2. Genes and proteins

- The sequence as a model of DNA
- Genes: from Mendel to molecular biology
- The genetic code
- A translation algorithm
- Implementing the genetic code
- Algorithms + data structures = programs
- The algorithm design trade-off
- DNA sequencing
- Whole genome sequencing
- **How to find genes?**

How to find genes?

Genome annotation

- Prediction of
  - gene locations
  - gene (i.e. protein) functions

- A bacterial genome
  - typically (*E. coli*) 4.5 Mb
  - ≈ 4,500 genes

- A human genome
  - 3.5 Gb
  - ≈ 20,000 genes « only »

What do we know about gene location?

- A gene (coding region) starts on a “start” codon: ATG

But

if inside a coding sequence, ATG codes also for Methyonine

- A gene always ends on a stop codon (TAA, TAG, or TGA)

But

start and stop codons occur ouside coding regions

in the so-called intergenic regions

And that’s not the end of it!

- A gene may be located on any of two strands
- A coding region is a succession of codons, i.e. triplets
  - Stop and Start codons must be separated by a number of bases multiple of 3
- There are 3 ways for grouping bases into triplets: 3 reading phases
  - starting on base i
  - starting on base i+1
  - starting on base i+2

And that’s not the end of it!

- A gene may be located on any of two strands
- A coding region is a succession of codons, i.e. triplets
  - Stop and Start codons must be separated by a number of bases multiple of 3
- There are 3 ways for grouping bases into triplets: 3 reading phases
  - starting on base i
  - starting on base i+1
  - starting on base i+2

CCTAGCTAATTGCTATTAATTGTGTCATGACGTCTAG
CCTAGCTAATTGCTATTAATTGTGTCATGACGTCTAG
CCTAGCTAATTGCTATTAATTGTGTCATGACGTCTAG

And that’s not the end of it!

- A gene may be located on any of two strands
- A coding region is a succession of codons, i.e. triplets
  - Stop and Start codons must be separated by a number of bases multiple of 3
- There are 3 ways for grouping bases into triplets: 3 reading phases
  - starting on base i
  - starting on base i+1
  - starting on base i+2

```
  CCTAGCTAATTGCTATTAATTGTGTCATGACGTCTAG
 CCTAGCTAATTGCTATTAATTGTGTCATGACGTCTAG
CCTAGCTAATTGCTATTAATTGTGTCATGACGTCTAG
```

- The start and the stop codons of a gene must be in the same phase