FACULTY MENTOR

Siavash Mirarab

PROJECT TITLE

Machine Learning Methods for Analyzing Biological Datasets

PROJECT DESCRIPTION

We will use various machine learning methods to integrate knowledge of the evolutionary history of organisms and their DNA sequences. Our goal will be to develop models that embed sequences in numerical spaces in ways that allow downstream applications. We will work closely with biological data.

This project will be in person.

INTERNS NEEDED

- 2

PREREQUISITES

- Advanced skills with Python and Pytorch.

FACULTY MENTOR

Siavash Mirarab

PROJECT TITLE

Spectral Analyses of Sequence kmers Using Probabilistic Modelling

PROJECT DESCRIPTION

One way to make sense of genomic sequences is to look at their kmer spectrums: a histogram showing how many small sequences of length k appear in the sequencing dataset. What impacts these spectra is a web of various factors, which are all stochastic in nature. Thus, our observations are a function of many different processes, and we are interested in disentangling those processes. This project will explore various methods to do so, applying them to real datasets.

This project will be in person.

INTERNS NEEDED

- 2

PREREQUISITES

- Comfort with math, especially basic linear algebra and probabilistic modeling.
- Some level of comfort with optimization techniques, particularly in terms of implementation using existing packages.
- Ability to implement mathematical algorithms into code (Python, perhaps).
- Interest in extensive empirical testing and tweaking of algorithms on data.