Blind dating – A phylogenetic approach to dating HIV reservoir sequences

Joshua Horacsek[1,2], Jeffrey B. Joy[2],
Zabrina L. Brumme[1,2], and Art F.Y. Poon[1,2,3]

June 26, 2015

Abstract

Background: The ability of HIV to persist within latent cellular reservoirs represents a major barrier to cure. The timing of establishment of individual viral reservoirs over the infection course could influence their susceptibility to elimination by immune-mediated or therapeutic approaches. However, dating methods to accurately estimate the age of reservoir sequences remain scarce. We propose a simple method to date suspected reservoir sequences using phylogenetic approaches.

Method: Simulated sequence data for model validation were generated using INDELible version 1.03. Published longitudinal clonal sequences from untreated HIV-infected individuals with estimated dates of infection were obtained from the Los Alamos National Laboratory database. Maximum-likelihood phylogenies were reconstructed with PhyML. Phylogenies were rooted by determining the location of the root that minimized the root-mean-square error between root-to-tip distances and known dates of sampling. The root-to-tip distances of latent sequences were mapped to the optimal regression line to estimate their establishment date, which was assumed to precede their sample dates by an unknown amount.

Results: We validated the root-to-tip method using simulated data and published longitudinal clonal sequence datasets from untreated HIV-1 infected individuals with known infection dates. For each dataset, arbitrary selections of up to 10

Conclusions: Given a known phylogeny comprised of longitudinal plasma HIV-1 RNA sequences, the establishment dates of unknown (reservoir) sequences can be reliably estimated when within-host HIV sequence evolution conforms to a molecular clock. Future studies will test this method on empirically derived longitudinal within-host HIV deep sequence data and extend the method to work under alternative clock models.

Keywords: HIV, Latency, Cellular Reservoirs, Phylogenetics, Linear Regression

1 Introduction

Latent viral reservoirs within HIV-1 infected individuals are a major obstacle to overcome on the path to a cure for HIV (Pace *et al.*, 2011). Reservoirs are cells that contain integrated viral DNA but have entered a dormant state that not only has a low output of virus, but can persist for years. Even though antiretroviral therapy (ART) can reduce a patient's virial load to below detectable levels, if a patient stops ART, these latent reservoirs can reseed an infection (Joos *et al.*, 2008; Pomerantz, 2003; Richman *et al.*, 2009).

Not much is known about when latent reservoirs are established during the course of infection. The timing of the establishment of reservoirs is of interest because the age of the stored virus may provide information as to the types of adaptations to the host's immune system the virus may have developed. Specifically, it may have an influence on how those infected cells react to immune-mediate or therapeutic treatments.

However, there hasn't been much work assigning dates to sequences that are suspected to be latent (there has been work on detecting latent reservoirs Immonen and Leitner (2014)). We propose a simple framework that utilizes the assumption of a strict molecular clock on the evolution of HIV within host, and that extracts timing information from the phylogenetic relationship between the virus sampled at various timepoints along the infection. We first construct a phylogeny for plasma and PBMC samples from a patient, then calibrate a strict clock for that tree based on solely the timing for the plasma sequences. We assume that PBMC cells are

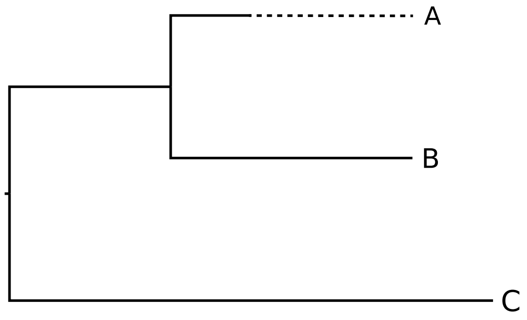

Figure 1: The dotted line in the above figure is an example of latency. It was archived at time t, yet was collected at the same time as B, that is, at time t'

latent, as they contain only integrated cells, and therefore have stopped evolving. We don't try to do any classification of these cells.

The main effect of latency on a phylogenetic tree, is that you'll probably end up seeing branches that terminate before their sampling date. If you want to estimate these dates, you can no longer make the assumption that sampling time represents when that element left the population. The death event of a birth death model is when that cDNA gets integrated into the cell. Therefore, we try to accept the parsimonious model, and argue that, when a strict clock is appropriate, we should use it.

We first tested our methodology on simulated data without any latent behavior to assess the sensitivity of reconstruction on the underlying data. When then used previously published data on plasma sequences only to to reconstruct plasma dates. Finally we applied our methodology to sequence data to X patients scraped off of the HIV lanl database.

Talk about why we didn't use beast and Bayesian approaches.

2 Methods

To validate our approach, we first simulated sequence data along a phylogeny with a strict clock. We reconstructed phylogenies and randomly chose tips to censor, then reconstructed those dates from the clock. We then tested our method on patient data that was RNA only, and did the same thing, finally, we look at a homogenous dataset with both plasma and PBMC data.

2.1 Simulated Data

Sequences were generated in INDELible 1.03 (Fletcher and Yang, 2009) along a seed phylogeny. This phylogeny was generated in TreeSim (**?**), and was given a clock from NESLI (**?**). Sequences were generated using an HKY85 substitution model with parameters set from a maximum likelihood estimate from published datasets (McCloskey *et al.*, 2014). A maximum likelihood phylogeny was then reconstructed from these sequences with RAxML (Stamatakis, 2014), which was then rooted using root-to-tip regression. 50 trees of 100 tips each. To test the effect of data on the clock calibration, we randomly censored the dates of 5, 10, 25 and 50 tips, then reconstructed those from the clock.

2.2 Data Collection

For our second set of experiments we used a previously harvested set of data (McCloskey *et al.*, 2014). These data had been harvested from the HIV lanl database. Our third experiment also used data from Los Alamos Intra-Patient search interface (Los, 2015), except this time we used it to identify patients with both plasma and PBMC sequences from the ENV region of the HIV-1 genome.

2.3 Patient Data

For our second set of experiments, we looked at RNA sequences from untreated patients from longitudinal studies with two or more clonal or single-genome (SGS) sequences available at each time point, with a known time- line relative to one of several reference points: HIV infection, seroconversion, presentation of symptomatic seroconversion illness, or birth. The seamples also wertr collected within 6 months (186 days) of this reference point, and where at least one of the subsequent (follow-up) time points occurred a minimum of 6 months after baseline. Known cases of superinfection were also excluded.

We selected published studies that had well-characterized partial env sequences generated via clonal sequencing methods or single-genome amplification from HIV RNA in blood plasma or integrated DNA from PBMC cells. There was no condition for the time between sample collections. Fifteen individuals were included with serial samples from at least 3 time points (2.3) (remove 3 from table). A total of X partial env

Patient ID	# of plasma samples	# of PBMC samples	Total # of Seqeunces	# Plasma Time points	# PBMC time points	Total Time points	Reference
820	50	87	137	5	10	15	Reference
821	76	192	268	7	17	17	Reference
822	32	98	130	3	10	10	Reference
824	100	107	207	7	9	13	Reference
10137	24	106	130	2	10	12	Reference
10138	82	119	201	6	13	16	Reference
10586	16	121	137	2	12	14	Reference
10769	229	96	325	10	4	11	Reference
10770	190	31	221	11	2	11	Reference
13889	151	132	283	13	14	18	Reference
16616	16	95	111	2	5	5	Reference
16617	16	89	105	2	5	5	Reference
16618	25	75	100	2	5	6	Reference
16619	13	60	73	2	6	6	Reference
34375	16	73	89	1	7	7	Reference
34382	37	5	42	3	1	4	Reference
34391	13	38	51	3	5	6	Reference
34393	25	74	99	2	6	7	Reference
34396	23	46	69	2	5	5	Reference
34397	26	25	51	2	2	4	Reference
34399	61	88	149	5	9	12	Reference
34400	54	23	77	4	1	5	Reference
34405	27	29	56	3	2	4	Reference
34408	5	73	78	2	8	8	Reference
34410	35	60	95	2	6	6	Reference
34411	25	43	68	3	3	6	Reference

Table 1: Patient data collected from the HIV LANL database

sequences blood plasma specimens and Y partial env sequences from PBMC cells were used in our analysis.

2.4 Sequence Aligment

All of our sequences were annotated in FASTA formats, each containing multiple time points. Our simulated sequences were not simulated with any insertions or deletions, and no alignment was necessary, additionally, the dataset we used for our second set of experiments was already cleaned and aligned (McCloskey *et al.*, 2014). For our third dataset, we used the built in MUSCLE (Edgar, 2004) interface within AliView (Larsson, 2014) to align the sequences, which were then visually inspected and modified.

2.5 Phylogeny Reconstruction

In all experiments all of our reconstructed phylogenies were maximum likelihood phylogenies reconstructed in RAxML (Stamatakis, 2014) using the GTR+Γ model. Synthetic data were solely rooted with root-to-tip regression. For all other patient data, both outgroup rooting (against the HIV-1 B ancestor consensus (**?**)) and root-to-tip regression Paradis *et al.* (2004) were used to root the tree. We did not attempt to use BEAST (Drummond and Rambaut, 2007) to build any phylogenies, as leaving dates unspecified on the amount of tips we had drastically increases the dimensionality of the problem, and would need a prohibitively large amount of samples from the MCMC chain to converge.

2.6 Date Reconstruction

Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date.

2.7 Patient Rejection

Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date.

3 Results

Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date.

3.1 Simulated Data

Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date.

3.2 RNA Only Data

Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date.

3.3 Patient Reconstruction

Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date. Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date. Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date.

4 Discussion

The simulated data are unsurprising, the RNA data too. Who knows with the patient stuff. Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date. Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date. Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date.

5 Conclusion

Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date. Once the trees had been rooted, a general linear model with the normal family was constructed with the with expected number of substitutions as the input, and sampling time as the response. For all experiments using patient data, this fit was only over the plasma data, and the PBMCs expected subs were used as input. From this, we collected the difference between the observed sampling date, and the predicted date. Talk about what we saw. Talk about what can still be done.

6 Acknowledgments

The authors would like to thank XXX.

References

(accessed June 24, 2015). *Los Alamos HIV Database*.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**(5), 1792–1797.

Fletcher, W. and Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, **26**(8), 1879–1888.

Immonen, T. T. and Leitner, T. (2014). Reduced evolutionary rates in HIV-1 reveal extensive latency periods among replicating lineages. *Retrovirology*, **11**, 81.

Joos, B., Fischer, M., Kuster, H., Pillai, S. K., Wong, J. K., Böni, J., Hirschel, B., Weber, R., Trkola, A., Günthard, H. F., *et al.* (2008). Hiv rebounds from latently infected cells, rather than from continuing low-level replication. *Proceedings of the National Academy of Sciences*, **105**(43), 16725–16730.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, **30**(22), 3276–3278.

McCloskey, R. M., Liang, R. H., Harrigan, P. R., Brumme, Z. L., and Poon, A. F. (2014). An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data. *J. Virol.*, **88**(11), 6181–6194.

Pace, M. J., Agosto, L., Graf, E. H., and O'Doherty, U. (2011). Hiv reservoirs and latency models. *Virology*, **411**(2), 344–354.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pomerantz, R. J. (2003). Reservoirs, sanctuaries, and residual disease: the hiding spots of hiv-1. *HIV clinical trials*, **4**(2), 137–143.

Richman, D. D., Margolis, D. M., Delaney, M., Greene, W. C., Hazuda, D., and Pomerantz, R. J. (2009). The challenge of finding a cure for hiv infection. *Science*, **323**(5919), 1304–1307.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9), 1312–1313.