BST281: Genomic Data Manipulation, Spring 2018

Monday 02: 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 representing, storing, and searching sequence data.

## Biopolymers

Nucleic acids and polypeptides 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.

## 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.

## Sequence homology

## Divergent evolution from a common ancestor leads to similarity among modern-day sequences (sequence homology). Homologous sequences differ due to nucleotide substitutions, insertions, and deletions. Identifying sequence homology is fundamental to many analyses of biological data (e.g. comparative genomics, protein functional assignment). Small-scale similarity (e.g. sequence motifs) can also result from convergent evolution.

## Sequence alignment

Sequence alignment detects homology between sequences by pairing corresponding nucleotide or amino acid sites. Alignment algorithms consider all possible alignment “paths” between two sequences, accounting for (e.g.) amino acid exchangeability and “block” sequence insertion/deletion. 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 is based on the assumption that well-aligning sequences contain short, gapless regions of high similarity (“seeds”). Seed matches are identified quickly from a pre-indexed database; the corresponding sequences are then locally aligned by traditional methods. BLAST alignments are scored by % identity/similarity, degree of locality (coverage), and statistical significance (*E*-value).

## 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.

# Suggested textbook reading

* Pevsner, Chapter 2, p19-37
* Pevsner, Chapter 3, p69-79, p96-112
* Pevsner, Chapter 4, p121-135, p138-151

# 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)
* [Qin, Junjie, et al. "A human gut microbial gene catalogue established by metagenomic sequencing." Nature 464.7285 (2010): 59-65.](http://www.nature.com/nature/journal/v464/n7285/abs/nature08821.html)