BST281: Genomic Data Manipulation, Spring 2019

Monday 02: Sequence Alignment

This lecture introduces methods for comparing biological sequences, searching databases for similar sequences to a query sequence, and mapping sequencing reads to reference databases.

## Sequence homology

## Divergent evolution from a common ancestor leads to similarity among modern-day sequences (homology). Homologous sequences differ due to (e.g.) nucleotide substitutions, insertions, and deletions. Detecting homology is fundamental to many biological data analyses (e.g. phylogenetics and gene function prediction).

## Sequence alignment

Sequence alignment quantifies homology between sequences by pairing (mapping) corresponding nucleotide or amino acid sites. Alignments are paths through a grid whose rows and columns represent the two sequences. Optimal alignments can efficiently “built up” using dynamic programming algorithms. Alignments may be global, in which all sites are aligned, or local, in which only the best-aligned sequence substrings are identified.

## BLAST

BLAST is a heuristic algorithm for performing homology-based search of a query sequence against a sequence database. BLAST assumes that related sequences contain short, identical regions called “seeds” that are rare in unrelated sequences. Indexing a database enables rapid identification of seeds, which are then extended to check for local alignments. Alignments are scored by percent identity, coverage, and statistical significance (*E*-value). Many BLAST variants exist for specialized tasks (e.g. mapping DNA sequences to potential protein products).

## Aligning short reads

Special considerations are needed when aligning short reads to a reference genome due in part to their vast numbers. Most methods (e.g. bowtie2) use a Burrows-Wheeler transformation of the reference, which reduces its memory footprint and facilitates fast query lookup. Alignments of short reads are reported using SAM/BAM format, which also doubles as a FASTQ alternative for storing sequence information.

## Mapping without alignment

# Homologous sequences can be identified without explicit sequence alignment using methods called alignment-free mapping or pseudoalignment. MinHash and derivatives compare sequences based on overlap in their *k*-mer content, focusing on a random subset of *k*-mers for computational efficiency. Kallisto maps sequencing reads to maximally informative *k*-mers in the assembly graph of potential transcripts.

# Suggested textbook reading

* Lesk, Chapter 5, p150-160 (sequence alignments and search)
* Pevsner, Chapter 3, p69-79, p96-112 (sequence alignment, skipping scoring matrices)
* Pevsner, Chapter 4, p121-151 (BLAST)
* Pevsner, Chapter 9, p399-406 (read mapping, SAM/BAM format)

# Related literature

* [Altschul, Stephen F., et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.](http://www.sciencedirect.com/science/article/pii/S0022283605803602)