Tools for Faster and More Sensitive Sequence Annotation, and Visualization of Those Annotations

Genevieve Krause[1], Dave Rich[1], Jack Roddy[1], **Travis Wheeler**[1*](travis.wheeler@umontana.edu)

Institutions: [1]University of Montana, Missoula

Website URLs:
https://github.com/TravisWheelerLab/hmmer/tree/frameshift
https://github.com/TravisWheelerLab/MMOREseqs
https://sodaviz.org/

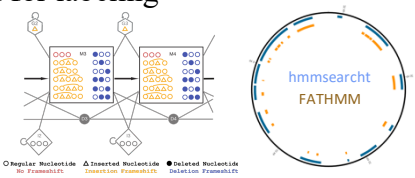
Project Goals

Microbial communities are ingrained in essentially every conceivable niche. We are particularly motivated by the need to understand soil communities that play a key role in the plant-soil dynamic, with impact on food and fuel crop production. To understand the roles of these microbial communities, it is vital that we maximally annotate their genomic and functional capacity. In this work, we are primarily concerned with (i) the problem of accurately annotating protein-coding DNA that contains insertions or deletions that induce frameshifts in the coding sequence, (ii) performing this annotation at high speed, and (iii) generating complex and interactive visualizations of annotations.

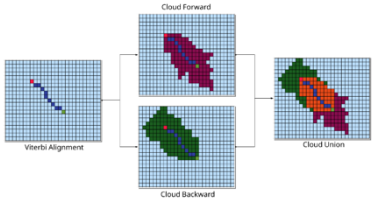
Abstract Text

We describe advances in methods for annotation with profile hidden Markov models (pHMMs):

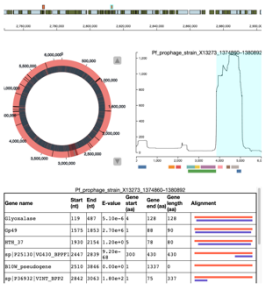
- We have developed an implementation of an extended pHMM for labeling protein-coding DNA with frameshifts. This tool, FATHMM, is substantially more sensitive than competing software methods in the fact of frameshifts, and is only slightly slower than frameshift-oblivious translated search with pHMMs.

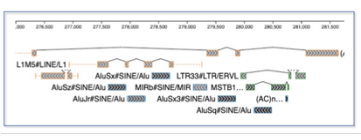


- We have developed a sparse dynamic programming algorithm to produce highly accurate Forward/Backward profile HMM alignments with 20-100x reduction in memory and runtime requirements (MMOREseqs);



- We have created an open-source TypeScript library that supports efficient development of custom, dynamic, and interactive visualizations of annotations of linear and circular genomic sequence (SODA).





Funding Statement: This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0021216, and NIH NIGMS R01GM132600.