Outline of proposed research

I am currently interested in solving large scale inference problems arising from phylogenetics. My research group is currently working on developing phylogenetic methods to be used in studying the diversity of human populations, which can be based on biological data as well as linguistics data. I am currently focusing on developing methods to analyze linguistics data, which is motivated by there being relatively fewer number of quantitative frameworks for analyzing large amounts of linguistic data compared to the abundance of methods available for analyzing biological data. This imparity is prohibiting the development of cohesive framework for adjoining the two fields for the scientific study of the diversity of human populations.

In historical linguistics, the comparative method is a common approach taken by linguists to find the evidences supporting hypothesis regarding an evolutionary relationship between a group of languages. The expectation is that if any two languages derive from a common ancestor, they should share remnants of the common ancestor through sound correspondences that regularly occur in the words of the two languages. Paramount to the task of establishing sound correspondences between languages is the detection of cognate sets. A cognate set is a set of words that share the same etymological origin and form the basic unit of data of the comparative method. Despite its importance, linguists are limited to relatively small number of cognate sets as the cognate set detection is a time consuming process. There has been attempts to automate cognate detection in the computational linguistics literature. However, the existing methods are limited to only two languages or require as further input a tree of languages that already describes the evolutionary relationships among the group of languages under study.

We are working on developing a probabilistic model that jointly infers both the language tree and the cognate sets given a list of vocabularies of the languages under study. In order to be of practical use, we require the vocabularies to be organized only by gloss, which eases the data preparation step and ultimately would allow the model to be used on analyzing large corpus. Aside from the computational tractability associated with processing large datasets, a further problem is that the signal-to-noise ratio in the form of false cognates may be considerably high. Our main objective is to build a computationally tractable model that can process large corpus and makes accurate inference amidst noise. To establish sound correspondences, we use model based on new approximation of large scale weighted finite state transducers.[1] To establish correct sound correspondences, we need to estimate the transition probabilities of sounds over time; we model this process using continuous time Markov chains and the rate matrix. Therefore, the statistical goal of this project is to derive a consistent estimator for the rate matrix, which quantifies the substitution rates between all possible sounds found in the languages under study. For inference, we use Sequential Monte Carlo method, which is an emerging, efficient method for exploring the space of trees.[2] We have already developed algorithms for maximizing computational resources to achieve accurate inference using SMC.[3] These methods are fully generic and we are currently working on releasing a software implementation of these methods built on top of Spark[4] for Monte Carlo practitioners tackling large scale inference problems in science and engineering. We are currently testing on the Austronesian language family data; we are collaborating with linguists to analyze large corpus on Athabaskan language family.

[1] Jason Eisner. Parameter estimation for probabilistic finite-state transducers. ACL, 2002.

[2] A. Bouchard-Côté, Sriram Sankararaman, and Michael I. Jordan. Phylogenetic inference via Sequential Monte Carlo. Systematic Biology, 2012.

[3] Refer to my contributions

[4] M. Zaharia et al. Spark: cluster computing with working sets. Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, 2010