Research Statement

Xiang Ji, Department of Mathematics, Tulane University

My research focuses on every component of statistical phylogenetics, from model development and advanced inference techniques to under-the-hood parallel computation libraries. My efforts have one central goal: solving biological questions. Advances in genome sequencing technology are generating genetic data at an ever-increasing pace. This burst of data provides opportunities to look at the underlying biological processes that generate evolutionary patterns. However, these opportunities are accompanied by both statistical and computational challenges that I combat with a mix of theory and practical implementations.

Theoretical phylogenetic method development

Scalable phylogenetic gradients. Likelihood evaluation is usually considered the computational bottleneck in phylogenetic studies. Even worse is calculating the likelihood gradient for parameter inference. Several groups have recognized that replacing one transition probability matrix with its differential and then completing a post-order traversal yields the derivative with respect to (w.r.t.) a single branch. Popular phylogenetic maximum likelihood estimation (MLE) software such as GARLI and RAxML (arguably the fastest available) employ this idea for local optimization. However, in this manner, an analytic gradient for optimization w.r.t. all branches requires $O(N^2)$ operations where N is the number of sequences. In Ji et al. (2020), we complement the post-order Felsenstein's pruning algorithm with its pre-order traversal that calculates the gradient w.r.t. all branches in $O(N)$. We find dramatic computational improvement (see Table 1).

Example	No. Rates	Speed-up
West Nile	206	126 ×
Lassa	420	168 ×
Dengue	702	234 ×

Table 1: MLE inference speedups per BFGS optimization iteration

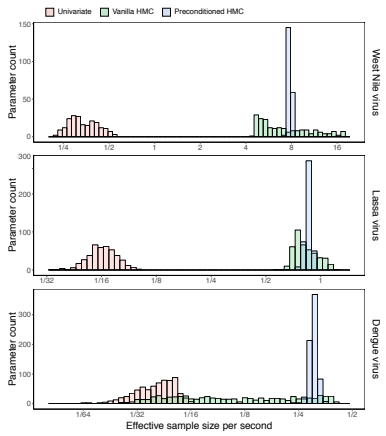

Figure 1: Posterior sampling efficiency on all branch-specific evolutionary rate We bin parameters by their ESS/s values.

Phylogenetic Hamiltonian Monte Carlo (HMC). HMC is an advanced Markov chain Monte Carlo method that employs deterministic dynamics to intelligently generate high-dimensional proposal states, after which a Metropolis accept-reject step with usually high acceptance rates ensures convergence to a target distribution of interest. HMC promises scalability, but only with inexpensive evaluations of the gradient. Our results on inferring the branch-specific evolutionary rates demonstrate that HMC outperforms the univariate Metropolis-Hastings transition kernels as employed in current mainstream software (Figure 1).

Efficient divergence time estimation through node height to ratio transform. To tackle divergence time estimation with HMC, we developed a reparameterization that transforms all internal node heights into a series of independent ratios bounded by $[0, 1]$ (Ji et al., 2023). The parameterization works for both concurrent and serially sampled data. Our method both resolves a mixing issue in the West Nile virus example (Figure 2) and improves inference efficiency by at least 5-fold for the Lassa and rabies virus examples.

Relaxed random walk models (RRW) at scale. RRW models of trait evolution introduce branch-specific rate multipliers to modulate the variance of a standard Brownian diffusion process along a phylogeny and more accurately model overdispersed biological data. In Fisher et al. (2021), we develop a scalable method that resembles the phylogenetic gradient for CTMCs on diffusion processes. We further extended the work with

shrinkage priors to resemble a random local clock model that is previously intractable for large datasets (Fisher et al., 2023).

Phylo-Hawkes models. Self-exciting spatiotemporal Hawkes processes have found increasing use in the study of large-scale public health threats. In Holbrook et al. (2022a), we developed a flexible Hawkes model that incorporates different levels of spatial uncertainty of sampling locations. In Holbrook et al. (2022b), we employed a Hawkes process to infer viral contagiousness in a Bayesian analysis of 23,421 viral cases from the 2014 to 2016 Ebola outbreak in West Africa, of which only a subset of 1610 samples have genomic information (Figure 4).

Gene conversion in multigene family evolution. Interlocus gene conversion (IGC) homogenizes repeats and induces evolutionary dependence between sequence positions. While genomes can be repeat-rich, the evolutionary importance of IGC is poorly understood, largely because of a lack of statistical tools. In Ji et al. (2016), we showed how to quantify IGC via a one-parameter extension to any existing substitution model. The key idea is to jointly treat corresponding positions in different paralogs so that codon (or nucleotide) substitutions originating with both point mutation and IGC could be considered. We evaluated the approach with 14 data sets of yeast ribosomal protein genes and found the percentage of codon substitutions that originate with IGC rather than point mutation to range from 20% to 38%. Conventionally, IGC is ignored and these substitutions would be misattributed to point mutation. Our subsequent unpublished work on duplicated protein-coding mammalian genes has estimated that slightly more than 10% of codon substitutions originate with IGC.

In Yang et al. (2023), we improved our 2016 model to assess how much paralog homogenization can be attributed to IGC mutation versus correlated selective pressure in paralogous genes. In that study, we detected substantial IGC following an ancient whole genome duplication in teleosts. Given that conventional treatments of molecular evolution ignore IGC and that repetitive DNA constitutes a high proportion of so many genomes, we are extending our IGC research in multiple directions. We have ongoing efforts to examine IGC in segmentally-duplicated primate genes. We have extended our 2016 model with a composite likelihood procedure that infers IGC tract length to model spatial correlations between sites and are investigating how IGC mutation rates are influenced by paralog divergence. My python-based open-source software is freely available on GitHub https://github.com/xji3/IGCexpansion.

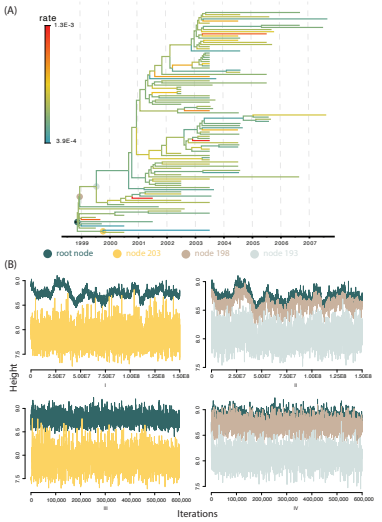

Figure 2: Trace plot of four height parameters indicated on a WNV phylogeny. I. and II. are trace plots with univariable samplers, III. and IV. are trace plots with an HMC sampler for equal runtime.

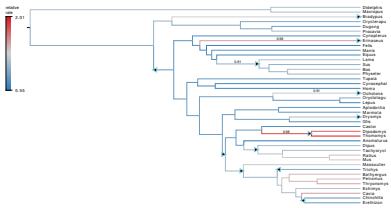

Figure 3: Maximum clade credibility tree under shrinkage-clock of mammalian and rodent radiation. Numbers on a branch indicates the posterior probability of a new clock. For comparison, local clocks of the random local clock (RLC) model are depicted as black triangles.

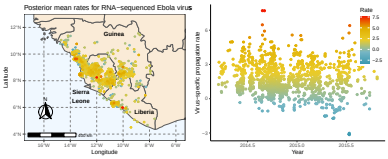

Figure 4: Hawkes model posterior mean rates for the 1367 (of 1610) RNA-sequenced viral samples for which date/location data are available in 2014 to 2016 Ebola outbreak in West Africa.

Parallel numerical implementation

Massive parallelization of pre- and post-order traversals. Parallelization is growing as a dominant theme in large-scale statistical inference with hardware ranging from clusters of independent compute nodes to multithreaded multicore processors to parallel coprocessors such as graphics processing units (GPUs). Capitalizing on these hardware features in software implementation is emerging as the most important task facing a computational statistician. To further advance our linear-time gradient algorithms, we have designed and implemented a double-bock architecture (Figure 5) for pre- and post-order partial updates and their product reductions on GPUs with a cuda implementation for NVIDIA products as well as an OpenCL implementation for AMD and Intel products in the software package BEAGLE (Gangavarapu et al., 2024). At the same time, a vectorized CPU implementation of our linear-order gradient algorithms in (Ji et al., 2020, 2023) using streaming SIMD extensions (SSE) delivers better performance than general-purpose automatic differentiation methods for the application of phylogenetic variational inference (Fourment et al., 2023).

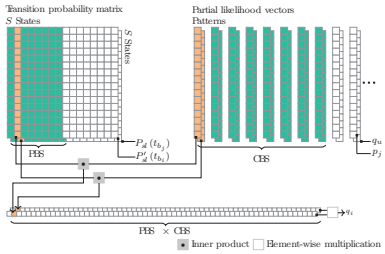

Figure 5: Double block design for pre- and post-order partial update and reductions on GPUs

Massively parallelized spatiotemporal Hawkes model gradient. To make our work on Phylo-Hawkes models computationally tractable, we implemented high-performance parallel implementations of the gradient of the log-likelihood w.r.t. spatial locations of our spatiotemporal Hawkes model (Figure 6) (Holbrook et al., 2022a,b).

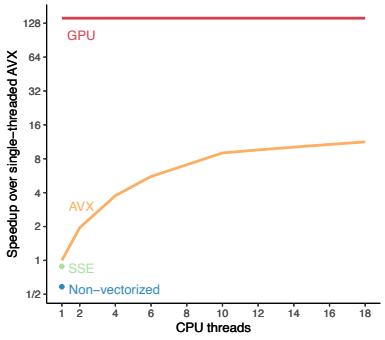

Figure 6: Speedups for the spatiotemporal Hawkes process gradient evaluations.

Collaborative work

Collaborations constitute a big proportion of my research. An applied statistician should never work alone — we advance biology by collaborating with experimental biologists, developing statistical methods tinkered towards specific biological hypothesis and providing software. I am collaborating with virologists interested in learning the evolution and diffusion patterns of various viruses through space and time and to test correlations of factors with key events in their evolution. I am also collaborating with systematists interested in population structure of sub-species and their geographic distribution models. Since joining Tulane University, I have established promising collaborations with cancer biologists using Drosophila model to study cancer evolution.

International consortia fighting infectious diseases. Genomic data furnish one major asset in the fight against infectious diseases. Historical information contained in viral sequences contributes to better insight into viral emergence and early transmission dynamics, even before systematic epidemiological surveillance can initiate. As a developer of the BEAST and BEAGLE development team, I provide scalable state-of-the-art statistical methods to help virologists and public health decision makers combat viral epidemics. For example, MCMC sampling of branch-specific CTMC rate-multipliers drains $>$ 95% of total computation effort while reconstructing the 2020 SARS-CoV-2 resurgence in Europe that benefits from the speedups of our linear-time algorithms and HMC development (Lemey et al., 2021).

Evolutionary biology fighting cancer. I am collaborating with Dr. Wu-Min Deng from the Medical School at Tulane University to provide bioinformatics assistance. Using a Drosophila tumor model, in which onco-

genic Notch drives tumorigenesis in an epithelial transition zone, we have found that tumor progression is driven by a combination of polyploid mitosis, endoreplication, and depolyploidization. These tumors show remarkable levels of DNA double-stranded breaks and chromosome instabilities. In these tumors, copy number variations and polyaneuploidy (hallmarks of lethal cancer) are increased in more advanced tumors that have been transplanted into host flies for multiple generations. We are developing statistical methods that draw information jointly from bulk and single-cell somatic DNA data to learn the evolutionary dynamics of these copy number variations in tumors.

References

Fisher, A. A., X. Ji, A. Nishimura, G. Baele, P. Lemey, and M. A. Suchard (2023). Shrinkage-based random local clocks with scalable inference. *Molecular biology and evolution 40*(11), msad242.

Fisher, A. A., X. Ji, Z. Zhang, P. Lemey, and M. A. Suchard (2021). Relaxed random walks at scale. *Systematic Biology 70*(2), 258–267.

Fourment, M., C. J. Swanepoel, J. G. Galloway, X. Ji, K. Gangavarapu, M. A. Suchard, and F. A. Matsen IV (2023). Automatic differentiation is no panacea for phylogenetic gradient computation. *Genome Biology and Evolution 15*(6), evad099.

Gangavarapu, K., X. Ji, G. Baele, M. Fourment, P. Lemey, F. A. Matsen IV, and M. A. Suchard (2024). Many-core algorithms for high-dimensional gradients on phylogenetic trees. *Bioinformatics 40*(2), btae030.

Holbrook, A. J., X. Ji, and M. A. Suchard (2022a). Bayesian mitigation of spatial coarsening for a hawkes model applied to gunfire, wildfire and viral contagion. *The Annals of Applied Statistics 16*(1), 573–595.

Holbrook, A. J., X. Ji, and M. A. Suchard (2022b). From viral evolution to spatial contagion: a biologically modulated hawkes model. *Bioinformatics 38*(7), 1846–1856.

Ji, X., A. A. Fisher, S. Su, J. L. Thorne, B. Potter, P. Lemey, G. Baele, and M. A. Suchard (2023). Scalable bayesian divergence time estimation with ratio transformations. *Systematic Biology*.

Ji, X., A. Griffing, and J. L. Thorne (2016). A phylogenetic approach finds abundant interlocus gene conversion in yeast. *Molecular Biology and Evolution 33*(9), 2469–2476.

Ji, X., Z. Zhang, A. Holbrook, A. Nishimura, G. Baele, A. Rambaut, P. Lemey, and M. A. Suchard (2020). Gradients do grow on trees: a linear-time O (N)-dimensional gradient for statistical phylogenetics. *Molecular Biology and Evolution 37*(10), 3047–3060.

Lemey, P., N. Ruktanonchai, S. L. Hong, V. Colizza, C. Poletto, F. Van den Broeck, M. S. Gill, X. Ji, A. Levasseur, B. B. Oude Munnink, et al. (2021). Untangling introductions and persistence in covid-19 resurgence in europe. *Nature 595*(7869), 713–717.

Yang, Y., T. Xu, G. Conant, H. Kishino, J. L. Thorne, and X. Ji (2023). Interlocus gene conversion, natural selection, and paralog homogenization. *Molecular Biology and Evolution*.