BST281: Genomic Data Manipulation, Spring 2019

Wednesday 01: Biological Sequences: Concepts and Data

This lecture introduces biological sequences (e.g. genomes, genes, transcripts, and proteins) as an important biological data type, as well as methods and tools for acquiring, representing, and organizing such data.

## Biopolymers

Nucleic acids (DNA and RNA) and polypeptides (proteins) are biological polymers comprised of specially arranged molecular subunits. This property allows biopolymers to store information. Biopolymers are associated via the central dogma of molecular biology. Biopolymers are represented as strings of 1-letter representations of their molecular subunits. Biopolymer sequences have a dual role in biology: 1) representing biological entities (e.g. genes and genomes) and 2) as a means of quantifying biological states (e.g. RNA-seq).

## DNA sequencing

DNA sequencing methods are largely based on the natural chemistry of DNA replication (“sequencing by synthesis”). Sanger sequencing is a long-standing method based on fragment termination and size separation. Modern, high-throughput methods use reversible termination and image processing to follow the stepwise synthesis of 10s of millions of DNA fragments simultaneously. Short (~100 nt) sequencing reads are generated in pairs from the ends of randomly sampled DNA fragments (“shotgun sequencing”). These reads can be assembled into biological sequences (e.g. genes and genomes) or used directly for quantification.

## Determining RNA and protein sequences

RNA sequencing is performed through reverse transcription to DNA molecules, followed by traditional DNA sequencing. Proteins cannot be directly sequenced in such a way: their sequences are discovered computationally by searching DNA sequences for open reading frames (ORFs) or through mass spectrometry.

## FASTA files

The FASTA file format represents biological sequence data. It consists of individual header lines, one per sequence and starting with “>,” followed by one or more lines of corresponding sequence information.

## FASTQ files

FASTQ files store raw sequencing reads and associated quality scores (4 lines/read). Quality (Phred) scores vary logarithmically with the probability of an error and are represented by ASCII characters. FASTQ files typically come in pairs representing the respective ends of sequenced DNA fragments.

## Sequence databases

The NCBI and EMBL-EBI maintain a variety of useful databases for exploring/retrieving biological sequence information, including Genbank, RefSeq, Ensembl, and UniProt. The NCBI Sequence Read Archive (SRA) is the primary source of raw sequencing reads from scientific studies.

# Suggested textbook reading

* Pevsner, Chapter 2, p19-37 (Databases)
* Pevsner, Chapter 9, p377-393 (Sequencing methods and FASTQ files)
* Lesk, Chapter 1, p1-32 (Introduction to genomics)
* Lesk, Chapter 3, p110-130 (Sequencing methods and databases)

# Related literature

* [F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *PNAS* 12:5463-5467 (1977).](https://www.pnas.org/content/74/12/5463)