"All models are wrong, but some are useful."
–George Box

Throughout my life, I have been fascinated by technology, which earned me the title of "hacker" by most people I knew. While technology was my personal hobby, my academic interests have revolved around human health and genetics. In my youth, these two passions were compartmentalized, but when I joined the Bioinformatics program at UC San Diego, I studied the applications of theoretical Computer Science to biological big data analysis, and I learned how to approach biomedical problems from an algorithmic mindset. Towards the beginning of my Ph.D. career, when I entered the mentorship of Dr. Siavash Mirarab, whose lab specializes in the development of novel phylogenetics methods targeting large-scale datasets, I was offered an HIV-related project. I immediately jumped at this opportunity to unite my passions of Computer Science and human health, not realizing that I would be stumbling upon the focus of my thesis, and more broadly, of my entire research career.

HIV evolves rapidly, so the use of molecular data can provide epidemiologists actionable information. For example, reconstructing the evolutionary history of the virus can provide key insights into its transmission history. Recently, multiple methods have been developed to infer properties of transmission networks from molecular data, but it is not always clear which method/setting combination performs best for a specific downstream use-case or for specific epidemiological conditions. More broadly, the effectiveness of these methods in helping achieve public health goals is the subject of ongoing clinical and theoretical research.

Transmission networks are difficult to study because the parameters that drive an epidemic (e.g. network shape, transmission rates) cannot be controlled. A relatively inexpensive method to investigate questions related to epidemics is via simulation, and although existing epidemic simulators exist (e.g. epinet, TreeSim, seedy, PANGEA.HIV.sim), yet no existing tools can perform a full end-to-end simulation (social contact network, disease transmissions, viral phylogenies and sequences, and real-world errors). Further, existing tools make model assumptions that may not be realistic in all conditions.

Thus, the primary project of my Ph.D. has been the development of FAVITES (FrAmework for VIral Transmission and Emission Simulation), an epidemic simulation tool that performs the full end-to-end simulation workflow. The significance of FAVITES is in its flexibility: the tool's workflow is designed as interactions of abstract modules that make no model assumptions, and each *implementation* of a given module incorporates a statistical model. Thus, the model sampled in a single FAVITES execution is defined by the user-selected module implementations.

In addition to FAVITES, my work in the Mirarab lab has led me to develop statistical models and tools. In the realm of statistical modeling, I developed a novel model of tree evolution, the *dual-birth* model, and I used it to estimate the number of active and inactive *Alu* retrotransposons in the human genome. In terms of tool development, my focus has been to provide scalable methods to fill various deficiencies I have experienced, namely in HIV epidemiology. Transmission clustering is a process by which epidemiologists cluster patients based on molecular distances of their viral samples, yet existing methods (e.g. Cluster Picker and HIV-TRACE) typically take hours to run on datasets with thousands of samples. I developed TreeCluster, a tool that performs transmission clustering on a viral phylogeny, and I used it to infer clusters on a tree of over 100,000 Influenza samples in seconds. HIV-TRACE is extensively used by epidemiologists to study HIV transmission clusters, but the tool is parametric in that it depends on a user-provided distance threshold, so I developed TreeN93, a Python tool that produces a data structure that, once constructed, permits the rapid querying of HIV-TRACE clusters of *any* threshold. Further, TreeN93 then outputs completely non-parametric clusters by optimizing an epidemiologically-motivated objective function on the resulting tree. Because all of my work is centered around tree structures in Python, I required the use of a Python module for storing, traversing, and manipulating trees efficiently. However, the existing options (e.g. DendroPy, ETE Toolkit) added overhead that slowed my analyses when performed on ultra-large datasets, so I developed TreeSwift, a lightweight Python module specifically focused on tree traversal and manipulations that is extremely scalable in terms of both operation runtime and memory consumption.

My future research will be centered around open computational problems that interest me in HIV epidemiology. For example, HIV samples are generally sequenced when an individual begins antiretroviral therapy (ART), so given *x* sequences with sequencing (and thus ART initiation) times, who are the *n* individuals who are expected to the virus the most? Solution efficacy can be explored using FAVITES: simulate an epidemic, run the proposed solution using samples obtained before a time *t*, and use the known transmission network to determine the efficacy of the proposed solution. Further, many of the current methods in HIV epidemiology rely on model assumptions, yet statistical models catered specifically towards HIV are limited, and there is much room for novel models to better capture features of its evolution (e.g. recombination, drug resistance sites, etc.). There are many open computational problems in viral epidemiology, and as a Lecturer with Potential Security of Employment (LPSOE) at UC San Diego, I will continue to develop scalable open source tools solving these problems for the community to utilize to combat HIV and other infectious diseases. Because of the modularity of my work, I hope to mentor undergraduates who are interested in computational problems with medical relevance.

Aside from HIV research, much of my scholarly activity involves the development of high-quality online educational materials (largely in the form of Massive Adaptive Interactive Texts, or MAITs) for use by instructors in flipped classes as well as for integration into Massive Open Online Courses (MOOCs). I discuss the nature of my online materials in my *Statement of Teaching*, and I have developed such materials for many Bioinformatics topics (focusing on Bioinformatics Algorithms) as well as for Data Structures (which has been integrated into UC San Diego's CSE 100 course as well as at other universities). As an LPSOE at UC San Diego, I hope to develop more MAITs to integrate into flipped courses and MAITs. The two next topics especially of interest to me are introductory programming in Python, which I feel will be especially useful to students pursuing Data Science (either as major/minor or as a career), and stochastic models in phylogenetics, which is of interest to me because of my research. I would also want to create more content (and potentially courses) intersecting computation and medicine.

References

Groendyke, C., Welch, D., & Hunter, D. R. (2012). A Network-based Analysis of the 1861 Hagelloch Measles Data. *Biometrics*, 68(3), 755-765. doi:10.1111/j.1541-0420.2012.01748.x

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6), 1635-1638. doi:10.1093/molbev/msw046

Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., & Ferguson, N. (2014). Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, 10(1). doi:10.1371/journal.pcbi.1003457

Moshiri, N., & Izhikevich, L. (2018). *Design and Analysis of Data Structures*. Amazon Kindle Direct Publishing.

Moshiri, N., & Mirarab, S. (2017). A Two-State Model of Tree Evolution and Its Applications to Alu Retrotransposition. *Systematic Biology*, 67(3), 475-489. doi:10.1093/sysbio/syx088

Moshiri, N., Ragonnet-Cronin, M., Wertheim, J. O., & Mirarab, S. (2018). FAVITES: Simultaneous simulation of transmission networks, phylogenetic trees, and sequences. *bioRxiv*. doi:10.1101/297267

Moshiri, N. (2018). TreeCluster: Massively scalable transmission clustering using phylogenetic trees. *bioRxiv*. doi:10.1101/261354

Moshiri, N. (2018). TreeSwift: A massively scalable Python tree package. *bioRxiv*. doi:10.1101/325522

Pond, S. L., Weaver, S., Brown, A. J., & Wertheim, J. O. (2018). HIV-TRACE (TRAnsmission Cluster Engine): A Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Molecular Biology and Evolution*, 35(7), 1812-1819. doi:10.1093/molbev/msy016

Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J., & Lycett, S. (2013). Automated analysis of phylogenetic clusters. BMC Bioinformatics, 14(1), 317. doi:10.1186/1471-2105-14-317

Ratmann, O., Hodcroft, E. B., Pickles, M., Cori, A., Hall, M., Lycett, S., . . . Fraser, C. (2016). Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. *Molecular Biology and Evolution*, 34(1), 185-203. doi:10.1093/molbev/msw217

Stadler, T., & Bonhoeffer, S. (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1614), 20120198-20120198. doi:10.1098/rstb.2012.0198

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, 26(12), 1569-1571. doi:10.1093/bioinformatics/btq228

Worby, C. J., & Read, T. D. (2015). SEEDY (Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens. *PLoS One*, 10(6). doi:10.1371/journal.pone.0129745