Biological Design Doctoral Defense

Statistical Sequence Alignment of Protein Coding Regions

School for Engineering of Matter, Transport and Energy

Juan José García Mesa

Advisor: Dr. Reed Cartwright

Abstract

"Sequence alignment is an essential method in bioinformatics and the basis of many analyses, including phylogenetic inference, ancestral sequence reconstruction, and gene annotation. Sequence artifacts and errors made in alignment reconstruction can impact downstream analyses, leading to erroneous conclusions in comparative and functional genomic studies. While such errors are eventually fixed in the reference genomes of model organisms, many genomes used by researchers contain these artifacts, often forcing researchers to discard large amounts of data to prevent artifacts from impacting results.I developed COATi, a statistical, codon-aware pairwise aligner designed to align protein-coding sequences in the presence of artifacts, such as early stop codons and abiological frameshifts. Unlike common sequence aligners, which rely on amino acid translations, only model insertion and deletions between codons, or lack a statistical model, COATi combines a codon substitution model specifically designed for protein-coding regions, a complex insertion-deletion model, and a sequencing base calling error step. The alignment algorithm is based on finite state transducers (FSTs), computational machines well-suited for modeling sequence evolution. I show that COATi outperforms available methods using a simulated empirical pairwise alignment dataset as a benchmark.The FST-based model and alignment algorithm in COATi is resource-intense for sequences longer than a few kilobases. To address this constraint, I developed an approximate model compatible with traditional dynamic programming alignment algorithms. I describe how the original model is marginalized to build the approximate model and how the alignment algorithm is implemented by modifying the popular Gotoh algorithm. I simulate a benchmark of alignments and measure how well marginal models approximate the original method.Finally, I present a novel tool for analyzing sequence alignments. Available metrics can measure the similarity between two alignments or the column uncertainty within an alignment but cannot produce a site-specific comparison of two or more alignments. AlnDotPlot is an R software package inspired by traditional dot plots that can provide valuable insights when comparing pairwise alignments. I describe AlnDotPlot and showcase the utility of the different available models."

October 25, 2023; 10 AM; Biodesign Auditorium - room B105