more important and ubiquitous in HIV-1

evolution than positive selection (Seo *et al*., 2002a). This observation can be explained by either the nearly neutral theory or the slightly deleterious model of molecular evolution (Ohta and Kimura, 1971; Ohta, 1987; Tachida, 1991; discussed in Gillespie, 1995). The difficulty is that the strict molecular clock hypothesis does not appear to survive careful scrutiny. For example, only 7 out of 50 RNA viruses fit a strict molecular clock when tested in one recent comprehensive study (Jenkins *et al*., 2002). However, the authors of that study went on to show by simulation that even for the viruses that did not obey a strict molecular clock, the substitution rates estimated could still be regarded as an accurate reflection of the average substitution rate. A more satisfactory solution to this problem is the recent development of 'relaxed clock' models of substitution (Thorne *et al*., 1998; Huelsenbeck *et al*., 2000). These models allow molecular evolutionary rates to vary over time and across lineages. In the future, incorporation of these methods into the analysis of temporally spaced sequence data should allow both estimation of average evolutionary rate and the extent of rate variation across lineages. This has already begun, with a recent description of an MCMC method (though without considering phylogenetic uncertainty) that allows for dated-tips and lineage-specific rate variation (Thorne and Kishino, 2002). The chief concern in the further development of tests of the molecular clock will be in assessing the relative merits of rate-per-lineage models and MRDT models in uncovering the trends in the variation of evolutionary rate.

5.3. Estimating Generation Length

If the ages of sequences are known in calendar units (for example, days or years) then it is possible to estimate the substitution rate per site per calendar unit. However, population genetic theory tells us that in a haploid population the expected genetic diversity, Θ, is two times the product of population size and mutation rate per *generation*. Hence in order to estimate population size we need to know the conversion factor τ, the number of calendar units per generation (i.e. the generation length). This problem can be turned on its head if the mutation rate is already known from some external source. In this case, one can estimate the generation length from serially sampled genetic data, given the mutation rate. A number of methods have been described to do this for HIV-1 (Rodrigo *et al*., 1999; Fu, 2001; Seo *et al*., 2002a) and all agree closely with methods based on viral load dynamics. This congruence between genetic methods and viral load dynamics is encouraging because it occurs despite completely different sources of data. The most recent of these methods, a pseudo-likelihood

method (Seo *et al.*, 2002a), was used to estimate the generation length of nine intra-patient data sets. Assuming a single underlying mutation rate, they estimated that generation length in HIV-1 varied from 0.73 to 2.43 days among the nine patients, again showing close congruence with early work.

5.4. Conclusion

Recent maximum likelihood and Bayesian methods of analysis have filled an important gap in the study of viral evolution. These methods both provide a wealth of options for hypothesis testing and model comparison by providing a solid statistical basis for genealogy-based inference of molecular rates, based on coalescent theory and likelihood models of molecular evolution. However, as mentioned above, the methods described here are still limited by a number of simplifying assumptions. Substantial population subdivision, recombination or selection may adversely affect analysis of temporally spaced viral sequences. Most of the methods described here assume single panmictic populations, free of recombination and selection. Therefore, extensions of the Bayesian inference framework described here to take into account migration between subpopulations, substantial recombination and selection effects are needed. Most of these processes fall squarely within the purview of population genetics and are already understood in the context of contemporaneous samples of sequences. We expect that in the near future methods that allow incorporation of all of these effects will exist for analysis of rapidly evolving viruses. In fact, very early on it was predicted that temporally spaced data would provide the opportunity to shed new light on these forces:

> "To sum up, selective trends will be detectable only if data from the past are available." (Cavalli Sforza and Edwards, 1967)

The use of the methods outlined in this article, and their derivatives, will assist in answering fundamental questions about the tempo and mode of viral molecular evolution.

Software packages for performing some of these analyses and links to other resources are available from http://evolve.zoo.ox.ac.uk/VirusRates/.

ACKNOWLEDGEMENTS

This work was funded by EPSRC and MRC (AJD), The Wellcome Trust (OGP) and The Royal Society (AER). Thanks to Ziheng Yang and an anonymous referee for helpful comments.

REFERENCES

Buonagurio, D.A., Nakada, S., Parvin, J.D., Krystal, M., Palese, P. and Fitch, W.M. (1986). Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. *Science* **232**, 980 982.

Cavalli Sforza, L.L. and Edwards, A.W.F. (1967). Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* **19**, 233 257.

Drummond, A. and Rodrigo, A.G. (2000). Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial sample UPGMA. *Molecular Biology and Evolution* **17**, 1807 1815.

Drummond, A., Forsberg, R. and Rodrigo, A.G. (2001). The inference of stepwise changes in substitution rates using serial sequence samples. *Molecular Biology and Evolution* **18**, 1365 1371.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G. and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307 1320.

Efron, B. and Tibshirani, R. (1993). An introduction to the bootstrap. London: Chapman and Hall.

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

Fitch, W.M., Leiter, J.M., Li, X.Q. and Palese, P. (1991). Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 4270 4274.

Fu, Y.X. (2001). Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Molecular Biology and Evolution* **18**, 620 626.

Gillespie, J.H. (1995). On Ohta's Hypothesis: most amino acid subsitutions are deleterious. *Journal of Molecular Evolution* **40**, 64 69.

Gojobori, T., Moriyama, E.N. and Kimura, M. (1990). Molecular clock of viral evolution, and the neutral theory. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 10015 10018.

Goulder, P.J.R., Brander, C., Tang, Y.H., Tremblay, C., Colbert, R.A., Addo, M.M., Rosenberg, E.S., Nguyen, T., Allen, R., Trocha, A., Altfeld, M., He, S.Q., Bunce, M., Funkhouser, R., Pelton, S.I., Burchett, S.K., McIntosh, K., Korber, B.T.M. and Walker, B.D. (2001). Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature* **412**, 334 338.

Harvey, P.H. and Pagel, M.D. (1991). The comparative method in evolutionary biology. In: *Oxford Studies in Ecology and Evolution* (R.M. May and P.H. Harvey, eds), Oxford: Oxford University Press.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97 109.

Hayashida, H., Toh, H., Kikuno, R. and Miyata, T. (1985). Evolution of influenza virus genes. *Molecular Biology and Evolution* **2**, 289 303.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. and Vandepol, S. (1982). Rapid evolution of RNA genomes. *Science* **215**, 1577 1585.

Hudson, R.R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**, 1 44.

Huelsenbeck, J.P. and Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227 232.

Huelsenbeck, J.P., Larget, B. and Swofford, D. (2000). A compound poisson process for relaxing the molecular clock. *Genetics* **154**, 1879 1892.

Jenkins, G.M., Rambaut, A., Pybus, O.G. and Holmes, E.C. (2002). Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of Molecular Evolution* **54**, 156 165.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275 276.

Kimura, M. (1987). Molecular evolutionary clock and the neutral theory. *Journal of Molecular Biology* **26**, 24 33.

Kimura, M. and Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467 469.

Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and their Applications* **13**, 235 248.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S. and Bhattacharya, T. (2000). Timing the ancestor of the HIV 1 pandemic strains. *Science* **288**, 1789 1796.

Krystal, M., Buonagurio, D., Young, J.F. and Palese, P. (1983). Sequential mutations in the NS genes of influenza virus field strains. *Journal of Virology* **45**, 547 554.

Kuhner, M.K., Yamato, J. and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis Hastings sampling. **140**, 1421 1430.

Lanciotti, R.S., Gubler, D.J. and Trent, D.W. (1997). Molecular evolution and phylogeny of dengue 4 viruses. *Journal of General Virology* **78**, 2279 2286.

Leitner, T. and Albert, J. (1999). The molecular clock of HIV 1 unveiled through analysis of a known transmission history. *Proceedings of the National Academy of Sciences, USA* **96**, 10752 10757.

Li, W. H., Tanimura, M. and Sharp, P.M. (1988). Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution* **5**, 313 330.

Mansky, L.M. and Temin, H.M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* **69**, 5087 5094.

Martinez, C., del Rio, L., Portela, A., Domingo, E. and Ortin, J. (1983). Evolution of the influenza virus neuraminidase gene during drift of the N2 subtype. *Virology* **130**, 539 545.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087 1091.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV 1 envelope gene. *Genetics* **148**, 929 936.

Nijhuis, M., Schuurman, R., de Jong, D., van Leeuwen, R., Lange, J., Danner, S., Keulen, W., de Groot, T. and Boucher, C.A.B. (1997). Lamivudine resistant human immunodeficiency virus type 1 variants (184V) require multiple amino acid changes to become co resistant to zidovudine in vivo. *Journal of Infectious Diseases* **176**, 398 405.

Ohta, T. (1987). Very slightly deleterious mutations and the molecular clock. *Journal of Molecular Evolution* **26**, 1 6.

Ohta, T. and Kimura, M. (1971). On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution* **1**, 18 25.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401**, 877 884.

Pybus, O.G., Rambaut, A. and Harvey, P.H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429 1437.

Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C. and Harvey, P.H. (2001). The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323 2325.

Rambaut, A. (2000). Estimating the rate of molecular evolution: Incorporating non contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395 399.

Rodrigo, A.G. and Felsenstein, J. (1999). Coalescent approaches to HIV population genetics. *In*: Molecular evolution of HIV (K. Crandall, ed), Baltimore, MD: Johns Hopkins University Press.

Rodrigo, A.G., Shpaer, E.G., Delwart, E.L., Iversen, A.K., Gallo, M.V., Brojatsch, J., Hirsch, M.S., Walker, B.D. and Mullins, J.I. (1999). Coalescent estimates of HIV 1 generation time in vivo. *Proceedings of the National Academy of Sciences of USA* **96**, 2187 2191.

Saitou, N. and Nei, M. (1986). Polymorphism and evolution of influenza A virus genes. *Molecular Biology and Evolution* **3**, 57 74.

Saitou, N. and Nei, M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 159 166.

Schröder, E. (1870). Vier Combinatorische Probleme. *Zeitschriften für Mathematik und Physik* **15**, 361 376.

Seo, T.K., Thorne, J.L., Hasegawa, M. and Kishino, H. (2002a). Estimation of effective population size of HIV 1 within a host. A pseudomaximum likelihood approach. *Genetics* **160**, 1283 1293.

Seo, T.K., Thorne, J.L., Hasegawa, M. and Kishino, H. (2002b). A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics* **18**, 115 123.

Shankarappa, R., Margolick, J.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X. L. and Mullins, J.I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology* **73**, 10489 10502.

Shields, G.F. and Wilson, A.C. (1987). Calibration of mitochondrial DNA evolution in geese. *Journal of Molecular Evolution* **24**, 212 217.

Smith, D.B. and Inglis, S.C. (1987). The mutation rate and variability of eukaryotic viruses an analytical review. *Journal of General Virology* **68**, 2729 2740.

Sokal, R.R. and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **38**, 1409 1438.

Swofford, D.L. (1998). PAUP 4.0: Phylogenetic analysis using parsimony (and other methods). 4.0b10 ed. Sunderland, MA: Sinauer Associates, Inc.

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996). Phylogenetic inference. 2nd ed. *In*: Molecular Systematics (D.M. Hillis, C. Moritz and B.K. Mable, eds), pp. 407 514. Sunderland: Sinauer Associates, Inc.

Tachida, H. (1991). A study on a nearly neutral mutation model in finite populations. *Genetics* **128**, 183 192.

Tanaka, Y., Hanada, K., Mizokami, M., Yeo, A.E.T., Shih, J.W. K., Gojobori, T. and Alter, H.J. (2002). A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence

in the United States will increase over the next two decades. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 15584 15589.
Thorne, J.L. and Kishino, H. (2002). Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**, 689 702.
Thorne, J.L., Kishino, H. and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**, 1647 1657.
Van Dooren, S., Salemi, M. and Vandamme, A. M. (2001). Dating the origin of the African human T cell lymphotropic virus Type I (HTLV I) subtypes. *Molecular Biology and Evolution* **18**, 661 671.
Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**, 60 62.
Wu, C.F.J. (1986). Jacknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* **14**, 1261 1295.
Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in Biosciences* **13**, 555 556.