Reconstructing Ancestor-Descendant Lineages from Serially-Sampled Data: A Comparison Study

Patricia Buendia[1], Timothy M. Collins[2], and Giri Narasimhan[1]

[1] Bioinformatics Research Group (BioRG),
School of Computing and Information Science,
Florida International University, Miami, FL 33199, USA
giri@cis.fiu.edu
[2] Department of Biological Sciences,
Florida International University, Miami, FL 33199, USA

Abstract. The recent accumulation of serially-sampled viral sequences in public databases attests to the need for development of algorithms that infer phylogenetic relationships among such data with the goal of elucidating patterns and processes of viral evolution. Phylogenetic methods are typically applied to contemporaneous taxa, and result in the taxa being placed at the tips or leaves of the tree. In a serial sampling scenario an evolutionary framework may offer a more meaningful alternative in which the rise, persistence, and extinction of different viral lineages is readily observable. Recently, algorithms have been developed to study such data. We evaluate the performance of 5 different methods in correctly inferring ancestor-descendant relationships by using empirical and simulated sequence data. Our results suggest that for inferring ancestor-descendant relationships among serially-sampled taxa, the MinPD program is an accurate and efficient method, and that traditional ML methods, while marginally more accurate, are far less efficient.

1 Introduction

The modeling of viral evolution can be greatly improved through the study of samples isolated at different periods in time which can lead to a better understanding of the diseases caused by pathogens such as HIV-1, human Influenza A and Hepatitis C. Understanding the within-host evolution of pathogens also has implications for the development of new therapies. An increasing amount of data from rapidly evolving viral organisms sampled serially from a single host is now available in the public databases. Unlike contemporaneous data, serially-sampled data contains taxa ancestral to other taxa, and may be placed at internal nodes of an evolutionary framework (see Fig. 1) providing a more specific hypothesis of evolutionary relationships.

Holmes et al. investigated the evolution of the V3 region of the HIV envelope gene by analyzing sequences of plasma viral RNA donated over a seven-year period by a single patient [1] and created the following "evolutionary framework" (Fig. 1), stating that *"unlike most molecular phylogenies, real ancestors may be present in the data and the framework expresses the postulated ancestor-descendent relationships."* [1].

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 807 – 814, 2006.

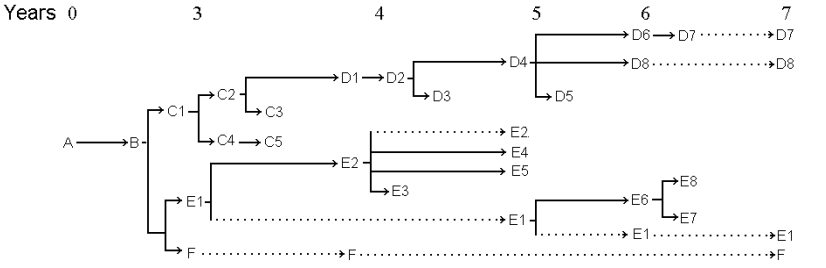

Fig. 1. Evolutionary Framework relating 24 different amino acid sequences found in the V3 loop. Redrawn in rectangular format from Holmes et al. Fig. 2 [1]. Time scale is given along the top. Dashed lines indicate identical sequences.

Hillis et al. pioneered the use of known molecular phylogenies, producing a known T7 phage phylogeny in the laboratory [2]. Cunningham et al. extended this work by serially propagating six bifurcating lineages of bacteriophage T7 according to the protocol of Hillis et al. resulting in a data set with known phylogeny [3]. We used Cunningham's experimental evolution data in our analysis as well as the Holmes viral data set and will henceforth refer to the HIV data as the Holmes92 data set, and the T7 phage sequences as the Cunningham97 data set.

More recently, several researchers have attempted to adapt existing phylogenetic methods to analyze serially-sampled data (see Methods section). The goal of our study is to compare five phylogenetic methods and assess how well they capture ancestor/descendant relationships in lineages of serially-sampled sequence data. The methods were tested on the two published phylogenies described above and on simulated data. The coalescent method was used to generate the simulated data sets. Since in practice only a small fraction of the total number of sequences representing a lineage is sampled, we incorporated a random sampling step into our simulation strategy.

Although it has been well documented that recombination is an important process in retroviral evolution [4, 5], there is only one existing method (MinPD) that directly addresses the study of serially-sampled data in the presence of recombination [6]. In this study however, MinPD's recombination detection feature was turned off.

2 Methods

In the last five years, several methods have been devised to study serially-sampled data, many of which are variants of phylogenetic methods for contemporaneous taxa. Three of the methods–sUPGMA, TipDate, and BEAST [7-9]–assume a molecular clock, i.e., assume a constant rate of evolution. They are further constrained by the traditional tree style of handling contemporaneous data. We are not including BEAST in this comparison study as it outputs a distribution of trees, while the computation of our performance score requires a single tree with branch lengths. Moreover, when choosing the tree with the highest likelihood, BEAST performs comparable to TipDate, but its computation time is at least twice as long as TipDate (see also Table 2).

Although traditional phylogenetic methods typically assume taxa are contemporaneous, it is possible to modify these methods to allow taxa to be designated as

ancestors. Therefore, we also chose fastDNAml, an efficient implementation of the traditional ML method [10, 11]. Below we provide a detailed description of each method.

Sequential-linking algorithm. (SeqLink) We chose to implement (in C language) the NJ version of the algorithm (as described in [12, 13]). The algorithm is based on the evolutionary framework published by Holmes et al. [1] and is based on two assumptions:

1. The sequence from time point n with the minimum distance to some sequence in sampling period n+1 is the ancestor of "all" sequences from sampling period n+1.
2. The ancestor of a sequence was sampled at the previous time period.

Ties are broken by using additional criteria involving NJ branch lengths. Distances were measured using the JC69 distance, as other distance measures decreased the accuracy of the algorithm.

MinPD. The distance-based MinPD method attempts to improve on the performance of the previous algorithm, SeqLink, by avoiding the strong assumptions noted above. MinPD calculates pairwise distances using the Tamura-Nei 93 method with gamma rate heterogeneity and finds a closest ancestor among all preceding sampling time periods by searching the distance matrix for minimum distances.

TipDate. (Version 1.2) was designed to compute Maximum Likelihood (ML) estimates of the mutation rate from a set of non-contemporaneous input sequences (dated tips) assuming a molecular clock and a known tree topology [8]. As tree input to TipDate we used the topology estimated by the fastDNAml method. TipDate recomputes the tree branch lengths to fit the molecular clock assumption.

sUPGMA. This is a distance-based program modified from the UPGMA method, which by definition assumes a constant rate of evolution [7]. The program was implemented as a command line script using the JAVA PAL 1.4 package available from the URL: http://bioweb.pasteur.fr/docs/PAL/. We used the rate estimated by TipDate as input for sUPGMA to analyze the data sets created with the clock model of evolution. To ensure a better performance for sUPGMA for the non-clock data sets, we settled for a mutation rate of 0.004.

Maximum Likelihood. (fastDNAml) As mentioned earlier, fastDNAml is an efficient implementation of the ML method chosen to examine how the best traditional phylogenetic methods perform with non-contemporaneous data [10, 11]. FastDNAml was used with default settings (HKY85, T_s/T_v of 2, and empirical base frequencies).

2.1 Evaluation Measures and Tools

We performed comprehensive experimentation to compare the five methods mentioned above for their ability to correctly infer ancestor-descendant lineages. By studying the placement of ancestor sequences in the trees of empirical and simulated data sets, we could observe that ancestors are often assigned to very short branches, and therefore devised a score based on branch lengths. This measure is referred to as the *Performance Score* and is based on the percentage of correctly inferred relationships. For a given taxon, the closest ancestral relative is defined as the closest

sequence from some previous sampling period corresponding to either the most recent sampled ancestor or to the closest sampled relative of an unsampled most recent ancestor.

As some methods output a phylogenetic tree without explicitly inferring ancestral relationships we created a program, Nwk2Ances, which reads in a phylogenetic tree in Newick format, and returns for each sequence the "closest" sequence from any previous sampling period. Given a phylogenetic tree with inferred lengths, Nwk2Ances uses an additive metric to search for the minimum path between a sampled sequence and a sampled ancestral sequence, where the path length is the sum of branch lengths along the path.

3 Results

3.1 Inferring Ancestor-Descendant Relationships: Empirical Data

The 31 T7 page Cunningham97 sequences were 2733 nucleotides long [3], a much more robust data set compared to the 89 short sequences (each of length 35 aa) used by Holmes et al. [1]. Other features, such as the presence of parallel evolution and the skewing of the mutational bias and the number of invariable sites by the mutagen, nitrosoguanidine, used in the propagation of the T7 page, presented additional challenges to phylogenetic reconstruction methods, especially those based on an assumption of a clocklike rate of evolution. Table 1 shows the performance scores for all five programs on the two empirical data sets. SeqLink recovered most of the Holmes92 relationships. It fared poorly with the Cunningham97 data set. The poor performance of the algorithm with the T7 page phylogeny may be due to the strong assumptions of the algorithm. MinPD recovered 100% of the Cunningham97 lineage relationships. As for the Holmes92 framework, the one notable difference was with sequence E8 sampled in year 6, for which MinPD chose sequence B (from year 3) as being a closer representative of its ancestral lineage over sequence E1 as postulated by Holmes et al. [1]. FastDNAml and the clock-based methods, TipDate and sUPGMA performed better on the Cunningham97 data set. Nwk2Ances was applied to the output trees of the ML and clock-based programs to calculate the performance score.

Table 1. Performance scores for empirical data

Programs	Performance Scores		
	Holmes92	Cunningham97	Average
MinPD	95.65%	100.00%	97.83%
fastDNAml	65.22%	96.43%	80.82%
TipDate	69.57%	92.86%	81.21%
sUPGMA PAL+TD rate	69.57%	75.00%	72.29%
Seq-Link	78.26%	10.71%	44.49%

3.2 Inferring Ancestor-Descendant Relationships: Simulated Data

A large number of DNA sequences were generated using the coalescent model of evolution [14]. These sequences were provided as input for the five programs under consideration. As before, evaluation was based on the Performance Score measure.

3.2.1 Generated Data Sets

The program Treevolve v1.3.2 was modified to return the sampled sequences from the internal nodes and the genealogy of only the sampled sequences [14]. The twister randomization function of SeqGen 1.2.7 was also added. Our modified version of Treevolve performed the following steps:

1. Generate the random tree with different combinations of tree generation parameters: Mutation rate, Recombination Rate, Clock, Number of Leaves (See Fig. 2).
2. Assign all nodes to sampling periods and randomly sample sequences from sampling periods using specified sampling parameters: Sample Size, Start of Sampling, Number of Periods (Fig. 2).
3. Output smaller tree containing only sampled sequences and linking nodes.

Twelve sets of 100 replicates each were generated for different parameter combinations. Parameters were selected based on information from published studies [15].

3.2.2 Simulation Results

Results were analyzed with the standard statistical software package, SPSS 13, by running 2-way ANOVAs on the program performance scores using one of the simulation or sampling parameters as a second variable. As variances were large and mostly overlapping, Post Hoc analysis (Bonferroni, $p<0.05$) was used to determine which differences in means were significant.

The data did not meet the assumptions of a normal distribution or homogeneous variances, suggesting an arcsine-sqrt transformation of the performance score variable. However, it is known that ANOVA is robust to violations of these assumptions as long as the data consists of large and roughly equal-sized samples [16]. Furthermore, boxplots (or normality plots) of the arcsine-sqrt score did not show any marked skewness, and the interaction and Post Hoc results remained identical with or without the transformation. Here we report the untransformed performance scores and show the graphs for the interaction effects that were found to be significant (see Fig. 2).

Clock: Among all our experiments, the only case when a clock method, TipDate, outperformed MinPD and ML, was when the data sets were generated assuming the clock-like model of evolution. Even for these data sets, the difference was within a margin of 4-6% (see Fig. 2) and was significant from MinPD's performance scores, but not from fastDNAML scores. The effect size, measured by the partial eta-squared value of 0.23, was the largest among all the interactions.

Recombination Rate: Picking one recombinant donor (parental sequence) out of two or more (a partial match) counted as a correct prediction. This scoring choice explains why the performances were not affected by the higher recombination rates, but also suggests that the programs placed one of the recombinant donors closer to the reference sequence about as frequently as they did with ancestors of non-recombinant sequences. Effect Size was only 1%. (MinPD's recombination detection feature was turned off).

Mutation Rate: MinPD and fastDNAml displayed improved performance at a mutation rate of 10^{-5}, while the clock methods achieved their best performance at a faster rate of 10^{-4}, which resulted in highly divergent sequences. At a slower rate of 2×10^{-6} (at 10^{-6} fastDNAml would severely slow down) all of the methods had difficulty inferring ancestor-descendant relationships. The effect size was small at 3%.

Leaves: More leaves imply a larger tree and therefore, a smaller sampling rate. The performance scores declined slightly for the larger trees. The partial eta-squared value was just 0.013. The performance scores of fastDNAml for the 50K leaves data sets were significantly better than that of the other programs, including MinPD.
First Sampling Period (Start): Earlier initiation of sampling resulted in sampling of sequences closer to the root. The plots confirmed that inferring relationships is more difficult when sampling starts after sequences have diverged and separated into different lineages. The interaction effect size was 9%.

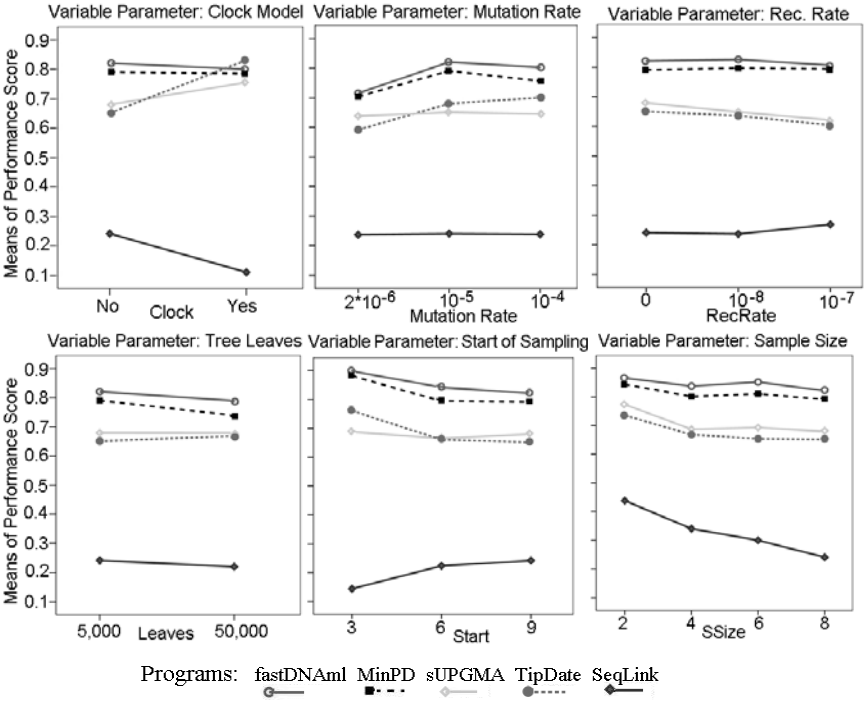

Fig. 2. Graphs showing dependence of the algorithm's performance scores on the different parameters. The analyses were based on experiments with 100 replicates generated with a modified version of Treevolve1.3.2. All parameters were fixed except the ones described in the horizontal x-axis. Wherever unspecified, parameters were fixed as follows: Mutation rate: 10^{-5}; Start: 9; Periods: 6; Clock: No; Sample size: 8; Recombination rate: 0; Leaves: 5000, Sequence Length:1000, Model HKY, T_s/T_v:4, Gamma:0.5.

Sample Size (SSize): The interaction effect was very low (1%). Only the decline in performance (-5%) from size 2 to 4 for the top three programs was deemed significant by the ANOVA analysis. The lower error rates for smaller sample sizes are attributed to the smaller pool of ancestors that the programs end up choosing from.

4 Discussions

The investigation of ancestor-descendant lineages from serially-sampled viral sequences data will provide new insights into evolutionary patterns and potential development of the viral population in a host. We studied five different phylogenetic methods on two significantly different sets of viral sequence data: the Holmes92 [1] and the Cunningham97 data set [3], as well as on simulated data. We carried out simulations to better understand the effect of sampling and quality of data sets on the performances of these five methods. The results show that for inferring ancestor-descendant relationships, the distance method MinPD outperforms methods specifically designed to analyze serially-sampled sequence data. An important characteristic of MinPD is that it does not require an explicit assumption of a molecular clock.

The only method that performed consistently better (but rarely was the difference statistically significant) than MinPD is the ML-based fastDNAml, which was designed for contemporaneous data. Due to the similarity in performance of MinPD and fastDNAml, they were often grouped together in the same homogeneous subset by the Tukey range test procedure. MinPD is, however, several orders of magnitude faster than fastDNAml. Table 2 shows the average computation time for the five methods.

Table 2. Comparison of computation times for all five methods on 10 runs with 80 sequences of length 1000 on a Pentium 4, 2.40 GHz CPU with 512MB of RAM running Windows XP

Program	Average Time	Standard Deviation
sUPGMA	<1 second	0
Tipdate	67 minutes	52.44
fastDNAml	38 minutes	4.47
MinPD	<1 second	0
SeqLink	<1 second	0

The results of our simulation studies showed that certain sampling practices can adversely affect the outcome of the programs. Sampling from a larger tree increased the error rate by about 5%. Sampling early on, before the sequences have considerably diverged, produces better-defined phylogenetic relationships. In contrast, different sample sizes did not significantly affect program performance. The presence of a high rate of recombination did not increase the error rate when partial matches were counted towards the score. There is a decrease in performance for faster or slower mutation rates. As expected the two clock-based methods performed well with data generated by a clock model.

MinPD had a better performance than fastDNAml with the two empirical data sets in this study, it is possible however, that our simulation sampling procedure may have provided a slight advantage to a traditional method like fastDNAml as most sampling started later, i.e., closer to the leaves of the final time line. By starting to sample later we tried to emulate the scenario in which sampling from a patient starts in the later stages of infection when the virus has had ample time to evolve and diverge into different lineages. More studies with empirical data and simulated data generated under new combinations of parameters may provide a better understanding into the effects of data set composition on program performance.

Acknowledgements

P.B. was supported by MBRS-RISE Fellowship (NIH/NIGMS R25GM61347). G.N. was supported in part by NIH Grant P01 DA15027-01.

References

1. Holmes, E.C., et al.: Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA.(89) 1992 p. 4835-4839
2. Hillis, D.M., et al.: Experimental phylogenetics: generation of a known phylogeny. Science. 255.(5044) 1992 p. 589-592
3. Cunningham, C.W., et al.: Parallel molecular evolution of deletions and nonsense mutations in bacteriophage T7. Mol Biol Evol. **14.**(1) 1997 p. 113-6
4. Robertson, D., et al.: Recombination in HIV-1. Nature. **374.** 1995 p. 124-6
5. Mikkelsen, J.G. and F.S. Pedersen: Genetic reassortment and patch-repair by recombination in retroviruses. Journal of Biomedical Sciences. **7.** 2000 p. 77-99
6. Buendia, P. and G. Narasimhan. MinPD: Distance-based Phylogenetic Analysis and Recombination Detection of Serially-Sampled HIV Quasispecies. In IEEE Computational Systems Bioinformatics Conference. Stanford, CA. (2005)
7. Drummond, A. and A.G. Rodrigo: Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA (sUPGMA). Molecular Biology and Evolution. **17.** 2000 p. 1807-1815
8. Rambaut, A.: Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics. **16.** 2000 p. 395-399
9. Drummond, A. and A. Rambaut 2003. BEAST v1.0. http://evolve.zoo.ox.ac.uk/beast/
10. Olsen, G.J., et al.: fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10.** 1994 p. 41-48
11. Felsenstein, J.: Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. **17.** 1981 p. 368-376
12. Ogishima, S., F. Ren, and H. Tanaka: Reconstruction and Analysis of Within-host Longitudinal HIV-1 Evolution by a Distance-based Sequential-linking Algorithm. Chem-Bio Informatics Journal. **1(2).** 2001 p. 73-83
13. Ren, F., S. Ogishima, and H. Tanaka: Longitudinal phylogenetic tree of within-host viral evolution from noncontemporaneous samples: a distance-based sequential-linking method. Gene. **317(1-2).** 2003 p. 89-95
14. Grassly, N., P. Harvey, and E. Holmes: Population dynamics of HIV-1 inferred from gene sequences. Genetics. **151.** 1999 p. 427-438
15. Shankarappa, R., et al.: Consistent Viral Evolutionary Changes Associated with the Progression of HIV 1 Infection. Journal of Virology. **73.**(12) 1999 p. 10489-10502
16. Glass, G., P. Peckham, and J.R. Sanders: Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research. **42.** 1972 p. 237-288