Cells contain genetic material → DNA → large molecules called nucleic acids only four of them: Adenine, Thymine, Guanine, Cytosine.

- Stored in long segments → chromosomes → sum of chromosomes → genome
- Parts of the genome get transcribed then translated into proteins → those are made of amino-acids (20 of them): Alanine, Threonine, Glycine, Cysteine, ...

Sequencing technologies can be used to identify nucleic acids (DNA)

- Important: we cannot directly sequence DNA of a cell → there is always a laboratory protocol of substantial complexity → library preparation
- Sequencing instruments sequence libraries.

Sequencing Technologies - perspective

1st generation: Frederic Sanger develops DNA sequencing technology. Latest versions 3 million bases/day, 1500bp long reads

2nd generation: (next-gen) sequencing started 2005 with the release of the 454 sequencing platform. 600 billion bases/week, 150bp long reads

3rd generation: single molecule (no DNA amplification required), these are not replacing but augmenting 2nd generation systems, longer reads, shorter turnarounds
The wide range of characteristics among available platforms provides opportunities both to conduct groundbreaking studies and to waste money on scales that were previously infeasible.

One of the most elusive formats: FASTA

- Seemingly trivial but it is also “under-specified”
- Many “custom” extensions
- Tools make assumptions
- Surprising number of problems
Alphabets

- International Union of Pure and Applied Chemistry (IUPAC) codes
- Nucleic acid sequences
- Peptide sequences → polypeptides could be proteins

A multi record FASTA

It is not clear what the sequence above contains nucleic acids or amino acids
(Feels like a nucleic acids because of having so many ACTG both those are also valid amino acids)

Pitfalls

- The length of sequences was not regulated (a gigantic oversight!)
  If FASTA files were set to say 80 character limits we easily index then randomly access any interval inside it!
- Strange things will happen if one were to flatten (linearize) a gigantic sequence - human chromosome 300 million bases – tools may break in spectacular ways

More considerations

- Many tools will embed extra information into either the identifier or the “free zone” the description section
- See the FASTA format wiki page
First step of any sequence processing step understand your FASTA file

• What’s is what in this file

1. How many sequences do we have

2. Are sequences all on a single line or over multiple lines

3. What is the identifier, what is embedded in the description

Understand your file

What is the identifier and descriptor?

A handy command line calculator

Written in 1975 arbitrary precision calculator
A short demo on numerical precision

Vary the numbers in cell B1: 1, 1000, 1E6, 1E20, 1E32

Cut out the beginning of each read

Produce the list of subsequences with counts

Homework 4

The facility reports that each sequence in the lec4.fa file contains a 10 base long barcode followed by a 10 base primer sequence.

There should be 4 barcodes and 1 primer across all sequences. The primer though may have a mismatch.

Verify this statement and report your findings.