LA-UR- 08-6992

Approved for public release; distribution is unlimited.

Title:	Phylogenetic Trees in Bioinformatics
Author(s):	T. Burr
Intended for:	Journal: Current Bioinformatics

Form 836 (7/06)

Phylogenetic Trees in Bioinformatics

Tom Burr
Statistics Group, Los Alamos National Laboratory
Mail Stop F600, Los Alamos National Laboratory, Los Alamos NM 87545
Ph: 505.665.7865, fax: 505.667.4470, email: tburr@lanl.gov

Abstract

Genetic data is often used to infer evolutionary relationships among a collection of viruses, bacteria, animal or plant species, or other operational taxonomic units (OTU). A phylogenetic tree depicts such relationships and provides a visual representation of the estimated branching order of the OTUs. Tree estimation is unique for several reasons, including: the types of data used to represent each OTU; the use of probabilistic nucleotide substitution models; the inference goals involving both tree topology and branch length, and the huge number of possible trees for a given sample of a very modest number of OTUs, which implies that finding the best tree(s) to describe the genetic data for each OTU is computationally demanding.

Bioinformatics is too large a field to review here. We focus on that aspect of bioinformatics that includes study of similarities in genetic data from multiple OTUs. Although research questions are diverse, a common underlying challenge is to estimate the evolutionary history of the OTUs. Therefore, this paper reviews the role of phylogenetic tree estimation in bioinformatics, available methods and software, and identifies areas for additional research and development.

INTRODUCTION

Genetic data is often used to infer evolutionary relationships among a collection of viruses, bacteria, animal or species, or other operational taxonomic units (OTU). A phylogenetic tree (Figure 1) is a visual representation of either the true or the estimated branching order of the OTUs, depending on the context. Because the taxa often cluster in agreement with auxiliary information, such as geographic or temporal isolation, one goal associated with tree estimation is to infer the number of groups (also known as clusters, clades, or subtypes) and group memberships. For example, one application for trees is to identify viral subtype and new recombinant subtypes arising from combinations of known subtypes, which both have important implications for drug and vaccine design [1]. Another application for trees is to use tree shape [1,2] to infer aspects of population history such as population growth rates or subdivision.

Bioinformatics is too large a field to review here. We focus on that aspect of bioinformatics that includes study of similarities in genetic data from multiple OTUs. Although research questions are diverse, a common underlying challenge is to estimate the evolutionary history of the OTUs. Therefore, this paper reviews the role of phylogenetic tree estimation in bioinformatics, available methods and software, and identifies areas for additional research and development.

Phylogenetic tree estimation and associated inference remains a research topic despite its long history dating prior to the 1970s when the first quantitative approach was developed [3]. For example, a search for "phylogenetic tree" in the text or abstract in any article in just one journal (Bioinformatics) yielded 553 papers from January 2000 through October 2008. Our purpose here is to review how trees are estimated and used in bioinformatics.

In Figure 1 we show: (A) one of the 105 possible rooted trees for five OTUs, and (B) another of the possible rooted trees for five OTUs. In (A), OTUs 1-5 are observed (external nodes), while all other taxa are unobserved (internal nodes), with taxa R being the root (ancestor) node. Time progresses forward from the root toward the external (present day) OTUs. The number N of rooted trees grows very rapidly with the number n of OTUs, $N = \prod_{k=1}^{n-1}(2k-1)$ [3]. For n =10, there are over three million possible rooted trees, and an exhaustive search of all possible trees is computationally prohibitive for more than about n =10.

Tree estimation is unique for several reasons, including: the types of data used to represent each OTU; the use of probabilistic nucleotide substitution models; the inference goals involving both tree topology and branch length, and the huge number of possible trees for a given sample of a very modest number of OTUs, which implies that finding the best tree(s) to describe the genetic data for each OTU is computationally demanding. Factors such as population subdivision, migration, and changing population size impact the tree representing a sample of OTUs [2]. Forces such as recombination, mutation, and selection impact the genetic data of a given tree. As an example of how population history impacts tree shape, Figure 2 (left plot) is the "best" tree fit to the env region of 100 contemporary human immunodeficiency viruses (HIV). The right subplots in Figure 2 shows corresponding simulated data using an implementation of coalescent theory (TreeEvolve, described in [2]) to illustrate the impact of the history of the infected population size N, with an exponential growth rate, zero growth, zero growth period followed by exponential, and a quadratic growth rate, in the top left, top right, bottom left, and bottom right subplots, respectively. Figure 2 will facilitate discussion of several topics in this review, including tree estimation and inference, tree shape, nucleotide substitution models, subtype identification using trees, and coalescent theory.

Tree estimation and interpretation are strongly linked to coalescent theory [4,5], which is a probabilistic treatment of sample genealogies. Working with a present-day sample of n OTUs, coalescent theory looks backward in time and probabilistically specifies the death process by which the number of common ancestors represented in the population decreases to n-1, n-2, ..., and finally to 1.

As an example, coalescent theory together with an assumption of constant mutation rate ("molecular clock") over time suggest via the "mitochondrial Eve" theory that all present-day humans are descendents of a woman living in Africa approximately 170,000 years ago [6, 7]. Application of coalescent theory in this case involves samples of diverse present-day human mitochondrial DNA, which is maternally inherited. Maternal inheritance simplifies the analysis from diploid (two sets of chromosomes) to haploid (one set of chromosomes), avoiding complications due to genetic mixing during reproduction. Assumptions and analyses of this human mitochondrial DNA data continue to be debated, as does the estimated coalescent time to "Eve." One of the technical issues involves the method to generate

the best tree and the method used to root the tree; generally, using an outgroup OTU that is distant, but not too distant from the other OTUs is most effective [7]. The "mitochondrial Eve" analysis has been widely misinterpreted as pointing to an Eve in the Biblical sense. An inevitable conclusion of coalescent theory is that all present day OTUs coalesce to a single ancestor who is one individual among many alive at the time of coalescence, and not an "Eve" in the Biblical sense of being the only living woman at the time.

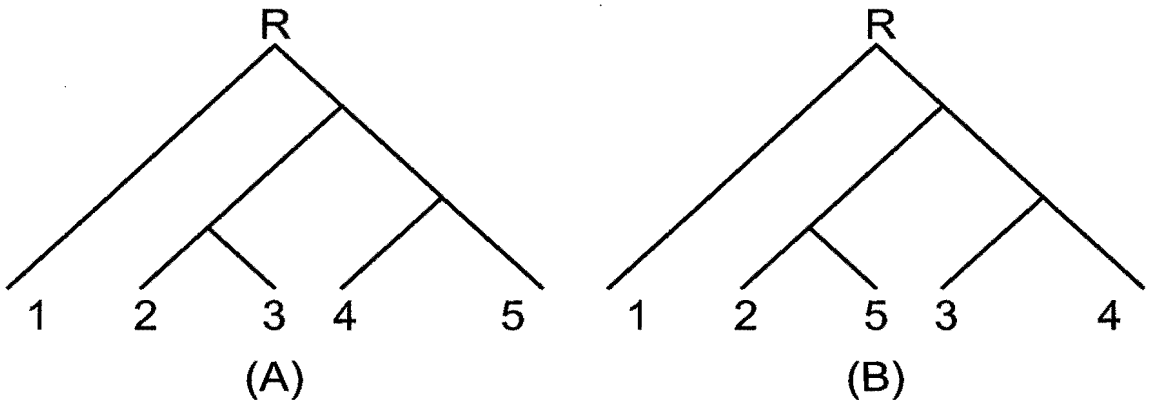

Figure 1. Two of 105 branching patterns for a rooted tree having five taxa (external nodes).

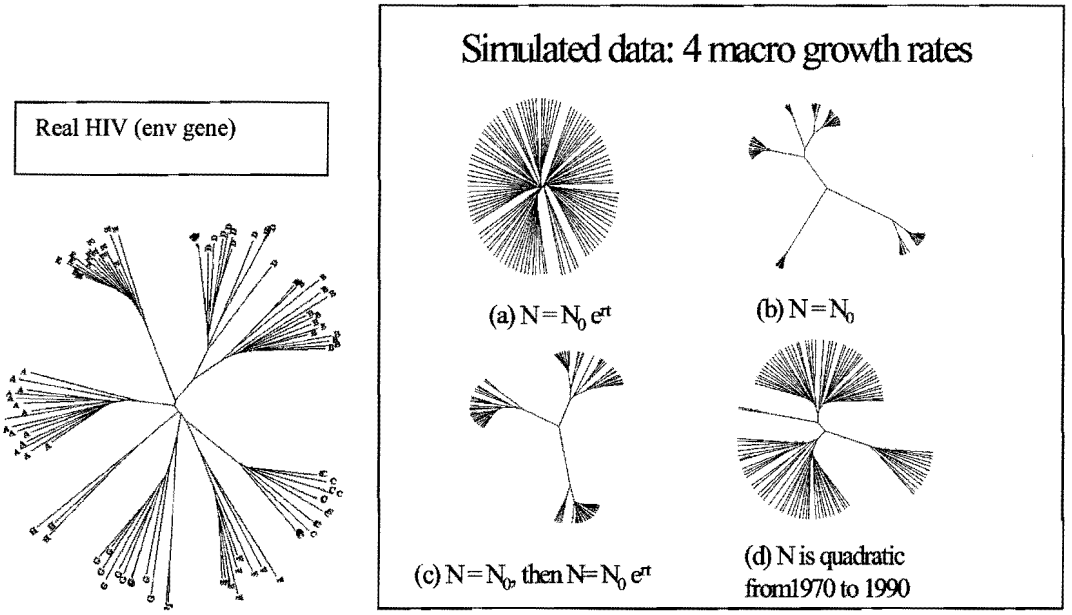

Figure 2. HIV, env gene region. Consensus trees (of 100 bootstrap samples) using maximum likelihood for real (the left plot) HIV (env gene) sequences and for coalescent-based simulated (the four right plots) sequences under different assumptions about the time behavior of the number of infected individuals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

The following sections include additional background, nucleotide substitution models, inference methods for tree estimation, methods to assign uncertainty to estimated trees, available software, applications of trees in bioinformatics, current limitations, fundamental limitations, discussion, and summary.

BACKGROUND

The first quantitative approach to estimate both the branching pattern (topology) and the ratios of branch lengths used gene frequencies and developed a maximum likelihood method based on some evolutionary model [3]. It is now common to have deoxyribonucleic acid (DNA) and/or amino acid data available from one or more regions of the genome of the observed taxa.

We focus on DNA, which usually consists of two complementary chains twisted around each other to form a right-handed helix. Each chain is a linear polynucleotide consisting of four nucleotides, two purines: Adenine (A), and Quanine (G), and two pyrimidines: Thymine (T) and Cytosine (C). In coding regions, each set of 3 nucleotides codes for one of 21 amino acids and each DNA site is called a codon. Here is example DNA data for the first 20 of the 63 informative sites from the data in [8].

OTU Code	**Mutually aligned DNA**
1	CCGGGCCTCGGCTGCGCACC...
2	CCGGGCCCTAGCCGTACACC...
3	TCGGGCCCCGGCCGCACACC...
4	TCAGGCCCCGACCGCACATC...
5	CTGAGCCCCGGCCGTATACC...
...	

The basic two concepts in the maximum likelihood (ML) approach for tree estimation are: (1) for a given topology, find branch lengths that maximize the probability of the data at the external nodes; and (2) choose the topology whose optimized branch lengths assign maximum probability to the observed data. It is important to remember that the likelihood of a given tree is not the probability that that tree is correct. Instead, it is simply the probability of observing the data if the tree and branch lengths are correct. Bayesian arguments are required for estimating the probability that a given tree topology is correct [8]. In tree estimation, several definition of "phylogenetic signal" have been suggested, each related in some way(s) to the confidence in the most likely tree(s) [9].

Tree estimation and interpretation are strongly linked to coalescent theory [1,2,4,5,10-13]. For example, coalescent theory can provide a prior probability for tree topologies [10-13]. Coalescent theory also can convert a posterior probability for topologies and branch lengths into inferences regarding population history [1,11,12] as the four right subplots in Figure 2 suggest might be possible, because the four population size growth rates each lead to quite different tree structures. More quantitatively, a skyline plot [12] is a "number of lineages through time" plot that can be used with coalescent theory to estimate the effective population size as a function of time in the

past. However, coalescent theory makes simplifying assumptions described below, so inference quality is sometimes unknown.

In some cases, we can apply a somewhat realistic coalescent-based model by defining a coalescent effective population size N_e which allows us to reference the real population to an equivalent idealized population with nonoverlapping generations and a birth-death process that is well modeled by specifying the generation time and the average and variance of the number of offspring per individual. The coalescent-effective population size N_e exists only if: (1) there is a linear, time-independent scaling of time in convergence to a coalescent process [14], and (2) the birth-death process is not linked to the nucleotide substitution model. Alternatively, perhaps recent extensions to coalescent theory can provide an acceptable approximation. All of these coalescent-based extensions make explicit and strong assumptions regarding subdivision, serial sampling, selection, and recombination. For example, see vCEBL [15].

EVOLUTIONARY MODELS

Any tree estimation method based on likelihood requires a nucleotide substitution model from which probabilities can be deduced. The most common currently used nucleotide substitution model specifies $\mu > 0$, and nonegative π_A, π_C, π_G, and π_T, which are the nucleotide relative frequencies (that sum to 1). The parameter μ determines the rate of change, and the time intervals between changes is assumed to be exponentially distributed with parameter μ. Equivalently, the number of changes in a given time is assumed to be Poisson distributed with mean $1/\mu$. The assumption that μ is constant over time is the well-known "molecular clock" assumption. Another parameter, typically denoted κ, allows for purine-to-purine or pyrimidine-to-pyrimidine mutations (called transitions) to be different than purine-to-pyrimidine or pyrimidine-to-purine (called transversions). The π's can be estimated using the observed nucleotide frequencies. The best way to estimate μ requires access to an outgroup OTU, or to the OTUs having a range of isolation times that is somewhat long compared to μ. The ratio of transitions to transversions can be used to estimate κ

A common nucleotide substitution model is as follows. Consider a pair of taxa denoted x and y. Define F_{xy} as

$$NF_{xy} = \begin{pmatrix} n_{AA}n_{AC}n_{AG}n_{AT} \\ n_{CA}n_{CC}n_{CG}n_{CT} \\ n_{GA}n_{GC}n_{GG}n_{GT} \\ n_{TA}n_{TC}n_{TG}n_{TT} \end{pmatrix}$$, where N is the number of base pairs (sites) in set of aligned sequences, n_{AA} is the number of sites with taxa x and y both having an A, n_{AC} is the number of sites with taxa x having an A and taxa y having a C, etc. Sequence alignment is a large topic that is beyond our scope, but is briefly mentioned in the Discussion section.

The most general time-reversible model (GTR) for which a distance measure has been defined [16,17] defines the distance between taxa x and y as d_{xy} = -trace$\{\Pi \log(\Pi^{-1}F_{xy})\}$ where Π is a diagonal matrix of the average base frequencies in taxa x and y and the trace is the sum of diagonal elements. The GTR is fully specified by 5 relative rate parameters (a, b, c, d, e, f) and 3

relative frequency parameters (π_A, π_C, and π_G with π_T determined via $\pi_A + \pi_C + \pi_G + \pi_T = 1$) in the rate matrix Q defined as

$$Q/\mu = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_G \\ c\pi_A & e\pi_C & f\pi_G & - \end{pmatrix}$$, where μ is the overall nucleotide substitution rate and each row sums to 0, so the diagonal entries are determined by the off-diagonal entries.. The rate matrix Q is related to the nucleotide substitution probability matrix P via $P_{ij}(t) = e^{Qt}$, where $P_{ij}(t)$ is the probability of a change from nucleotide i to j in time t and $P_{ij}(t)$ satisfies the time reversibility and stationarity criteria: $\pi_i P_{ij} = \pi_j P_{ji}$. Commonly used models such as Jukes-Cantor [16] assume that $a = b = c = d = e = 1$ and $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$. For the Jukes-Cantor model, it follows that $P_{ij}(t) = 0.25 + 0.75e^{-\mu t}$ and that the distance between taxa x and y is $-3/4 \log(1 - 4/3D)$ where D is the percentage of sites where x and y differ (regardless of what kind of difference because all relative substitution rates and base frequencies are assumed to be equal).

Important model generalizations include allowing unequal relative frequencies and/or rate parameters, and to allow the rate μ to vary across DNA sites. Allowing μ to vary across sites via a gamma-distributed rate parameter is one way to model the fact that sites often have different observed rates. If the rate μ is assumed to follow a gamma distribution with shape parameter γ then these "gamma distances" can be obtained from the original distances by replacing the function $\log(x)$ with $\gamma(1-x^{-1/\gamma})$ in the $d_{xy} = -\text{trace}\{\Pi \log(\Pi^{-1}F_{xy})\}$ formula [16]. Generally, this rate heterogeneity and the fact that multiple substitutions at the same site tend to saturate any distance measure make it a practical challenge to find a metric such that the distance between any two taxa increases linearly with time. Another generalization is to simply use the most general possible from for Q/μ, which is $Q/\mu = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ g\pi_A & - & d\pi_G & e\pi_T \\ h\pi_A & i\pi_C & - & f\pi_G \\ j\pi_A & k\pi_C & l\pi_G & - \end{pmatrix}$, where again each row sums to 0 so the diagonal entries are determined by the off-diagonal entries. The total number of parameters in this model is the 3 relative frequencies (π_A, π_C , π_G) plus the 11 relative rate parameters, a-k, with $l = 1$). To our knowledge, [18] is the only published case that used this most general model; results were indistinguishable from results using the GTR model.

At best, these simple nucleotide substitution models are a useful approximation. For example, at present, DNA sites are assumed to evolve independently, which is not likely to be true. For example, in regions that are GC rich or TA rich in the chloropolast genome, the nucleotide substitution pattern is influenced by the two nucleotides flanking the substitution site [19]. Partly to avoid context-dependent nucleotide substitution models, some studies use only DNA from each third position in the 3-codon reading frame that converts DNA to amino acids [16]. Other effects such as recombination and insertions and deletions can also be important, as mentioned in the Discussion section and in [20].

A favorable trend in phylogenetic analysis of DNA data is to choose the substitution model using goodness of fit or likelihood ratio tests [21]. However, the substitution model is likely to

depend on the region of the genome [5, 22]. For example, DNA regions that code for amino acids are more constrained over time due to selective pressure and therefore are expected to have a smaller rate of change than non-coding sequences.

INFERENCE METHODS

There are many tree-building methods in use, but ML remains the most preferred for small or modest numbers of taxa such as in most of the trees in, for example, an on-line tree database (http://www.treebase.org/treebase) which contains more than 1000 estimated trees. As a thumb rule, trees having less than 10, 10 to 100, and more than 100 OTUs are small, medium, and large, respectively. For 10 or more taxa, heuristic searches attempt to find a high likelihood (probably not maximum) tree topology, associated branch lengths, and to estimate approximately 2-15 parameters that describe the nucleotide substitution model (see the previous section).

Recently, Bayesian methods have become increasingly used. See Mr. Bayes and a few other Bayesian method implementations at http://mrbayes.csit.fsu.edu. Bayesian methods are closely related to ML in that both Bayesian and ML require an explicit nucleotide substitution model. In order to perform nearly any except the simplest Bayesian analysis, Markov Chain Monte Carlo (MCMC) is a heavily-used method that uses likelihood ratios for different sets of model parameters to generate samples from the posterior probability distribution [23]. The result is an estimated probability distribution on both tree topology (branching order) and branch lengths.

Bayesian methods typically involve the need to specify a prior probability which is updated using the likelihood to obtain a posterior probability on topology and branch lengths. The default "noninformative" prior in this context [8] is that all possible topologies are equally likely, then given a topology T, the branching times are uniformly distributed over the range defined by branch length orderings imposed by T. And, as in the ML approach, the distribution of nucleotides at each site in the root node is the stationary distribution given by the natural estimate of $(\pi_A, \pi_C, \pi_G, \pi_T)$; nucleotides at different sites evolve independently, and evolutionary processes along different branches are independent.

Having to specify a prior probability distribution is both good and bad. It is good if prior information is known and should be incorporated into the analysis. It is bad if there is essentially no apriori information, in which case one should assess sensitivity of the posterior probability to the prior, even so-called "noninformative" priors. The assumption that all tree topologies are equally-likely makes Bayesian methods similar to the ML method, but arguably superior in terms of uncertainty assessment (next section). If prior information is assumed, then coalescent methods [10] or a birth-death process [24], depending on the context could provide effective prior probabilities for topologies and branch lengths.

The Bayesian-based estimate of the posterior probability distribution can easily be used to assess confidence in any desired aspect of the estimated tree, such as whether pre-specified groups are monophyletic. A group is monophyletic if each of its member OTUs coalesces to a common ancestor before any OTU from outside the group coalesces to that ancestor. In contrast, to assign confidence to any desired aspect of the estimated tree using ML, a well-known resampling technique called the bootstrap is used in conjunction with repeated application of the ML algorithm. An example is described in the next section.

The main two alternatives to ML or Bayesian methods are distance-based methods and parsimony methods [16]. For nearly all of the currently used nucleotide substitution models, there is an associated distance measure. Therefore, if a distance method is required because there are too many OTUs for a Bayesian method or the ML method, the distance can be chosen on the basis of the best fitting substitution model, perhaps chosen using a subset of the OTUs if necessary for computational reasons. The chosen distance measure can be used to compute the distance between each pair of OTUs and various hierarchical clustering methods can then be applied. An interesting blend of likelihood and distance-based methods is provided in a weighted neighborhood method [25].

It is important to recognize that the genetic distance between any two OTUs is a function of the time since the OTUs shared a common ancestor. This time depends on the population size N, and for that reason, many studies simply attempt to estimate the effective mutation rate, $N\mu$.

Maximum parsimony is among the computationally simplest methods. It involves finding the topology that corresponds to the smallest number of substitutions required to explain the observed DNA data for each OTU [16]. Weighted parsimony methods can also involve likelihood by using appropriate weights when counting the required number of substitutions.

METHODS TO ASSIGN UNCERTAINTY TO ESTIMATED TREES

The previous section mentioned the use of ML plus bootstrap to assign confidence to various aspects of an estimated tree. Or, if MCMC is used in a Bayesian approach, then any summary of the posterior tree distribution will naturally include an uncertainty estimate. Nearly any method can at least in principle be combined with bootstrap resampling to estimate confidence in the estimated tree.

Felsenstein [26] introduced the bootstrap for assessing confidence in estimated trees. The basic bootstrap strategy is to sample each site with replacement as bootstrap sample 1. Then, apply the tree estimation algorithm to each of (typically) 100 to 1000 bootstrap samples and summarize results over all bootstrap samples. Several complications with ML plus bootstrap arise in practice. First, the number of unrooted or rooted topologies for T taxa grows prohibitively large for more than approximately 50 taxa. Therefore, unless there are only a few taxa, existing ML algorithms do not evaluate all topologies and choose the topology having ML. Instead, heuristic methods limit the search to a small number of candidate topologies. Sophisticated stochastic searches such as those using simulated annealing [27] improve matters, but do not remove the complication that the tree-finding algorithm cannot evaluate all topologies.

A second complication with the bootstrap involves interpretation and performance. Both [28] and [29] reported empirical evidence using simulated data that bootstrap proportions provide unbiased but imprecise estimates of repeatability but biased estimates of accuracy. If the phylogenetic estimation method is consistent, meaning that the estimated topology converges to the true topology in the limit as more DNA sites are used, the bootstrap appeared to underestimate high accuracies and overestimate low accuracies. Results in [28] and [29] indicated an empirically observed bootstrap bias (in the direction of underestimating confidence) in simulated data in addition to overdispersion. Results in [30] and [31] presented theoretical results to refute the bias claim. However, both [30] and [31] used theoretical arguments, largely invoking a simpler situation involving sampling from a normal distribution and inferring whether its mean was positive on the basis of the sample mean. For example, these toy but potentially insightful examples lead to the conclusion that the ML plus bootstrap based estimate of whether

a prespecified group is monophyletic can be biased low if the DNA data is highly informative and has high probability to correctly identify specified groups. Alternatively, if the DNA data is not highly informative, the ML plus bootstrap estimate can be biased high.

Reference [30] addressed the question of whether the discrete aspect of the topological space would render bootstrap estimates unreliable, and concluded that although the standard bootstrap could be improved with a very computationally intensive double-stage bootstrap (bootstrapping the bootstrap), the bootstrap is still valid (unbiased and not too imprecise) in the discrete space of tree topologies and, as [31] also concluded, does not lead to systematic over or under estimate of confidence. The discrete nature of the decision space for choosing topology 1 versus topology 2 means that the estimated topology cannot change in a smooth way as the input sequences vary. Newton [32] established a large deviation result for the bootstrap empirical distribution in a finite sample space, thereby validating nonparametric and parametric bootstrapping in some phylogenetic inference settings. Previous to this result, existing theory did not support the bootstrap because of the discrete nature of the space of tree topologies and because of the types of questions involved.

In contrast to the ML plus bootstrap approach, current Bayesian strategies use MCMC to search the posterior distribution over branch lengths and topologies. Generallly, both topology and branch lengths are estimated and while the concept of uncertainty in branch lengths is straightforward, the measure of closeness of the estimated topology to the true topology is a large topic. Several metrics have been proposed with a tendency to favor metrics that penalize mistakes near the root or center of the tree more than mistakes near the tips [33]. One common distance metric is the triples distance, defined for two trees each representing the same n OTUs as the number of the $\binom{n}{3}$ triples for which the trees have different branching order. The triples distance is strongly impacted by disagreements near the root between two trees.

The only comparisons of Bayesian confidence statements with ML + bootstrap confidence statements that we are aware of are [34, 35], which both found mild differences in confidence estimates in several real and several simulated DNA sequences. We believe that further experiments comparing Bayesian confidence measures to bootstrap confidence measures would be valuable. Such comparisons were not available when [28-31] were published. Also, reference [36] illustrated that in estimating the time to the most recent common ancestor, t_{MRCA}, using the "genetic distance" versus "isolation time" approach in [37], tree structure such as distinct clades requires a specialized bootstrap that has not been developed. Alternatively, TipDate [38] properly accounts for tree structure when estimating t_{MRCA} and associated confidence intervals. For example, the estimated confidence interval for t_{MRCA} in simulated data corresponding to real HIV sequences using TipDate [38] was much wider and more accurate for the env region of the HIV genome than the corresponding bootstrap-based confidence interval estimate applied to the "genetic distance" versus "isolation time" approach [36].

Reference [39] showed that use of a wrong substitution model leads to underdispersion and/or bias. As an example, it is well known that "long branches attract," which means the long branches tend to be grouped together regardless of whether that is the correct topology. This phenomenon will impact any method (Bayesian or ML plus bootstrap) so [34] deliberately avoided it in simulated data by using the same model to estimate the tree as to simulate the sequences. However, one can never be sure it is not in effect with real data because real data

never follow any model exactly. As mentioned, it is now more common for likelihood ratio tests or Bayes factors to be used to select models [16,22,40], but even in the best of real-data cases, we expect at least some degree of model misspecification.

One difficulty in evaluating related publications is that methods perform differently according to how closely the assumed model agrees with the actual model. For example, a ML method using the "wrong model" leads to biases, which favor recovering certain favored topologies (such as long branches tending to group together). If the true topology happens to agree with one of these "favored" topologies then the variation among ML fits on bootstrap samples will be unrealistically low. Therefore, [34] used the same model (the F84 model [16] with no rate heterogeneity and a constant substitution rate over time ("molecular clock")) in BAMBE [41] for a Bayesian method via MCMC and DNAMLk (for ML plus boottstrap in the PHYLIP suite of codes [42]) to estimate the phylogeny and also in the SeqGen code [43] to generate the data for a given topology. This means that the results with the simulated data represented a safe comparison and results on real data are as safe as possible with real data.

One main summary of a ML plus bootstrap analysis is a "consensus tree," which can be defined in several reasonable ways. The most conservative definition is a strict consensus tree that contains only groups that are exactly represented in each tree (each tree arises from a different bootstrap sample). Similarly, the Bayesian posterior can be summarized by finding a maximum posterior tree and confidence region. The confidence region could include only trees that are within a specified distance, as defined for example using the triples distance of the maximum posterior tree.

Another common summary of a tree estimation application is whether a prespecified group is monophyletic. Again, the ML plus bootstrap or the Bayesian posterior can be used to assign confidence in whether a group is monophyletic [34,44].

AVAILABLE SOFTWARE

The main software tools for performing ML tree estimation (such as DNALMk [42]) have several known limitations: (1) they use relatively restrictive substitution models; (2) they make no attempt to find the global maximum because of the so-called NP-completeness of the search; and (3) there are no diagnostics given to judge how close the estimates are to the global maximum. Nevertheless, likelihood methods are generally preferred because they: (1) have the ability to model a variety of factors thought to affect nucleotide sequence evolution; (2) have some robustness to violations of its model assumptions; and (3) have more resistance to long branch attraction than does parsimony [45]. Likelihood also makes use of more information in the data than do other optimality criteria. Also, if researchers have access to parallel computing platforms, parallel versions of FastDNAML or of PAxML [45] have been applied to ML tree estimation using up to 1000 OTUs.

The main software tools for Bayesian estimation (which also involves the likelihood calculation) include BAMBE [41], BEAST [46], and Mr Bayes [48]. It appears that Mr Bayes has overtaken BAMBE in terms of ease of use and range of available nucleotide substitution models.

BENCHMARK DATA SETS

The relative lack of real benchmark data sets for which the true branching order is at least approximately known is a current limitation for pylogenetic study. Available data sets for which the true branching order is known or approximately known include: (1) a lab-generated phylogeny using T7 bacteriophage which has been analyzed in several publications [49]; (2) protein-coding portion consisting of 12,234 base pairs of the mitochondrial genome of 19 OTUs whoe interrelationships (true branching order) are widely accepted [50] (3) an HIV transmission chain having known transmission history due to impressive contact tracing; this "known" transmission history refers to the sexual contact history rather than to the actual branching order of the sampled HIV genes [20]. Differences in the branching order of the genes and the corresponding sexual history are analogous to differences in gene versus species trees, and (4) an HIV transmission chain that is partially known due to established links between a Florida dentist and several patients who were deliberately infected during surgery [51].

Dozens of published studies have evaluated tree estimation methods using simulated data sets. To our knowledge, all such simulations follow the assumed "independent sites" substitution models.

APPLICATIONS OF PHYLOGENETIC TREES IN BIOINFORMATICS

Trees have many applications in bioinformatics, including enabling estimation of evolutionary history, population history, transmission history, rates of evolution, disease origins, identification of viral subtype, and prediction of sequence function.

A) Estimation of Evolutionary History

Simply put, a phylogenetic tree estimate acknowledges the non-independence of the DNA sequence data among OTUs due to their shared evolutionary history. Some publications use shortcut analyses that ignore the correlations among the sequences, or at least attempt to use sequences that are relatively independent by virtue of approach on "opposite sides of the estimated tree" [52]. Tree structure acknowledges shared evolution and the corresponding correlations. Reference [52] is another assessment of the coalescent time to the human Eve, but it reports much larger uncertainty in the estimate than do most other studies.

B) Estimation of Population History

Coalescent theory has provided tremendous insights and generated many research questions regarding interpretation of tree shape. The four right subplots in Figure 2 show examples that suggest how tree shape might be used, for example, to infer population history in terms of effective population size over time for the population from which a set of OTUs was sampled.

The basic concept for using tree shape to infer aspects of population history is as follows. The distribution of DNA sequences in a sample from a population involves the coalescent process, population history, and the nucleotide substitution model specifying how genetic data changes probabilistically over time. Most genetic data analyses rely on a forward model that specifies evolutionary forces and associated probabilities describing how offspring are generated. For simulating samples from a population, the state of art invokes coalescent theory [4], which uses

simplified models of the forward evolutionary process. These simplifications allow inverse analytical solutions and corresponding simulation software [2, 4, 42], commensurate with relatively low computational overhead. This is done in order to avoid having to simulate directly from the forward model and track the evolutionary histories. Sample genealogies can instead be simulated by running time from the present toward the past and tracking probabilistically when lineages coalesce to share a common ancestor. These coalescent-based simulated sample units are then used to infer how a population is evolving using features of the associated phylogenetic tree [46].

C) Estimation of Transmission History: HIV Example

Reference [20] reports results of the best four of seven common nonBayesian phylogenetic tree estimation methods applied to HIV sequences for patients linked by a known HIV transmission history. The results are encouraging in that the best fitting tree(s) are quite close to the true true. For example, some of the results in [20] were partially duplicated for this review. Figure 3 is the ML plus bootstrap consensus tree and Figure 4 is the maximum Bayesian posterior tree using BAMBE. Figure 5 is the known true tree. Disagreements are relatively minor.

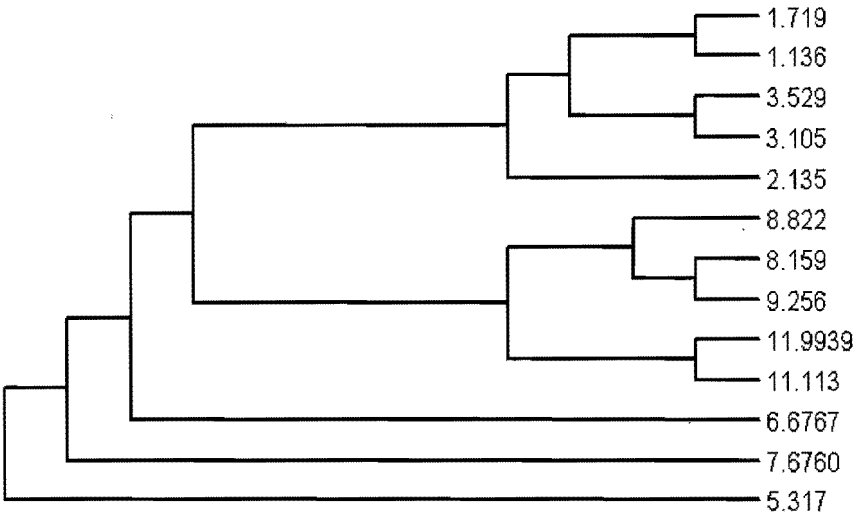

Figure 3. ML plus bootstrap consensus tree for a known HIV transmission history.

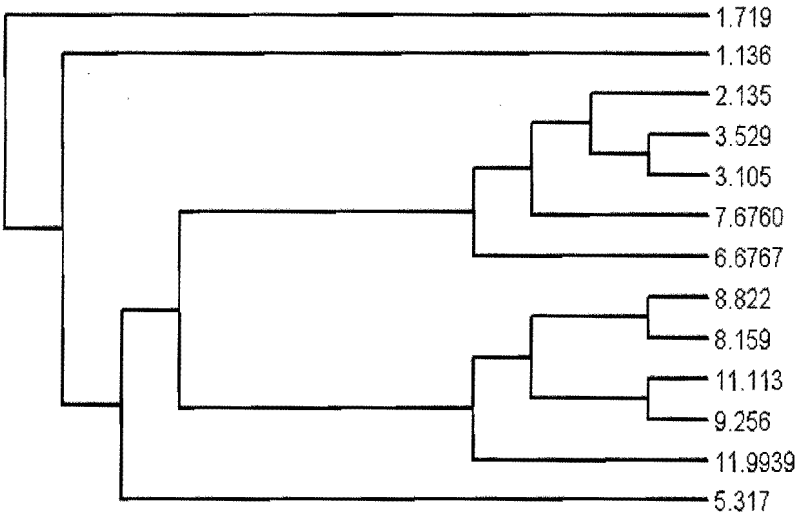

Figure 4. Maximum Bayesian posterior probability tree for a known HIV transmission history.

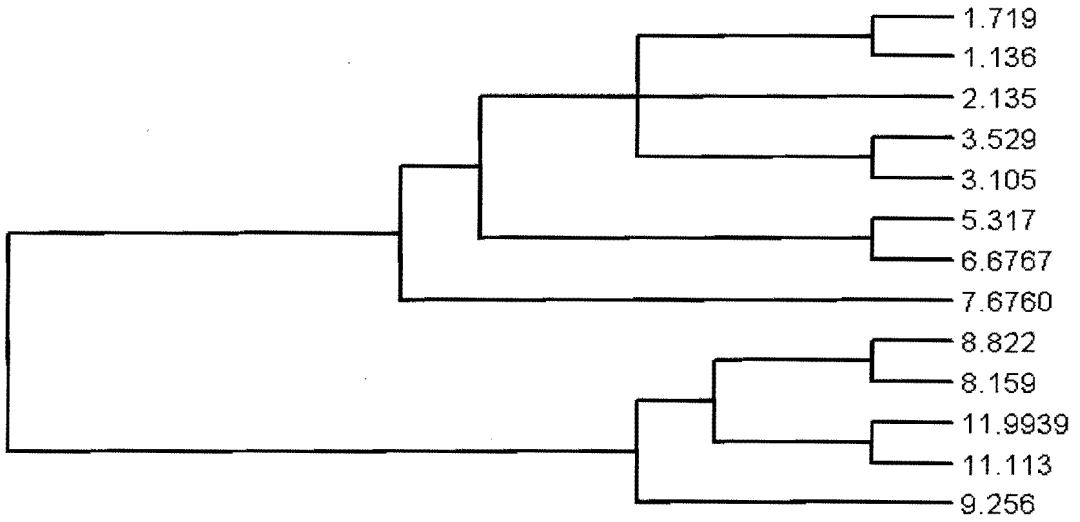

Figure 5. True tree corresponding to Figures 3 and 4.

D) Estimation of Nucleotide Substitution Rates

The nucleotide substitution model describing how DNA sequences evolve specifies the probability for various base pair substitutions as a function of the time since a given pair of OTUs shared a common ancestor. This time to a common ancestor is unknown. But, of course, a main goal in tree estimation is to estimate the branch lengths separating each OTU pair. Reference [53] used Monte Carlo (MC) sampling to address the problem of simultaneously estimating the tree topology and branch lengths and the nucleotide substitution model. The MC sampling needs to be incorporated into the standard ML + bootstrap to properly inflate variance

estimates for model parameters to account for having to estimate the topology, branch lengths, and nucleotide substitution models.

Alternatively, the BEAST implementation of Bayesian MCMC combines all model parameters into the MCMC, and clearly distinguishes five components of the full estimation problem: the nucleotide substitution model parameters (whose number depends on which model is used); the model for substitution rate variation among sites; the model for rate variation among tree branches (lineages) and branch lengths; the tree topology, and the prior probability for tree topology and branch lengths.

When the goal is tree estimation only, having to estimate the parameters of the nucleotide substitution model is a nuisance source of variation. However, model selection, interpretation, and associated parameter estimation are key goals of bioinformatics that are enabled by tree estimation. Currently, Modeltest [40] is probably the most common way to choose among candidate substitution models using likelihood ratio testing. The analogous test in Bayesian methods is the Bayes factor to choose among models [54, 55]. Modeltest currently can use trees created by ML, weighbor, or Mr Bayes.

E) Estimation of Disease Origins

Phylogenetic analysis suggests that HIV made at least two cross species transmissions from monkeys (simian immunodeficiency virus) to human resulting in HIV types 1 and 2 [56]. HIV-1 is closest to SIV strains isolated from chimpanzees and HIV-2 is most closely related to SIV from sooty mangabeys. Within HIV-1, phylogenetic analyses identified approximately 10 major subtypes (A-D, F, H, J, and K) found globally [57]. Within HIV-2, subtypes A and B are epidemic and subtypes C-G are nonepidemic.

Estimation of the timing of the cross-species transmissions is more problematic and has caused considerable confusion, analogous to the "mitochondrial Eve" confusion. For example, approximately 100 sequences spanning from approximately 1985 to 2000 isolated from HIV-1 subtype M patients have been shown in several studies to suggest a that t_{MRCA} is approximately 1930 [36, 57]. However, an estimate of t_{MRCA} is no indication when the cross-species transmission occurred. It has long been suggested that the rate of non-synonymous substitutions might increase after cross-species transmission of a pathogen, reflecting adaptation to the new host. A synonymous mutation is a "silent" DNA mutation that does not change the amino acid due to redundancy in the amino acid code, with only 21 amino acids codes for in codon triples. However, reference [58] attempted to use the rate of synonymous mutations to non-synonymous mutations to date the cross species transmission, but found ambiguous results.

F) Identification of Viral Subtype

Identification of viral subtypes in circulation and of circulating recombinants is crucial for disease mitigation strategy development such as antiviral drugs and/or vaccine [1]. Historically, tree estimates and reasonable but ad hoc methods were used to choose the number of subtypes (groups) [56],. A model-based clustering method has been investigated to choose the number of clusters (subtypes) and cluster memberships in the context of investigating how many subtypes might occur in a simulated HIV-type epidemic [5, 59].

Given a collection of subtypes defined by prior analysis such as model-based clustering [59], reference [60] showed how branch lengths can be used to determine whether newly acquired DNA sequences from test OTUs belong to existing subtypes or represent a new circulating recombinant form.

G) Prediction of Sequence Function

Generally, cross-species comparisons are thought to reveal regulatory regions of genomes. This belief is based on the hypothesis that non-coding genomic sequences having a sequence-specific function will be preferentially conserved and so have a slower substitution rate. More specifically, phylogenetic footprinting is a method to identify transcription factor binding sites within a non-coding region of DNA [61]. Such footprinting assumes that regulatory elements in non-coding genome regions are subject to purifying selection, so will be more conserved than surrounding neutral regions.

Within the conserved regions, one can screen for individual transcription factor binding sites for example [61]. A successful sequence alignment is required to align homologous (having the same ancestral site) sites and then phylogenetic analysis allows hypotheses to be generated and tested regarding gene/protein function and relationship.

CURRENT LIMITATIONS

Currently there are shortcomings or limitations of all aspects of tree estimation which we briefly describe next.

A) Limitations of Models

One of the most obvious limitations of current base substitution models is that nucleotides at different sites are assumed to evolve independently. Also, models for insertions and deletions are used during sequence alignment, but typically not during tree estimation.

When tree estimation results from different genome regions and/or different analyses of the same genome region are in conflict, it is natural to consider whether systematic estimation error due to poorly specified models is partly to blame (16, 62).

B) Limitations of Inference Methods

The most accepted inference methods use an explicit model for the nucleotide substitution process and some model comparison method such as ModelTest to choose the best available substitution model. This explicit use of a likelihood leads to either ML + bootstrap or Bayesian analysis and estimation via MCMC.

The ML + bootstrap, particularly if augmented with Monte Carlo as described in [salter] to account for the need to simultaneously estimate the tree topology and the nucleotide substitution, is limited to modest-sized trees of at most a few hundred taxa. Efforts to speed up and parallelize ML have been reported, and great speed gains have been achieved [42, 44].

Bayesian methods that use MCMC are also limited in practice to modest-sized trees. And, although uncertainty estimation is an inherent component of an MCMC calculation of the

Bayesian posterior, users must consider diagnostics to provide evidence that the MCMC chain has converged to a true realization from the desired posterior probability. The need for convergence diagnostics is not unique to tree estimation, but [63] has shown that for tree estimation, convergence can be very slow.

Finally, although the finding seems to have not drawn the attention of practitioners, reference [64] showed a nonidentifiability issue involving the use of the gamma model to accommodate rate variation across DNA sites. Nonidentifiability is problematic because it means that two or more sets of parameter values have the same likelihood. Users typically report the results of a likelihood ratio test comparing the "constant rates across sites" model to the "variable rates across sites" model and often reject the "constant rates across sites" model.

C) Limitations of Coalescent Models used to Infer Population Histories From Trees

Limitations of the coalescent techniques include: (a) Little is known concerning accuracy and robustness of coalescent theory's restrictive assumptions in many settings, although some forward models are known not to be well approximated by any coalescent model (depending on the relative time scales of various evolutionary effects such as drift, migration, and selection) [ref], and (b) Inference methods [2,9] invoke coalescent approximations to estimate the probability of candidate branching orders as part of the inference process. This leads to the undesirable situation of forcing a zero mismatch between the inference method's assumptions and the assumptions regarding how the population is evolving.

Tree estimates support inferences regarding, for example, whether a virus strain appears to be a natural branch from historical strains, or whether the strain seems to have made an unnatural leap indicating bioengineering. However, key coalescent assumptions that are violated by, for example, both HIV and influenza viruses are that all subtypes are equally transmissible and there is no recombination. Therefore, although to a limited extent and under restrictive assumptions, extensions [15] to coalescent theory have been made to accommodate recombination, selection, overlapping generations, and population subdivision, there are cases where the theory is either inadequate or the sensitivity of its conclusions to its assumptions is unknown. The corresponding inference quality using estimated trees is also unknown; the state of the art is therefore to quantify precision, but not accuracy, of inferences using coalescent theory.

The forward model is a key component of total uncertainty associated with population genetics inferences. The current approach is: specify an amenable-to-coalescent-theory forward model for how a population is evolving that includes for example, population size, structure, and selection effects; identify the coalescent effective population size N_e [14] in the nearest available coalescent model, which is often a complicated task; use the closest coalescent model to simulate sample genealogies under restrictive assumptions about the population and the sampling process. Coalescent theory was originally applied to macroscopic populations (such as plants and animals); more recently, it has been applied to microscopic populations such as virus populations such as HIV as mentioned and other viruses as described in [65]. Coalescent methods focus on histories of samples and provide a probabilistic description of sample genealogies for example by assuming a sample of nonrecombining sequences from a population with nonoverlapping generations that mixes randomly with no selection. Coalescent theory will continue to provide insight into evolutionary processes; however, there is no way to assess how robust our inferences are with respect to model violations.

It would be good to consider the extent to which growth rates, effective population sizes, selection effects, population bottlenecks, etc., can be inferred from molecular data, regardless of whether coalescent-based simulated data or our tool's simulated data is used. An example [5] involves whether the 8 to 10 approximately equidistant subtypes of HIV-1 (type M) could have arisen under available models of how HIV is evolving (left plot in Figure 2). To examine this, we used coalescent theory with questionable assumptions to simulate DNA data from a forward model of how HIV is evolving at both the macro and micro levels. This provided a reference distribution against which to compare the data. If the observed data is in the tail of the coalescent-theory-based reference distribution, then forward model used to simulate the data is not credible. A proposed new and better way to simulate sequences is to track each HIV case by geographic region including all known transmission routes such as sex, needles, blood transfusions, and mother-to-child, and track the genealogy of each case. We then sample ~100 simulated sequences from around the world or in specified regions at a snapshot in time, or distributed in time, and distributed spatially in either case. With careful bookkeeping we could deduce the sample genealogy (which 2 samples coalesced first to their most recent common ancestor (MRCA), which samples coalesced next, etc.) back in time until all 100 sequences coalesced to the single MRCA. This would produce 99 coalescent times and sample identities, which define the genealogy of the sample. This genealogy could also be thought of as the true evolutionary tree for the sample to be compared to coalescent-based genealogies.

The proposed method to track the transmission history of all units having offspring in the population is a huge computational challenge. So, we expect that coalescent theory with its unrealistic assumptions will continue in the indefinite future to provide the computational shortcut. However, we have coded an initial "brute-force" approach in Matlab and results are presented in [36], plus we are aware of a related first effort using C++/Python [66]. Still, this approach might not be sufficient, because for example with HIV, each patient carries many viral strains or sub-species and it might be important to model the bottleneck effects during transmission events. Also, we do not want to make the restrictive exchangeability assumption that all genotypes are equally transmissible. Therefore, it would be better to link the macro and micro in an explicit model that tracks viral strains within all infected individuals.

Improved models should track the macroscopic progression of infection in a host population, while also maintaining a microscopic model of pathogen evolution, giving a fidelity and capability not previously available in this area. Note that to provide a sampling option (either sampling at one time or at multiple times), one needs clever bookkeeping in order to know the true genealogy. One computational savings is that taxa that do not have offspring are killed and not tracked further.

D) Limitations of Available Software

For tree estimation and/or coalescent modeling, there is an impressive amount of shareware software written almost entirely by university or laboratory staff, often without direct support of "professional" programmers. Good web sites are evolution.genetics.washington.edu/phylip.html, evolve.zoo.ox.ac.uk, beast.bio.ed.ac.uk/Main_Page, and mrbayes.net, and many of the links from those sites.

Typical limitations of available tree estimation software include a limited range of available nucleotide substitution models, and a relatively low upper limit for the number of OTUs. Also,

model selection software is still in its infancy, although ModelTest is a very good first step [40]. Model selection research is ongoing in all areas of Bayesian analysis; for example, see [54]. Limitations of coalescent theory software implementations are as described in the previous section. We emphasize that many extensions to coalescent theory are available [2, 67, 68, 69], but not yet in a comprehensive software tool That is, currently, no coalescent-based software includes all of the now standard evolutionary factors such as overlapping generations, selection, serial sampling, recombination, and geographic isolation.

FUNDAMENTAL LIMITATIONS

There are of course fundamental limitations in phylogenetics, including: (a) ambiguous phylogenetic signal due to recombination or horizontal gene transfer; (b) possible mismatch between gene and species trees; (c) inability to infer transmission direction; (d) inability to infer when cross species transmission occurred; (e) additional uncertainty due to the need to align sequences into homologous positions by allowing insertion and deletion events;

A) Ambiguous phylogenetic signal due to recombination or horizontal gene transfer

It is well known that recombination or horizontal gene transfer cause ambiguity in tree estimation. Qualitatively, such events imply that there is not just one true tree for the DNA region of interest, which obviously confuses the analysis. Typically, recombinant sites are identified manually and avoided, or trees are fit to different sub-regions of the region of interest and tests for tree similarity [33] are used. Some of the effects of recombination are quantified in [70].

B) Possible mismatch between gene and species trees

One goal in tree estimation is sometimes to infer the branching order of various species. Of course the estimated tree actually estimates the branching order of the sampled genes, which could disagree with the branching order of the species [20].

C) Inability to infer transmission direction

Reference [20is a rich example involving estimation of a known transmission history among sexually linked HIV individuals in Sweden. Reference [50] is another good example involving a partially known transmission history linking an HIV dentist to some of his patients. It is important to realize however that there is nothing in phylogenetic tree estimation that enables inference of the direction of transmission.

D) Inability to infer when cross species transmission occurred

The subsection on Estimation of Disease Origins mentioned that the 1930 estimate of t_{MRCA} for HIV-1 subtype M does not help infer when HIV jumped from chimpanzees to humans. Related analyses involving the relative rates of synonymous and non-synonymous substitutions are believed to provide some signal for estimating the date of the jump, but [58] provided a brief analysis suggesting there was no detection signal in the example of interest.

E) Additional uncertainty due to the need to align sequences into homologous positions by allowing insertion and deletion event

Reference [71] is one example of an analysis that considers both alignment and phylogenetic analysis simultaneously. Typically, an alignment is done first allowing insertion and deletion events, and although the possibility of alignment errors is well recognized, the second step involving phylogenetic tree estimation proceeds as if there are no alignment errors. Alignment errors will always be a possibility, but allowing for insertion and deletions in the nucleotide substitution model might be worth pursuing as a way to combine the linked tasks of alignment and tree estimation.

DISCUSSION

It is important to distinguish fundamental limitations from current limitations. We must always consider the extent to which growth rates, effective population sizes, selection effects, t_{MRCA}, population bottlenecks, etc., can be inferred from molecular data, regardless of what inference approaches are used.

Despite the somewhat pessimistic list of inference and model flaws in phylogenetics, the successes have been steady and promising for decades. Particularly in the last decade when MCMC implementations have flourished, analyses can begin to more completely consider all relevant effects and noise sources associated with phylogenetic inference. For the few known or approximately known phylogenies, inference methods have performed reasonably well, perhaps as well as possible for the inference goals and sample sizes as defined by both the number of OTUs and the number of DNA sites [8, 16].

Having more real benchmark data sets with known phylogenies would be welcome. Software to simulate known phylogenies could also benefit from more realistic forward models that relax some of the coalescent theory assumptions.

SUMMARY

Phylogenetic tree estimation is central to population genetics and all aspects of bioinformatics that depend in any way on evolutionary history. The attempts to model evolution expose both our knowledge and our ignorance, and point us toward ways to improve our understanding of evolution.

References

[1] Grenfell B, Pybus O, Gog J, Wood J, Daly J, Mumford J, Holmes E., Unifying the epidemiological and evolutionary dynamics of pathogens, *Science* **2004**; 303:327-332.

[2] Grassley N, Harvey P, Holmes E., Population dynamics of HIV-1 inferred from gene sequences. *Genetics* **1999**; 151: 427-438.

[3] Felsenstein J., Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution* **1981**; 3: 368-376.

[4] Kingman J. On the genealogy of large populations, *Journal of Applied Probability* **1982;** 19: 27-43,

[5] Burr T, Myers G, Hyman J. The Origin of AIDS Darwinian or Lamarkian? *Philosophical Transactions of the Royal. Society of London B.* **2001**; 356:877-887.

[6] Cann R, Stoneking M,Wilson A., Mitochondrial DNA and human evolution, *Nature* **1987;** 325(1): 31-36.

[7] Ingman, M. Kaessmann, H, Paabo, S. and Gyllensten, U., Mitochondrial genome variation and the origin of modern humans," *Nature* **2000**; 408: 708-713.

[8] Li S, Pearl D, Doss H. Phylogenetic tree construction using Markov Chain Monte Carlo, *Journal of the American Statistical Association* **2000**; 95**:** 493-508.

[9] Revel L, Harmon L, Collar D., Phylogenetic signal, evolutionary process, and rate, *Systematic Biology* **2008**; 57(4): 591-601.

[10] Lemey P, Pybus O, Rambaut A, Drummond A, Robertson D, Roques P, Worobey M, Vandamme A., The molecular population genetics of HIV-1 group O, *Genetics* **2004**; 167: 1059-1068.

[11] Pybus O, Rambaut A, Holmes E, Harvey P., New inferences from tree shape: numbers of missing taxa and population growth rates, *Systematic Biology* **2002**; 51(6):881-888.

[12] Strimer K, Pybus O., Exploring the demographic history of DNA sequences using the generalized skyline plot, *Molecular Biology and. Evolution* **2001**; 18: 2298 - 2305.

[13] Tavare, S., Balding, D., Griffiths, F., and Donnely, P. 1997 "Inferring Coalescent Times from DNA Sequence Data," *Genetics* **145**: 505-518.

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

[15] Drummond A, Goode M., Virtual Computer and Evolutionary Biology Laboratory (vCEBL) http://www.cebl.auckland.ac.nz/pages/vcebl.html

[16] Swofford D, Olsen G, Waddell P, Hillis D. Phylogenetic inference in molecular systematics, 2nd edition, **1996**; pp. 407-514 (Hillis et al., eds.) Sunderland, Massachusetts: Sinauer Associates.

[17] Waddell P, Steel M., General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites, *Molecular Phylogenetics and Evolution* **1997**; 8(3): 398-414.

[18] Press W, Robins H., Isochores exhibit evidence of genes interacting with the large-scale genomic environment, *Genetics* **2006**; 174: 1029-1040.

[19] Yang Y, Chen Y, Li W., The Influence of adjacent nucleotides on the pattern of nucleotide substitution in mitochondrial introns of angiosperms, *Journal of Molecular Evolution* **2002**; 55: 111-115.

[20] Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J., Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis, *Proceedings of the National Academy of Sciences USA* **1996**; 93:10864-10869.

[21] Huelsenbeck J, Rannala B., Phylogenetic methods come of age: testing hypotheses in an evolutionary context, *Science* **1997**; 276: 227-232.

[22] Leitner T, Kumar, S, Albert J., Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in HIV type 1 populations with a known transmission history, *Virology* **71**: 4761-4770.

[23] Gilks W, Richardson S, Spiegelhalter D., Markov Chain Monte Carlo in practice, *Chapman and Hall/CRC:* New York, **1996.**

[24] Aldous D., Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, *Statistical Science* **2001**; 16(1):23-34.

[25] Bruno W, Socci N, Halpern A., Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction, *Molecular Biololgy and Evolution* **2000**; 17(1): 189-197.

[26] Felsenstein J., Confidence limits on phylogenies: an approach using the bootstrap, *Evolution* **1985**; 39: 783-791.

[27] Lewis P., A Genetic algorithm for maximum likelihood phylogeny inference using nucleotide sequence data, *Molecular Biology and Evolution* **1998**; 15(3): 277-283.

[28] Zharkikh A, Li, W., Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock, *Molecular Biology and Evolution* **1992**; 9: 1119-1147.

[29] Hillis D, Bull J., An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis, *Sysematic. Biology* **1993**; 42:182 -192.

[30] Efron B, Halloran E, Holmes S., Bootstrap confidence levels for phylogenetic trees *Proceedings of the National Academy of Sciences USA* **1996**; 93: 13429-13434.

[31] Felsenstein J, Kishino H., Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull, *Sysematic Biology* **1993;** 42: 193-200.

[32] Newton M., Bootstrapping phylogenies: large deviations and disperson effects, *Biometrika* **1996**; 83 (2): 315-328 .

[33] Critchlow D, Pearl D, Qian C., The triples distance for rooted bifurcating phylogenetic trees, *Systematic Biology* **1996**; *45: 323-334.*

[34] Burr T, Doak J, Gattiker, J, Stanbro, W. Assessing confidence in phylogenetic trees: bootstrap versus Markov Chain Monte Carlo. *Proceedings of the International Confereence on Mathematical and Engineering Techniques in Medicine and Biological Scienes* **2002;** 1:181-187.

[35] Mar J, Harlow T, Ragan M., Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation, *BMC Evolutionary Biology* **2005**; 5:8, doi:10.1186/1471-2148-5-8.

[36] Burr T, Gattiker J, Gerrish P., An investigation of error sources and their impact in estimating the time to the most recent ancestor of spatially and temporally distributed HIV sequences, *Statistics in Medicine* **2003**; 22:1495-1516.

[37] Korber B, Muldoon M, Theiler J, Gao R, Gupta R, Lapedes A, Hahn B, Wolinsky W, Bhattacharya T., Timing the ancestor of the HIV-1 pandemic strains, *Science* **2000**; 288: 1789-1796.

[38] Rambaut A., Estimating the rate of molecular evolution: incorporating non-conttemporaneous sequences into maximum likelihood phylogenies," *Bioinformatics* **2000**; 16(4): 395-399.

[39] Bruno W, Halpern A.,Topological bias and inconsistency of maximum likelihood using wrong models, *Molecular Biology and Evolution* **1999**; 16(4): 564-566.

[40] Posada D, Crandall K., Modeltest: testing the model of DNA substitution. *Bioinformatics* **1998**;14(9):817-818.

[41] Mau B, Newton M, Larget B., Bayesian phylogenetic inference via Markov Chain Monte Carlo method *Biometrics* **1999**; 55: 1-12.

[42] http://evolution.genetics.washington.edu/phylip.html.

[43] http://tree.bio.ed.ac.uk/software/seqgen/

[44] Stamatakis et al. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees, *Bioinformatics* **2004**; 21(4): 456-463.

[45] Burr T, Skourikhine A, Bruno W, Macken C. Confidence Measures for Evolutionary Trees: Applications to Molecular Epidemiology, *Proceedings of the. IEEE International Conference. on Information, Intelligence and Systems, Genetics and Evolution Section* **1999**; 107-114.

[46] Bergsten J., A review of long-branch attraction, *Cladistics* **2005**; 21(2): 163-193.

[47] Drummond A., Rambaut A., Beast: Bayesian evolutionary analysis by sampling trees, *BMC Evolutionary Biology* **2007**; 7:214-227.

[48] MrBayes, Bayesian inference of phylogeny (http://mrbayes.csit.fsu.edu/index.php) *Bioinformatics* **2001**; 17(8): 754-755.

[49] Hillis D, Bull J, White M, Badgett M, Molineus I., Experimental phylogenetics: generation of a known phylogeny, *Science* **1992**; 225: 589-592.

[50] Naylor G, Brown W., Structural biology and phylogenetic estimation, *Nature* **1997**; 388:528.

[51] Ou e. al., Molecular epidemiology of HIV transmission in a dental practice, *Science* **1992;** 256(50600);1165-1171.

[52] Templeton R., The 'Eve' hypothesis: a genetic critique and reanalysis, *American Anthropologist* **1993**; 95(1): 51–72.

[53] Wang Q, Salter L, Pearl D. Estimation of evolutionary parameters with phylogenetic trees, *Journal of Molecular Evolution* **2002**; 55:684-695.

[54] Doss H., Some thoughts on future directions in Bayesian model selection, *Statistica Sinica* **2007**; 17: 413-421

[55] Posada D, Buckley T., Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. *Systematic Biology* **2004;** 53(5): 793-808

[56] Hahn B, Shaw G, De Cock K, Sharp P., AIDS as a zoonosis: scientific and public health implication, *Science* **2000**; 287(5453):607-14.

[57] Korber B, Myers G., Signature pattern analysis: a method for assessing viral sequence relatedness, *AIDS Research and Human Retroviruses* **1992;**. 8, 1549-1560.

[58] Sharp P, Bailes E, Chaudhuri R, Rodenburg C, Santiago M, Hahn B., The origins of acquired immune deficienty syndrome viruses: where and when? *Philosophical Transactions of the Royal. Society of London B.* **2001**; 356: 867-876.

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

[60] Hraber P, Kuiken C, Waugh M, Geer S, Bruno W, Leitner T., Classificatin of hepatitis C virus and human immunodeficiency virus-1 sequence with the branching index, *Journal of General Virology* **2008**; 89: 2098-2107.

[61] Tagle D, Koop B., Goodman M., Slightom J, Hess D, Jones R., Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints, *Journal of Molecular Biology* **1988**; 203: 439-455.

[62] Brocchieri L, Phylogenetic inferences from molecular sequences: review and critique, *Theoretical Population Biology* **2001**; 59: 27-40.

[63] Mossel E, Vigoda E. Limitations of Markov chain monte carlo algorithms for Bayesian inference of phylogeny, *The Annals of Applied Probability* **2006**; 16(4): 2215-2234.

[64] Evans S, Warnow T., Unidentifiable divergence times in rates-across-sites models, *IEEE Transactions on Computational Biology and Bioinformatics* **2004**; 1(1): 130-134.

[65] Jenkins G, Rambaut A, Pybus O, Holmes E., Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis, *Journal of Molecular Evolution* **2002**; 54: 156-165.

[66] Peng B, Kimmel M. SimuPOP: a Forward-time Population Genetics Simulation Environment, *Bioinformatics* **2005;** 21: 3686-3687.

[67] Fu Y, Li, W., Coalescing into the 21st Century: An Overview and Prospects of Coalescent Theory, *Theoretical Population Biology* **1999**; 56:1-10.

[68] Felsenstein J, Kuhner M, Yamato, J, Beerli P. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data **1999**; 33:163-185 in Statistics in Molecular Biology and Genetics, ed. Francoise Seillier-Moiseiwitsch. IMS Lecture Notes-Monograph Series, Volume 33. Inst. of Math. Statistics and American Mathematical Society, Hayward, California.

[69] Pybus O, Rambaut A., GENIE: estimation demographic history from molecular phylogenies *Bioinformatcs* **2002**; 18(10): 1404-1405.

[70] Schierup M, Hein J., Consequences of recombination on traditional phylogenetic analysis, *Genetics* **2000**; 156: 879-891.

[71] Satija R, Pachter L, Hein J., Combining statistical alignment and phylogenetic footprinting to detect regulatory elements, *Bioinformatics* **2008**; 24(10): 1236-1242.