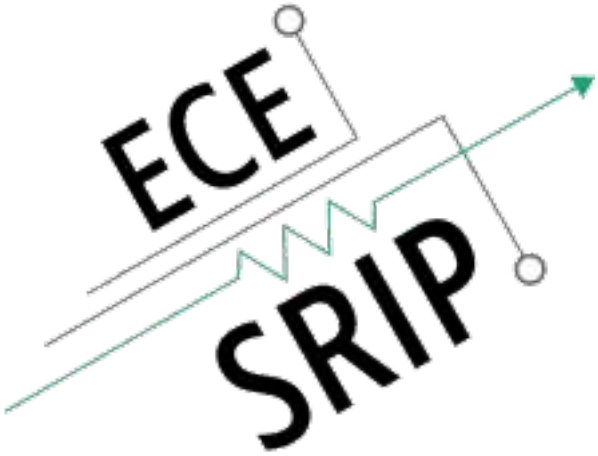


FACULTY MENTOR

Siavash Mirarab

PROJECT TITLE

Reconstructing evolutionary distances from bits and pieces of strings

PROJECT DESCRIPTION

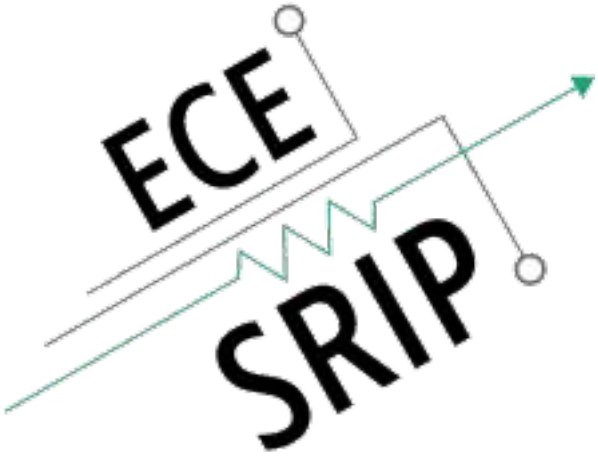
Description: A genome skim is a low-pass sequencing of a genome, producing many short reads across the genome, covering it at low coverage. Our recent work has shown this approach holds much promise in addressing important questions in understanding biodiversity, ecology, and even food provenance. Central to these questions is the ability to compute the average nucleotide identity (ANI) between a query and reference genome skims in an assembly-free and alignment-free manner. We have developed accurate k-mer-based methods in the past for this goal. In this project, our goal is to develop, implement, and test new algorithms to compute ANI from genome skims. We have several algorithmic ideas, which need to be implemented (in C++, python or Java), improved, and tested. The project will involve coding, algorithm development, and extensive testing on data.

INTERNS NEEDED

1 BS and 1 MS

PREREQUISITES

Required Qualifications:
1.Comfortable with programming in either C or Java or python.
2. Ability to use Unix systems.
3. Eager to learn other languages.
4. Any algorithmic skills will be a big plus.



FACULTY MENTOR

Siavash Mirarab

PROJECT TITLE

Expectation-Maximization (EM) algorithms for inferring ancient evolutionary times

PROJECT DESCRIPTION

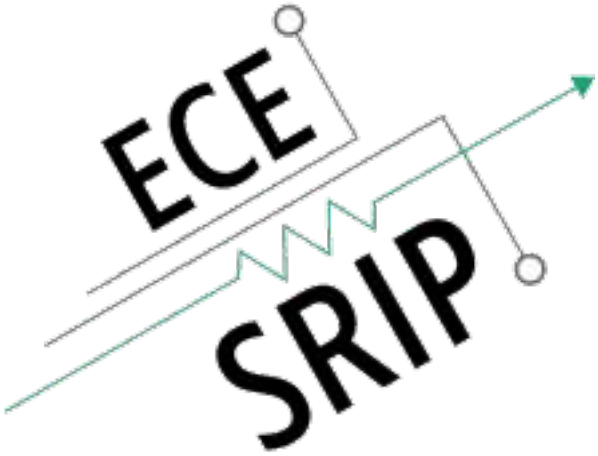
Description: A fundamental problem in evolutionary analyses is dating: inferring times when evolutionary events happened based on changes to genetic data. These analyses require combining data from carbon-dating fossil (or patient sampling time for viral evolution) with genomic data. Assuming a clock model, dating can be formulated as estimating a set of hyperparameters to maximize a likelihood function (i.e. maximum likelihood (ML)-based methods). ML-based methods face two difficulties: (1) the correct clock model that determines the likelihood function is not known and (2) optimizing the likelihood function is difficult because it involves an integral over the continuous domain of the rates. To tackle these problems, here we propose a nonparametric representation of the distribution of the rates. We discretize the unknown distribution and assume the rates are drawn from a multinomial distribution parameterized based on these bins. We use the Expectation-Maximization (EM) algorithm to estimate the probability mass of each bin from the observed data. Given a large enough number of rate categories, our method can approximate any distribution of the rates and therefore, can be applied to any clock model that assumes the rates are i.i.d.

INTERNS NEEDED

2 MS and 1 BS

PREREQUISITES

Required Qualifications:
1. Understanding of statistics, ideally knowing the EM algorithm
2. Ability to code in C++, Java, or python.



FACULTY MENTOR
Siavash Mirarab

PROJECT TITLE
Statistical models of amino acid evolution that account for biochemical properties

PROJECT DESCRIPTION
Description: Statistical models of sequence evolution are used to model evolutionary change across the tree of life. However, the connection of these changes to the biochemical properties of amino acids is often ignored. In this project, we build on recent advances in statistical modeling of sequence evolution to build a software tool that can infer models of sequence evolution while accounting for biochemical properties of amino acids. The project involves understanding a statistical model, implementing in C and python, and extensive data analyses to test the method. The project will involve a collaboration with a biologist and analyses of both real and simulated data.

INTERNS NEEDED
1 MS and 2 BS

PREREQUISITES
Required Qualifications:
1. Ability to build software pipelines