Research Statement

Xiang Ji*, *University of California, Los Angeles

My research focuses on every component of statistical phylogenetics, from model development, advanced inference technique to under-the-hood parallel computation libraries with one central goal: solving biological questions. Advances in genome sequencing technology are generating genetic data at an ever-increasing pace. The burst of data provides us opportunities to look at the underlying biological processes that generate the evolutionary pattern. However, opportunities come with challenges from both a statistical and computational perspective that merits theoretical thinking and practical implementations.

Theoretical phylogenetic method development

Modeling gene conversion in multigene family evolution. Interlocus gene conversion (IGC) homogenizes repeats. 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 introduced an approach to expand any existing substitution model for quantifying IGC. The key idea was to jointly treat corresponding positions in different paralogs so that codon 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%[a].

In our recent work Ji and Thorne (2019), we extended our 2016 model with a composite likelihood procedure that disentangles IGC tract length and initiation rate. Our estimates from protein-coding data concerning the mean length of fixed IGC tracts were unexpectedly low and are associated with high degrees of uncertainty. In contrast, our estimates from the primate intron data had lengths in the general range expected from IGC mutation studies.

Good theory is pointless without user-friendly software. While our python-based software is freely available[b], we are developing new IGC modules in BEAST[c] to test more biologically interesting hypothesis, e.g. if IGC rate decreases as paralogs diverge, how genomic distance between paralogs affect the IGC rate between them.

Scalable phylogenetic gradients. The likelihood evaluation is usually considered the computational bottleneck in phylogenetic studies. Even worse is its gradient calculation. Several groups have recognized that replacing one transition probability matrix with its differential and 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., 2019), we complement the post-order pruning algorithm with

Example	# taxa	Speed-up
West Nile	104	116 ×
Lassa	212	252 ×
Dengue	353	701 ×

Table 1: **Two orders-of-magnitude improved molecular clock rate MLE via exact scalable gradients.** Preliminary implementation compares L-BFGS optimization using analytic vs numeric gradients. Numeric approximations compute a finite (central) difference for each dimension. Speed-ups are relative per iteration; numeric approximations generally take more iterations. Further implementation remains to compare performance with coordinate-wise optimization popular in phylogenetic software.

[a] Note that these substitutions would be attributed to point mutations if gene conversion is not considered

[b] https://github.com/xji3/IGCexpansion

[c] Bayesian Evolutionary Analysis Sampling Trees (http://beast.community/). Under development on gene conversion branch of BEAST

its pre-order traversal that calculates the gradient w.r.t. all branches in $O(N)$ (Table 1).

Phylogenetic Hamiltonian Monte Carlo (HMC). HMC is an advanced MCMC 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 preliminary results on inferring the branch-specific evolutionary rates demonstrate that HMC outperforms the univariate Metropolis-Hastings transition kernels as employed in current software (Figure 1).

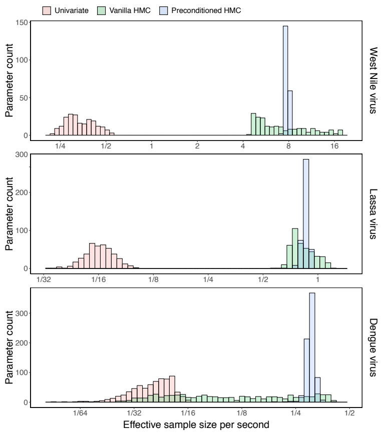

Figure 1: Posterior sampling efficiency on all branch-specific evolutionary rate for the West Nile virus, Lassa virus and Dengue virus examples. We bin parameters by their ESS/s values. The three transition kernels employed in the MCMC are color-coded: a univariate transition kernel, a 'vanilla' HMC transition kernel with an identity mass matrix and a 'preconditioned' HMC transition kernel with an adaptive mass matrix informed by the diagonal elements of the Hessian matrix.

Efficient divergence time estimations through node height to ratio transform. To tackle on divergence time estimation with HMC, one of the central theme of phylogenetics, we developed a reparameterization that transforms all internal node heights into a series of independent ratios bounded by $[0, 1]$. The parameterization works for both concurrent and serially sampled data. Our preliminary results show a 6-fold speed-up with vanilla HMC on the ratio space for Bayesian divergence time estimations on a 17-taxa tree of serially sampled Dengue virus sequences[d].

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. (2019) we develop a scalable method that resembles the phylogenetic gradient for CTMCs on diffusion processes.

Numerical implementation

Massive parallelization of pre- and post-order traversals. Parallelization is growing as a dominant theme in large-scale statistical inference and in hardware includes clusters of independent compute nodes, multithreaded multicore processors and parallel coprocessors such as graphics processing units (GPUs). We have implemented a central processing unit (CPU) version of the above algorithms in the software package BEAGLE[e]. We are implementing a CUDA-based GPU version in BEAGLE.

Higher level programing language for prototyping. While lower-level libraries gain the maximum computational efficiency, a higher-level language provides an easy-to-implement environment for prototyping and serves as an interface to benefit from both worlds. We use python and R for light-weight proof-of-concept projects.

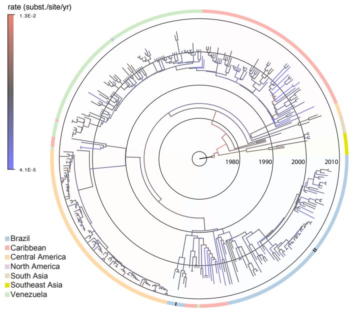

Figure 2: Maximum clade credible tree of the Dengue virus example. The data set consists of 352 sequences of the serotype 3 of the Dengue virus. Branches are color-coded by the posterior means of the branch-specific evolutionary rates according to the color bar on the top left. The concentric circles indicate the time scale with the year numbers.

[d] Under development on hmc-clock branch of BEAST

[e] On hmc-clock branch of BEAGLE library

Collaborative work

Collaborations constitute a big proportion of my research. A computational biologist should never work alone — we serve the biological sciences community by collaborating with empirical biologists, developing statistical methods and providing software. In my recent years, I collaborate with systematicists interested in population structure of sub-species and their geographic distribution models and virologists interested in viral evolutionary histories with the population dynamics and geographic transmission histories.

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 member of the ARTIC network[f], we aim to produce a cheap, mobile virus sequencing system, supported by statistically rigorous analysis frameworks, and information sharing platforms, to prepare for the future outbreaks and ensure that viral genome sequencing is positioned to have full impact on the public health response. I am collaborating with researchers in the Center for Viral Systems Biology[g] on statistical modeling and researchers from Nanjing Agricultural University on phylodynamic and phylogeographic analyses of pig coronaviruses too.

[f] http://artic.network/index.html

ARTICnetwork
Putting genomic surveillance at the heart of viral epidemic response

[g] https://cvisb.org

CV•SB Center for Viral Systems Biology

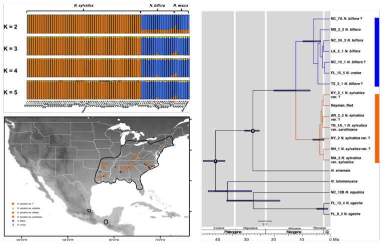

Figure 3: Consulting project with Dr. Jenny Xiang on *Nyssa sylvatica* complex Zhou et al. (2018).

Fun collaborations. I collaborated with Dr. Jenny Xiang at North Carolina State University for a consulting project that tried to resolve taxonomically difficult species relationship and biogeographic history with RAD-seq sequence data (Figure 3).
I am collaborating with Dr. Frederick Matsen and Dr. Mathieu Fourment, who use my node height to ratio transformation and gradient implementation in BEAGLE for their phylogenetic variational inference project. Besides, I am collaborating with Dr. Hirohisa Kishino and his colleague Dr. Yasuyuki Goto at University of Tokyo to study the impact of interlocus gene conversion in shaping the evolution of tandemly repeated antigens by using my software.

References

Fisher, A. A., X. Ji, P. Lemey, and M. A. Suchard (2019). Relaxed random walks at scale. *arXiv preprint arXiv:1906.04834*.

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. and J. L. Thorne (2019). A phylogenetic approach disentangles interlocus gene conversion tract length and initiation rate. *arXiv preprint arXiv:1908.08608*.

Ji, X., Z. Zhang, A. Holbrook, A. Nishimura, G. Baele, A. Rambaut, P. Lemey, and M. A. Suchard (2019). Gradients do grow on trees: a linear-time o(n)-dimensional gradient for statistical phylogenetics. *arXiv preprint arXiv:1905.12146*.

Zhou, W., X. Ji, S. Obata, A. Pais, Y. Dong, R. Peet, and Q.-Y. J. Xiang (2018). Resolving relationships and phylogeographic history of the Nyssa sylvatica complex using data from RAD-seq and species distribution modeling. *Molecular phylogenetics and evolution 126*, 1–16.