Ben Teo: Thesis Research

Context: Phylogenetics studies the evolutionary relationships among a set of species. Methods are developed to reconstruct their genealogy (*phylogeny*) from a common ancestor, and to fit evolutionary models that explain their observed traits in light of this shared history. The phylogeny is represented as a directed acyclic graph (DAG), whose nodes represent either present-day or ancestral species, and whose edges indicate descent. The phylogeny can be a tree, where each node other than the root (the common ancestor of all nodes in the graph) has one parent node, or more generally a network, where some nodes (generally called 'reticulate' or 'hybrid' nodes) have multiple parents to represent hybridization, migration and subsequent gene flow between populations, recombination between virus strains, etc. Traits can be discrete or continuous (e.g. DNA sequences, flower color, body size, gene expression level, strain virulence) and are modeled as evolving from the root state down the phylogeny according to some stochastic process (e.g. Brownian motion, Ornstein-Uhlenbeck process). Models that accurately capture their evolution are important both for reconstructing ancestral states and predicting potential trait trajectories, and should be evaluated on larger, more realistic phylogenies, that often contain many reticulations. Yet existing methods to fit such models do not scale well as the phylogeny grows in size (number of nodes) and complexity (number of hybrid nodes).

Research focus: Statistical and computational techniques to efficiently infer the parameters of models for the evolution of continuous traits on large, complex phylogenetic networks.

- **Previous work (Section 1):** Extended existing continuous trait models for phylogenetic networks to account for *intraspecific* (within-species) variation.
- **Current work (Section 2-3):** (1) Reframe inference for continuous trait models in terms of *belief propagation* (BP) and *loopy belief propagation* (loopy BP) from the Graphical Models literature. (2) Build a Julia package to fit Gaussian trait models on phylogenetic networks using BP and loopy BP.
- **Proposed work (Section 4):** Specialize BP and loopy BP to fit Gaussian trait models on phylogenetic networks using exact and approximate likelihood gradient computations.

Impact: Methods that scale well with the complexity of the phylogeny *will address a critical bottleneck* in the current confluence of (1) advances in methods to reconstruct complex phylogenies [27, 14] and (2) the growing potential for large-scale comparative studies in the literature [11, 26]. Further, efforts to develop these methods *complement recent efforts* to adapt dimension-reduction techniques (e.g. latent factor models, structural equation models) to high-dimensional continuous trait models on phylogenetic trees [7, 25]. The integration of both these components (methods that manage network complexity and trait dimension) will *significantly advance model selection for the evolution of continuous traits.*

1 Intraspecific variation for Gaussian trait models on a network

Context: Regression methods have been adapted to test hypotheses on trait evolution by modeling a response trait as a linear function of predictor traits among a set of species, and residual variation by an evolutionary model on their phylogeny to capture correlated deviations. The trait values for each species are typically sample aggregates that are influenced by intraspecific variation. Failing to account for such variation adversely affects parameter estimation and model selection for these methods on phylogenetic trees [20, 3]. However, its *impact given a phylogenetic network had not been documented.* In fact, there was *no available regression method implementation for phylogenetic networks* that could model the variable effects of intraspecific variation for different species (e.g. some species can be less densely sampled than others).

Approach: We extended existing regression methods that account for intraspecific variation in the response trait along a tree, to networks [24]. Our method assumes equal intraspecific variation in the response trait across species, and jointly estimates the regression coefficients for predictor traits with interspecific and intraspecific variation in the response trait. It is now implemented in the `PhyloNetworks` Julia package [21]. We proved that the computational complexity of our method scales with the number of species instead of the number of individuals, so that either individual-level or species-level data can be supplied. We ran simulations to demonstrate our method's accuracy and its robustness to assumption violations, and applied it to study leaflet size evolution in *Polemonium*, whose history was shown to involve hybridization [19].

2 Exact and approximate methods to fit trait models

Context: Trait evolution is typically modeled by a Markov process that evolves down the phylogeny and induces a joint distribution over its nodes. The joint distribution can be expressed as a product of factors, one over each node family (a child node and its parent(s)). For example, the joint distribution over a tree can be written as a product of edge factors. To fit trait models, we often need to compute the *data likelihood* by evaluating the joint distribution at the trait values observed at the tips (leaves) and marginalizing out unobserved node states. Several *pruning algorithms* have been invented to efficiently compute this quantity in a postorder traversal (of branches, hence the term "pruning") of a tree [4, 8, 16]. These algorithms essentially rewrite the multiple integral as a nested sequence of smaller integrals. Unfortunately, *pruning cannot be applied to networks*, and *alternative algorithms to efficiently compute the data likelihood on a network are lacking*. For example, current implementations that fit Gaussian trait models on a network build the covariance matrix for all nodes and invert a submatrix for the tip nodes each time the data likelihood is evaluated for different parameter values [1]. This becomes *increasingly intractable* as the network size grows.

Approach: Algorithms to efficiently marginalize a joint distribution that factors over a directed acylic graph (DAG) have been extensively studied in the Graphical Models literature. A graph data structure called a *clique tree* (also known as *tree decomposition*, *join tree*, *junction tree*) can be constructed from the phylogeny and trait model, and the data likelihood can be computed in a postorder traversal of the clique tree under a message-passing scheme called belief propagation (BP) [12].

We cast the various pruning algorithms as specialized instances of BP, and demonstrate the usefulness of this framework for fitting complex evolutionary models that induce a more complex graph representation than the original phylogeny. For example, models that account for *incomplete lineage sorting* (loosely speaking, a specific mechanism by which the observed traits and trait model can appear discordant with an accurate phylogeny) cannot be described by a Markov process for the trait value on the original phylogeny [18], but may be reformulated as a Markov process over a larger state space on a larger DAG that contains the original phylogeny. BP can be run on a clique tree for this larger DAG (typically a network) to compute the data likelihood. Our goal is to alert or remind the phylogenetics community of the versatility and applicability of these tools to phylogenetic networks, so that efforts can be directed towards adapting or extending them instead of reinventing the wheel.

While BP computes the exact likelihood, loopy belief propagation (loopy BP), which applies the same message-passing scheme on a different graph data structure called a *cluster graph*, approximates the likelihood but scales better with the complexity of the phylogeny [12]. Structural features of the phylogeny are informative about the computational gains of loopy BP over BP, and a survey of real and simulated phylogenies with attention to these features suggests that these gains will grow as phylogenies grow in complexity and comparative studies grow in scale. Thus, we assert the importance of investigating loopy BP in the phylogenetics context. To start, we highlight and propose solutions to pertinent initialization details (e.g. initial messages, message schedule) that are unique to Gaussian trait models on phylogenetic networks.

3 PhyloGaussianBeliefProp.jl: Fit Gaussian trait models on phylogenetic networks using Belief Propagation

Context: Biologists who work with *continuous trait data over huge sets of species* will need implementations of *more scalable approaches to fit Gaussian trait models* on these datasets, using a phylogenetic network.

Approach: We build a Julia package to fit Gaussian trait models on a phylogenetic tree or network using BP or loopy BP. The package allows for various branch length transformations (e.g. Pagel's λ or δ, early-burst, accelerating-decelerating [6]) on the phylogeny, and parameter changes across the phylogeny (e.g. added shifts in trait value at hybrid nodes [1]) to improve model fit. Parameter optimization proceeds via either likelihood maximization using BP or factored-energy minimization using loopy BP. As a byproduct of the message-passing framework, ancestral states are simultaneously and efficiently reconstructed. For loopy BP, we implement several cluster graph options (e.g. *factor graph*, *layered trees running intersection property*, *join-graph structuring*) for handling phylogenetic networks of varying complexity [15, 22]. We build in Julia to capitalize on its rich scientific computing ecosystem, especially for future extensions (e.g. `Turing` for Bayesian analysis [5]), which is a strength that current standalone approaches (e.g. `BEAST` [23]) do not share.

I intend to present Sections 2-3 at the Evolution 2024 conference to publicize these methods to biologists and establish potential methodological and empirical collaborations.

4 Gradient propagation and cluster graphs for phylogenetics

Context: For biologists who want to do *model selection for the evolution of continuous traits on large, complex phylogenetic networks*, BP and loopy BP provide a means to manage computational costs at the exploratory phase (Section 2). Parameter optimization and model selection can be made even more efficient if the gradient (with respect to the model parameters) of the data likelihood can be computed, especially in the Bayesian setting where more efficient schemes to sample from the posterior distribution (e.g. *Hamiltonian Monte Carlo*) require such gradients [17]. Conveniently, BP can be adapted to compute the likelihood gradient on a general DAG using a clique tree [12], while loopy BP can been adapted to approximate the likelihood gradient using a cluster graph [13]. However, *neither BP nor loopy BP has been specialized to compute gradients on a phylogenetic network*. Leveraging such methods to compute gradients on networks is a *natural continuation of recent work to optimize gradient computations on trees* [9, 2].

The factor graph has remained the main default for general implementations due to its convenient construction and initialization, and good empirical performance in most (exclusively non-phylogenetic) applications. Yet, preliminary simulations show that the *factor graph generally performs poorly on more complicated phylogenetic networks (ancestral recombination graphs) from the recent literature*. An added complication is that the approximation error and rate of convergence for loopy BP depend on the cluster graph and message schedule used, and this dependence has been less well quantified when node states (e.g. species traits) are multivariate and alternative cluster graphs are used.

Approach: *Adapting gradient techniques to networks:* I will study BP and loopy BP innovations to compute or approximate the likelihood gradient for Gaussian models on a general DAG, and compare it with recent developments in phylogenetics for efficient gradient computation on trees [2]. I will extend `PhyloGaussianBeliefProp` (Section 3) to incorporate gradient methods, and benchmark performance differences (e.g. accuracy, runtime, memory usage) against existing methods.

Adapting cluster graphs and message schedules to networks for loopy BP: Phylogenetic networks have special and well-studied characteristics, which I hypothesize can be used to guide cluster graph construction. For example, each network can be divided into components known as *blocks*, and cluster graphs for separate blocks can be connected to form a cluster graph for the whole network, so that different cluster graph options can be combined. I will conduct a thorough simulation study using `PhyloGaussianBeliefProp` to study how cluster graphs and message schedules can be tailored to phylogenetic networks.

- The recent `SiPhyNetwork` package can simulate phylogenetic networks under realistic evolutionary processes, with various degrees of structural complexity [10]. `PhyloNetworks` can then simulate trait data on these networks using multivariate Gaussian trait models [21].
- `PhyloGaussianBeliefProp` (in development) can be extended to allow *local* (to a block) cluster graphs, which may use different constructions, to be pieced together.
- Benchmarking the performance (over a range of evolutionary settings) of different cluster graph constructions against one another and a clique tree (BP) will help with developing adaptive schemes for cluster graph construction and message scheduling, tuning heuristics to decide between BP and loopy BP, and building intuition in efforts to prove theoretical convergence results.

References

[1] Bastide, P., C. Solís-Lemus, R. Kriebel, K. William Sparks, and C. Ané, Phylogenetic Comparative Methods on Phylogenetic Networks with Reticulations, *Systematic Biology*, *67*(5), 800–820, doi:10.1093/sysbio/syy033, 2018.

[2] Bastide, P., L. S. T. Ho, G. Baele, P. Lemey, and M. A. Suchard, Efficient Bayesian inference of general Gaussian models on large phylogenetic trees, *The Annals of Applied Statistics*, *15*(2), 971–997, doi:10.1214/20-AOAS1419, 2021.

[3] Cooper, N., G. H. Thomas, C. Venditti, A. Meade, and R. P. Freckleton, A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies, *Biological Journal of the Linnean Society*, *118*(1), 64–77, doi:10.1111/bij.12701, 2016.

[4] Felsenstein, J., Maximum-likelihood estimation of evolutionary trees from continuous characters., *American journal of human genetics*, *25*(5), 471, 1973.

[5] Ge, H., K. Xu, and Z. Ghahramani, Turing: a language for flexible probabilistic inference, in *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pp. 1682–1690, 2018.

[6] Harmon, L. J., *Phylogenetic comparative methods*, EcoEvoRxiv, doi:10.32942/osf.io/e3xnr, 2019.

[7] Hassler, G. W., B. Gallone, L. Aristide, W. L. Allen, M. R. Tolkoff, A. J. Holbrook, G. Baele, P. Lemey, and M. A. Suchard, Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis, *Methods in Ecology and Evolution*, *13*(10), 2181–2197, doi:10.1111/2041-210X.13920, 2022.

[8] Ho, L. S. T., and C. Ané, A linear-time algorithm for Gaussian and non-Gaussian trait evolution models, *Systematic Biology*, *63*(3), 397–408, doi:10.1093/sysbio/syu005, 2014.

[9] Ji, X., Z. Zhang, A. Holbrook, A. Nishimura, G. Baele, A. Rambaut, P. Lemey, and M. A. Suchard, Gradients do grow on trees: a linear-time O (N)-dimensional gradient for statistical phylogenetics, *Molecular biology and evolution*, *37*(10), 3047–3060, doi:10.1093/molbev/msaa130, 2020.

[10] Justison, J. A., C. Solis-Lemus, and T. A. Heath, SiPhyNetwork: An R package for simulating phylogenetic networks, *Methods in Ecology and Evolution*, *14*(7), 1687–1698, doi:10.1111/2041-210X.14116, 2023.

[11] Kattge, J., et al., TRY plant trait database–enhanced coverage and open access, *Global change biology*, *26*(1), 119–188, doi:10.1111/gcb.14904, 2020.

[12] Koller, D., and N. Friedman, *Probabilistic graphical models: principles and techniques*, MIT Press, 2009.

[13] Lu, Y., Z. Liu, and B. Huang, Block belief propagation for parameter learning in markov random fields, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4448–4455, doi:10.1609/aaai.v33i01.33014448, 2019.

[14] Maier, R., P. Flegontov, O. Flegontova, U. Isildak, P. Changmai, and D. Reich, On the limits of fitting complex models of population history to f-statistics, *Elife*, *12*, e85,492, doi:10.7554/eLife.85492, 2023.

[15] Mateescu, R., K. Kask, V. Gogate, and R. Dechter, Join-graph propagation algorithms, *Journal of Artificial Intelligence Research*, *37*, 279–328, doi:10.1613/jair.2842, 2010.

[16] Mitov, V., K. Bartoszek, G. Asimomitis, and T. Stadler, Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts, *Theoretical Population Biology*, *131*, 66–78, doi:10.1016/j.tpb.2019.11.005, 2020.

[17] Neal, R. M., MCMC using Hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X. Meng, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, 2011.

[18] Rabier, C.-E., V. Berry, M. Stoltz, J. D. Santos, W. Wang, J.-C. Glaszmann, F. Pardi, and C. Scornavacca, On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo, *PLOS Computational Biology*, *17*(9), 1–39, doi:10.1371/journal.pcbi.1008380, 2021.

[19] Rose, J. P., C. A. P. Toledo, E. M. Lemmon, A. R. Lemmon, and K. J. Sytsma, Out of sight, out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal, *Systematic Biology*, *70*(1), 162–180, doi:10.1093/sysbio/syaa049, 2021.

[20] Silvestro, D., A. Kostikova, G. Litsios, P. B. Pearman, and N. Salamin, Measurement errors should always be incorporated in phylogenetic comparative analysis, *Methods in Ecology and Evolution*, *6*(3), 340–346, doi:10.1111/2041-210X.12337, 2015.

[21] Solís-Lemus, C., P. Bastide, and C. Ané, Phylonetworks: a package for phylogenetic networks, *Molecular Biology and Evolution*, *34*(12), 3292–3298, doi:10.1093/molbev/msx235, 2017.

[22] Streicher, S., and J. du Preez, Graph coloring: Comparing cluster graphs to factor graphs, in *Proceedings of the ACM Multimedia 2017 Workshop on South African Academic Participation*, pp. 35–42, doi:10.1145/3132711.3132717, 2017.

[23] Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using beast 1.10, *Virus evolution*, *4*(1), vey016, doi:10.1093/ve/vey016, 2018.

[24] Teo, B., J. Rose, P. Bastide, and C. Ané, Accounting for Within-Species Variation in Continuous Trait Evolution on a Phylogenetic Network, *Bulletin of the Society of Systematic Biologists*, *2*(3), 1–29, doi:10.18061/bssb.v2i3.8977, 2023.

[25] Thorson, J. T., et al., Identifying direct and indirect associations among traits by merging phylogenetic comparative methods and structural equation models, *Methods in Ecology and Evolution*, *14*(5), 1259–1275, doi:10.1111/2041-210X.14076, 2023.

[26] Tobias, J. A., et al., AVONET: morphological, ecological and geographical data for all birds, *Ecology Letters*, *25*(3), 581–597, doi:10.1111/ele.13898, 2022.

[27] Vaughan, T. G., D. Welch, A. J. Drummond, P. J. Biggs, T. George, and N. P. French, Inferring ancestral recombination graphs from bacterial genomic data, *Genetics*, *205*(2), 857–870, doi:10.1534/genetics.116.193425, 2017.