Design Document: Phylogenetics, Phylodynamics, and Phylogeography in Pyro

First edit 2020-07-30 - Last edit 2020-12-20
Authors: @fritzo

Objective

Enable flexible scalable Bayesian phylogenetic, phylodynamic, and phylogeographic modeling in Pyro.

Specific feature requests (prioritized)

+ Scale to the 250k-sample (and growing) COVID-19 dataset at gisaid.org.
+ Support phylogeographic modeling where each sample is timestamped and geotagged, and we have weak prior information on inter-region transit rates.
+ Support scRNA-seq data where each read represents only a partially observed genome, sometimes mixed.
+ Support within-host variants similar to (Miao et al. 2018).
+ Support superspreading / multifurcation models along the lines of (Hoscheit & Plybus 2019).
+ Support active learning / experimental design: "who should we sequence?"

Out of scope

- Modeling crossover, incomplete lineage sorting, diploid genetics, species trees.
- Selection pressure?
- Inhomogeneous mutation rate (e.g. cancer)?

Design Overview

We aim to implement as little new machinery as possible, and instead rely on existing Pyro machinery composed in new ways and with a few new components and existing external tools. We will focus on four stages of tasks, each building on the previous:

Prediction / Counting

Given a posterior distribution over trees and other variables, make it easy for users to answer novel queries such as "what portion of my city's new infections have internal vs external source?", "when was the first

infection in my city?", or "what regional attributes correlate with superspreading?". Some of these tasks require a little inference, but often they are simple aggregation operations. This effort will result in new tutorials and possibly new helpers like Predictive, ForecastingModel.predict(), or CompartmentalModel.predict().

Inference given a fixed tree

In cases where the phylogeny is only loosely coupled to other latent variables, we would like to use Pyro to infer latent variables conditioned on a fixed tree (or bag of trees) generated by another tool such as Beast/Beast2. This is the approach taken by (Fourment and Darling 2019), who perform variational inference in STAN. An early example of this in Pyro is the CoalescentRateLikelihood in contrib.epidemiology. This effort will result in new tutorials, possibly new constraints e.g. for inferring branch lengths, possibly new likelihoods e.g. for birth/death and branching processes, possibly new data structure wrappers to interface e.g. tskit objects with PyTorch/JAX.

Joint inference involving trees

Build a new Coalescent distribution (or similar component) that is compatible with variational inference and subsampling, and thus Pyro's most mature automatic inference algorithms. This should allow upstream dependency on global parameters (e.g. GTR mutation model) and downstream dependency on other observations (e.g. Pyro's CoalescentRateLikelihood).

Experimental design

Optimally allocate limited sequencing testing capacity among a population so as to most accurately answer specific questions of interest to policymakers. Who should we sequence?

Work Plan

Work will be driven by a combination of tutorials, analyses, and experiments.

Inference given a fixed tree

Analysis: Region-dependent superspreading.

Split say COVID data into regions (e.g. countries, states, provinces) and use an existing tool to infer phylogenies within each region. Then construct a joint model using contrib.epidemiology and CoalescentRateLikelihood, and hierarchically share some parameters across regions. Infer both R0 and superspreading dispersion k across all regions. Optionally then extend the model to search for relevant region-dependent covariates on which R0 and k might depend (e.g. population density, policy, temperature), where coefficients are fit via joint Bayesian regression.

Joint inference involving trees

Tutorial: Joint biogeographic inference.

Possible data source: Geo-tagged genomes from gsaid.org.

Tutorial: Jointly fit epidemiological parameters and a tree.

Fit epidemiological parameters (R_0=basic reproductive number, k=superspreader dispersion, possibly even a α=Beta-coalescent stability) jointly with a phylogeny. Possible data source: Nextstrain COVID phylogenies.

User interviews

2020-08-03 applications to within-host single cell sequencing

Attendees: Mehrtash, Nick, Pyro folks

- What are the unique challenges of single cell sequencing?
    - sample sequence reads are only very sparsely observed
    - samples may be mixed within each cell (e.g. viruses)
- What are some open questions that could be resolved by phylogeny inference?
    - Within a single host, are mutations completely random or under selective pressure?

2020-08-05 applications to virus tracking

Attendees: Sabeti lab folks, Pyro folks

- What do you do now?
    - Run either Beast or faster ML tools.
    - Beast requires expertise.
- What features do you require of existing tools?
    - Existing datasets do not handle large datasets.
    - Would like to handle multifurcation.
    - We need easier ways to compute summary statistics on (bags of) trees.
        - counting
        - inference
- What new features would be useful?
    - We'd like to estimate superspreader parameter k.
        - Ideally per locally outbreak.
- Data scale?
    - Typically 10s to 100s.
    - 500-1000 samples, each in 10k samples (each 32kb in size)
- Does topological uncertainty matter? How much is there?
    - There is a lot of uncertainty.

2020-08-06 applications to virus tracking

Attendees: CZ folks, Pyro folks

- Does topological uncertainty matter? How much is there?
    - There is a lot of uncertainty.
    - What matters most is uncertainty at the root, or at a geographic region's root. We want the distribution over the time of initial infection, globally and within each geographic region.
- What is your biggest feature request?
    - Scalability is a huge issue for phylogeny inference.
- Do you need to model recombination / crossover?
    - Yes, recombination can complicate models, e.g. polio can exchange among strains and even among other polio-like viruses. Sometimes crossover is an issue, e.g. flu can exchange a subset of its multiple segments. However there is as yet no evidence for crossover in COVID.
    - Sometimes we handle this in preprocessing, restricting attention to a single gene and inferring a gene genealogy.
- Do you need to model selection pressure?
    - We don't usually model selection pressure. Sometimes in flu.
    - Birth-death processes can model selection, but also there may be ways to adapt coalescent models under selection.
- What kind of complex models do you consider?
    - Phylogeographic models are currently important. We want to fuse mobility data with phylogenetic data. E.g. Beast2 has a GML model where you can enter flight pattern data.
- Does multifurcation matter?
    - It may matter more in small populations. Coalescent models typically perform worse when sample proportion is high; in that case birth-death models perform better. It may be that multifurcating coalescent models may perform better in small-population high-sample-proportion settings.

Background

Related tools

Domain-specific phylogenetic and phylodynamic tools

- Beast (LGPL) - MCMC, robust
- Beast2 (LGPL) - MCMC, has many custom modules
- MrBayes (GPL3) - MCMC
- PhyML (GPL3) - maximum likelihood phylogenetic inference
- ExaML (GPL3) - maximum-likelihood phylogenetic inference on supercomputers
- RAxML-NG (AGPL3) - maximum-likelihood phylogenetic inference (example)
- IQ-TREE (GPL2) - maximum likelihood phylogenetic inference with ultrafast bootstrap

- BitSeq - cancer
- PhyloWGS (GPL3) - cancer phylogenies
- PHYLIP - Joe Felsensein's phylogeny inference software

The Cyberinfrastructure for Phylogenetic Research hosts a number of scalable tools.

PPL-based phylogenetic and phylodynamic tools

- RevBayes (GPL3) - a phylogenetics-specific Bayes net modeling language + MCMC inference.
- Birch, WebPPL. Ronquist et al. (2020) develop SMC algorithms for phylodynamic inference.

Tree data structures and low-level libraries

Apache-compatible licenses:

- tskit (MIT) - low level data structure for multi-trees with crossover (MIA talk)
- ARGWeaver (MIT) - phylogeny inference with crossover
- LiCHEE (MIT) - fast scalable maximum likelihood inference of cancer phylogenies (paper)

Non-Apache compatible licenses:

- msprime (GPL) - annotates tskit trees with mutations
- fwdpp (GPL), SLiM (GPL) - flexible forward simulators for trees (uses tskit & msprime)
- tsinfer (GPL) - infers correlated genealogies from genetic variation data
- libpll (AGPL3) - low level library for phylogenetic analysis. backend for PLL and RAxML-NG
- fastARG (no license)
- BAMSE (GPL 3) - BAyesian Model SElection for Inferring the subclonal history of tumor samples (paper)

Inference algorithms

Scalable maximum likelihood, maximum parsimony, and bootstrap algorithms

Saitou and Nei (1987) defined the classic neighbor joining algorithm for greedily constructing phylogenies. Guindon et al. (2010) describe PhyML, a maximum likelihood inference tool. Minh et al. (2013) and Hoang et al. (2017) implement an "ultrafast" approximation to Efron's nonparametric bootstrap to estimate uncertainty in phylogenetic reconstruction; their algorithm is implemented in the IQ-TREE system.

TODO many others: RAxML, …

Joint Bayesian inference algorithms

Schiffman et al. (2018) define a tree model whose tree generalizes a Dirichlet diffusion tree and whose observations (scRNA-seq transcriptomes) are noisy; their **MCMC** inference strategy uses belief propagation to draw joint cell state samples condition on trees; after tree topology moves they sample state before MH rejection, which empirically increases acceptance rate. Ronquist et al. (2020) express phylogenetic models as probabilistic programs involving BirthDeath distributions, and implement **SMC** inference algorithms in two different universal PPLs: WebPPL and Birch. Zhang and Matsen (2019) develop two full **variational inference** strategies for inferring tree topology + branch time, both strategies restricting to a small support of

high-probability topologies: (1) multi-sample KP(q,p) minimization using a VIMCO estimator for discrete- and reparameterization for continuous- RVs, and (2) multi-sample KL(p,q) minimization using RWS; they find VIMCO performs better. Corro and Titov (2019) develop a perturb-and-parse algorithm that generalizes the Gumbel-softmax trick to projective dependency trees; they use this trick to define a **semi-supervised VAE** for dependency parsing. Zhang (2020) enhances variational bayesian phylogenetic inference (VBPI) by introducing permutation equivariant **normalizing flows** for the branch length distributions.

MCMC on trees

Alfaro et al. (2003) compare Bayesian MCMC with maximum-likelihood- and maximum-parsimony-bootstrap; they find MCMC is slightly better than ML-bootstrap, which is much better than MP-bootstrap. Dinh et al. (2017) adapt HMC to a non-euclidean space composed by gluing together many locally euclidean simplices; they adapt this variant of HMC to infer tree structures. Palacios et al. (2019) perform MCMC over a space of "Tajima trees" which are quotiented versions of Kingman's coalescent; while the hypothesis space is smaller, it remains to be seen whether this allows scaling to thousands of observations. Yuan et al. (2015) (BitPhylogeny) use MCMC to infer cancer evolution in a single host by solving a joint problem of clustering noisy samples and inferring a phylogeny under a tree-structured stick breaking process model; they analyze both bulk tissue and single cell data; their algorithm scale to ~100 samples and trees with 10s of nodes.
...many others...

Phylogeography / biogeography

Notohara (1990) introduces the structured coalescent model. Ree et al. (2005) introduce a phylogeographic model where location is treated as a discrete variable (as a character) and unobserved locations marginalized out in a continuous Markov process; the limiting cost was the matrix exponential for large numbers of states. Landis et al. (2013) reduce the cost of Ree et al. (2005) by treating location as an auxiliary variable (at cost of mixing speed).
TODO Lemey et al. (2014) ?
Vaughan et al. (2014) describe new MCMC kernels for the structured coalescent, as implemented in BEAST2; they apply this to H3N2 migration; they echo the advice of Ewing et al. (2004) that phylogeographic inference requires informative priors; their Supplementary Material clearly explains the differences between Kingman's coalescent and the structured coalescent. De Maio et al. (2015) compare three phylogeographic inference algorithms: multi-type tree (MTT) is an auxiliary variable MCMC method; discrete-trait analysis marginalizes our latent histories, treating migration between discrete locations ("demes") as if it were mutation; and a new method BASTA that approximates the MTT model; they show BASTA and MTT largely agree, whereas DTA is subject to sampling bias (it assumes genetic samples are observed randomly in the population, which is strongly violated in viral sequencing). Kühnert et al. (2016) define a **multitype birth death process** with MCMC inference, similar to a compartmental model with multiple regions and migration and/or transmission across regions; types can be interpreted as either geographic demes or other partitions like risk group; they claim that **discrete type analysis** (DTA) models do not account for heterogeneous population size among demes, and that **structured coalescent** approaches do not perform well in the early outbreak regime of stochastic exponential growth. Müller et al. (2016) distinguish **mugration** approaches (that assume uniform sampling and permit collapsed inference as with mutation-models) from **structured coalescent** approaches (that allow non-uniform sampling but for which inference is more challenging); they implement an exact (but

slow) inference method for the structured coalescent and show that popular approximate approaches can yield qualitatively wrong inferences. Volz and Siveroni (2018) demonstrate phylodynamic inference of epidemiological parameters in Ebola and Influenza A, using BEAST2. Baele et al. (2018) survey software for exact and approximate phylodynamics with applications to virology and epidemiology. Lundgren and Ralph (2019) compare resistance-based and coalescent-based models of biogeography; they show that the two classes are equivalent if the hitting-time matrix is symmetric (a stronger condition than symmetric dynamics), but that in non-symmetric settings resistance-based models can be inaccurate. Lemey et al. (2020) analyze COVID-19 spread by sampling phylogenies of 284 genomes from 28 countries, augmented with recent travel history of each individual's genome; they find that the analysis is made difficult due to highly biased sampling rates (e.g. much higher in UK than in China). Deshwar et al. (2015) describe their method PhyloWGS for constructing small phylogenies (<=5 subpopulations) from bulk WGS samples; they account for simple somatic mutations (SSMs), copy number variations (CNVs), structural variations (SVs), interactions among those three, and the challenging statistical inference of mixture components under bulk sequencing.

A common issue among phylogeographic inference algorithms is how to integrate over migration histories. Common approaches are: MCMC and variable elimination (for a small number of demes or Brownian motion).

Alignment free phylogeny inference

Zielezinski et al (2019) compare many alignment-free sequence comparison methods (with no focus on phylogeny inference).

Sampling uniform random spanning trees

Schild (2017) proves that near-linear-time samplers exist. Harvey & Xu (2016) describe a practical algorithm that samples in matrix multiply time $O(n^{2.38})$ and could probably be implemented in PyTorch. Pyro's SpanningTree distribution implements an $O(n^3)$ algorithm on top of PyTorch, which parallelizes to $O(n^2 \log(n))$ parallel time with perfect efficiency. Paulus et al. (2020) generalize the Gumbel-softmax trick to a variety of structured discrete models including undirected spanning tree and rooted directed spanning tree; these permit incorporation in VAEs as in Corro and Titov (2019).

Sampling random matchings

Conditioned on times and genetic sequences of internal nodes, the phylogeny problem reduces to a random 2-matching, for which there is much literature mostly focusing on the related problem of random 1-matchings i.e. permutations. Huang & Jebra (2009) provide a fast Bethe approximation of the matrix permanent (i.e. the partition function of the random perfect matching problem); while the exact partition function is #P-complete, their algorithm scales as $O(n^2)$ per iteration and converges in an empirically constant number of iterations (about 42 iterations to tolerance of 1e-10 on random matrices); it should be easy to adapt this algorithm to the random 2-matching problem. Vontobel (2012) further explores the Bethe permanent, showing that it is a lower bound of the true permanent (hence would provide a true ELBO in VI) and that it should converge quickly. Chen (2018) applies the permanent approximations of Roos (2018) to multiple target tracking; these involve both first and second-order approximations with known error bounds. Volkovs & Zemel (2012) define an

efficient MCMC sampler for bipartite matchings; this might be adapted to sample 2-matchings, perhaps initialized to a MAP matching computed via max-product BP (Bayati et al. 2011), (Huang & Jebra 2011).

Complexity results

Computational complexity:
Sebastien Roch (2005) gives a short proof that maximum likelihood tree estimation is NP-hard and that ML is even hard to approximate within a constant factor; however the latter claim seems weak since we instead generally want to approximate to within a constant log-likelihood shift *per datum*.

Sample efficiency: how many characters (fix n, vary p)?
Mossel (2003), Daskalakis et al. (2006), and Daskalakis et al. (2011) identify a phase transition of phylogenetic identifiability: when mutation rate is too high, identification requires polynomial(n)-many characters at each leaf (variant sites), whereas mutation rate is sufficiently low, identification requires only O(log(n))-many characters; however in virus tracking we are often limited by genome size and too-low mutation rate (relative to speed of spreading). Roch and Sly (2018) tighten those bounds; they also define a combinatorial distance metric between trees and use it to prove bounds on leaf character distributions. (Roch and Wang 2017) and Fan and Roch (2018) characterize the difficulty of root reconstruction; a critical point is known as the Kesten-Stigum threshold.

Sample efficiency: how many taxa (vary n, fix p)?
Zwickl (2002) and Heath et al. (2008) basically argue that "more data leads to better inferences", in particular, increasing the number of taxa = samples = leaves n reduces errors; they classify types of errors. Guten et al. (2007) and Susko and Roger (2012) refine these observations by asking "which taxa would improve inference" in the framework of **experimental design**, with objective functions including Fisher information and probability-of-correct-ML-estimate.

Inference in combinatorial spaces

Bouchard-Cote and Jordan (2010) develop variational inference algorithms (BP, MF, tree-reweighted) for approximate inference in a wide class of combinatorial models with linear binary potentials and local tractable hard constraints on sets thereof; they apply this framework to bipartite matchings and multiple sequence alignment. Tarlow et al (2012) develop fast exact inference algorithms for binary combinatorial models with arbitrary cardinality potentials. Djolonga, Jegelka, and Kraus (2018) develop variational inference algorithms with provable upper bounds on an inclusive Renyi divergence, for a wide class of combinatorial models obeying a submodularity property; however their bounds . Kuleza and Taskar (2012) apply determinantal point processes to machine learning problems, covering a wide class of models that include negative local potentials including some soft-constrained combinatorial problems.

Hyperbolic VAEs and Tree-VAEs

Vikram, Hoffman, & Johnson (2018) define a LORACs prior and **subsampling** strategy for undirected tree-structured data, whereby they learn an "inducing point tree" (200-2000 nodes) conditioned on which real data is independent; this could be adapted to large-scale phylogeny inference by learning a "latent induced

phylogeny" conditioned on which all observations are independent; their geometry is euclidean. Nickel & Kiela (2018) propose to use **Lorentz coordinates** of hyperbolic space for learning hierarchical structured data; they provide computations for exponential maps and use these for Riemannian gradient updates. Nagano et al. (2019) define an exponential-map **transported Gaussian** distribution on hyperbolic space, and provide algorithms for reparametrized sampling and log_density computations (including log-abs-det-jacobian of the exponential map), which could be implemented as a Transform. Mathieu et al. (2019) propose Poncare VAEs whose latent space has **hyperbolic** geometry and is thus suitable as a continuous representation of hierarchical structures; this would be an appropriate geometry for variational tree posteriors based on embeddings; they develop a hyperbolic decoder that can serve as a template for a neural net layer that *inputs* points in hyperbolic space. Bose et al. (2020) develop **normalizing flows** on hyperbolic space, which could serve as more expressive variational posteriors than the warped Gaussians of Mathieu et al. (2019).

Covid-specific data characteristics

Skums et al. (2020) observe that (as of March 2020) Sars-CoV-2 samples exhibited few deviations from a **perfect phylogeny** (violations of the 4-gamete rule); they handle these cases specially and then construct Camin-Sokal phylogenies, which allow each mutation to occur independently at most twice.

Misc

Fourment and Darling (2019) implement a complex model of many phylogenetic parameters *conditioned on a fixed tree*, and compare different continuous inference strategies including VI in Stan and existing MCMC methods; they cite Zhang and Matsen (2019) as interesting future work. Gavryushkin and Drummond (2016) define a number of **distance metrics** on trees, and show they result in different mean (consensus, summary) trees (see Roch and Sly (2018) for another metric). David Duvenaud's course surveys methods of learning discrete structures. De Maio et al. (2018) leverage within-host genetic variants of viruses to improve the accuracy of transmission inference; this assumes a weak bottleneck so that multiple strains can simultaneously transmit across hosts; they use simulations on an Ebola outbreak. Minka (2004) describes the **Dirichlet-tree distribution** which serves as a conjugate prior to observations of tree-structured categoricals. El-Kebir et al. (2015) pose the perfect phylogeny reconstruction problem as the optimization of a certain binary matrix. Qi, Pradhan, and El-Kebir (2019) provide a number of complexity results and an approximation algorithm for phylogenetic deconvolution of bulk cancer samples. He et al (2019) define VAEs with learned structure among latent variables.

TODO relaxed perturb-and-MAP https://arxiv.org/abs/2001.04437

References

- Vu Dinh, Arman Bilge, Cheng Zhang, Frederick A. Matsen IV (2017)
  Probabilistic Path Hamiltonian Monte Carlo (site | pdf | supplement)
- Bernard Fichet (1998)
  The Lp-product of ultrametric spaces and the corresponding product of hierarchies

- MATHIEU FOURMENT, AARON E. DARLING (2019)
  EVALUATING PROBABILISTIC PROGRAMMING AND FAST VARIATIONAL BAYESIAN INFERENCE IN PHYLOGENETICS (abs)
- Nicholas J. A. Harvey, Keyulu Xu (2016)
  Generating Random Spanning Trees via Fast Matrix Multiplication (pdf)
- Jiawei He, Yu Gong, Joseph Marino, Greg Mori, Andreas M. Lehrmann (2019)
  Variational autoencoders with jointly optimized latent dependency structure (pdf)
- Patrick Hoscheit, Oliver G. Pybus (2019)
  The multifurcating skyline plot (abs)
- Nicola De Maio, Colin J. Worby, Daniel J. Wilson, Nicole Stoesser (2018)
  Bayesian reconstruction of transmission within outbreaks using genomic variants (abs)
- Fredrik Ronquist, Jan Kudlicka, Viktor Senderov, Johannes Borgström, Nicolas Lartillot, Daniel Lundén, Lawrence Murray, Thomas B. Schön, David Broman (2020)
  Probabilistic programming: a powerful new approach to statistical phylogenetics (abs)
- Julia A. Palacios, Amandine Véber, Lorenzo Cappello, Zhangyuan Wang, John Wakeley and Sohini Ramachandran (2019)
  Bayesian Estimation of Population Size Changes by Sampling Tajima's Trees (abs)
- N Saitou, M Nei (1987)
  The neighbor-joining method: a new method for reconstructing phylogenetic trees. (pdf)
- Miriam Shiffman, William T. Stephenson, Geoffrey Schiebinger, Jonathan Huggins, Trevor Campbell, Aviv Regev, Tamara Broderick (2018)
  Reconstructing probabilistic trees of cellular differentiation from single-cell RNA-seq data (abs)
- Aaron Schild (2017)
  An almost-linear time algorithm for uniform random spanning tree generation (abs)
- Sharad Vikram, Matthew D. Hoffman, Matthew J. Johnson (2018)
  The LORACs prior for VAEs: Letting the Trees Speak for the Data (abs)
- Tasfia Zahin, Md. Hasin Abrar, Mizanur Rahman, Tahrina Tasnim, Md. Shamsuzzoha Bayzid, Atif Rahman (2019)
  An Alignment-free Method for Phylogeny Estimation using Maximum Likelihood (pdf)
- Cheng Zhang, Frederick A. Matsen IV (2019)
  VARIATIONAL BAYESIAN PHYLOGENETIC INFERENCE (pdf, reviews)
- Andrzej Zielezinski, Hani Z. Girgis, Guillaume Bernard, Chris-Andre Leimeister, Kujin Tang, Thomas Dencker, Anna Katharina Lau, Sophie Röhling, Jae Jin Choi, Michael S. Waterman, Matteo Comin, Sung-Hou Kim, Susana Vinga, Jonas S. Almeida, Cheong Xin Chan, Benjamin T. James, Fengzhu Sun, Burkhard Morgenstern, Wojciech M. Karlowski (2019)
  Benchmarking of alignment-free sequence comparison methods (abs)