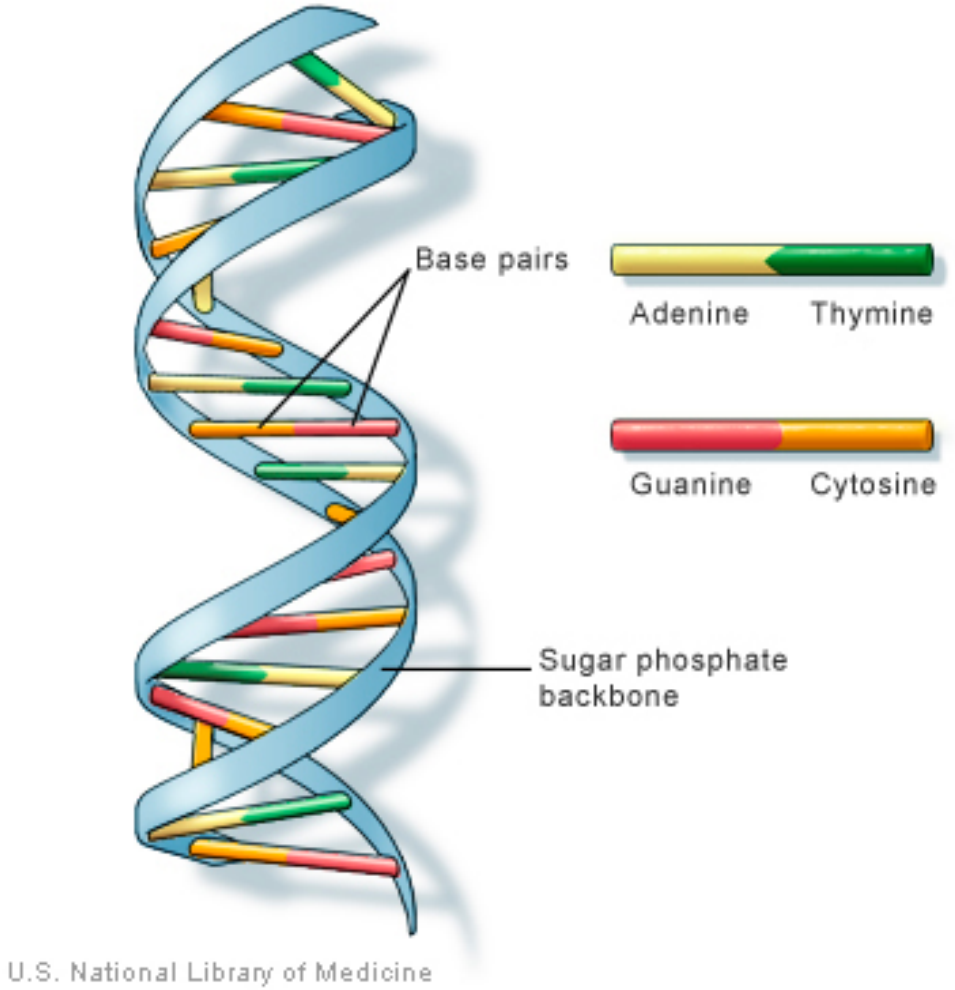
CMSC 423:
Finding Biological Signals

Part 1

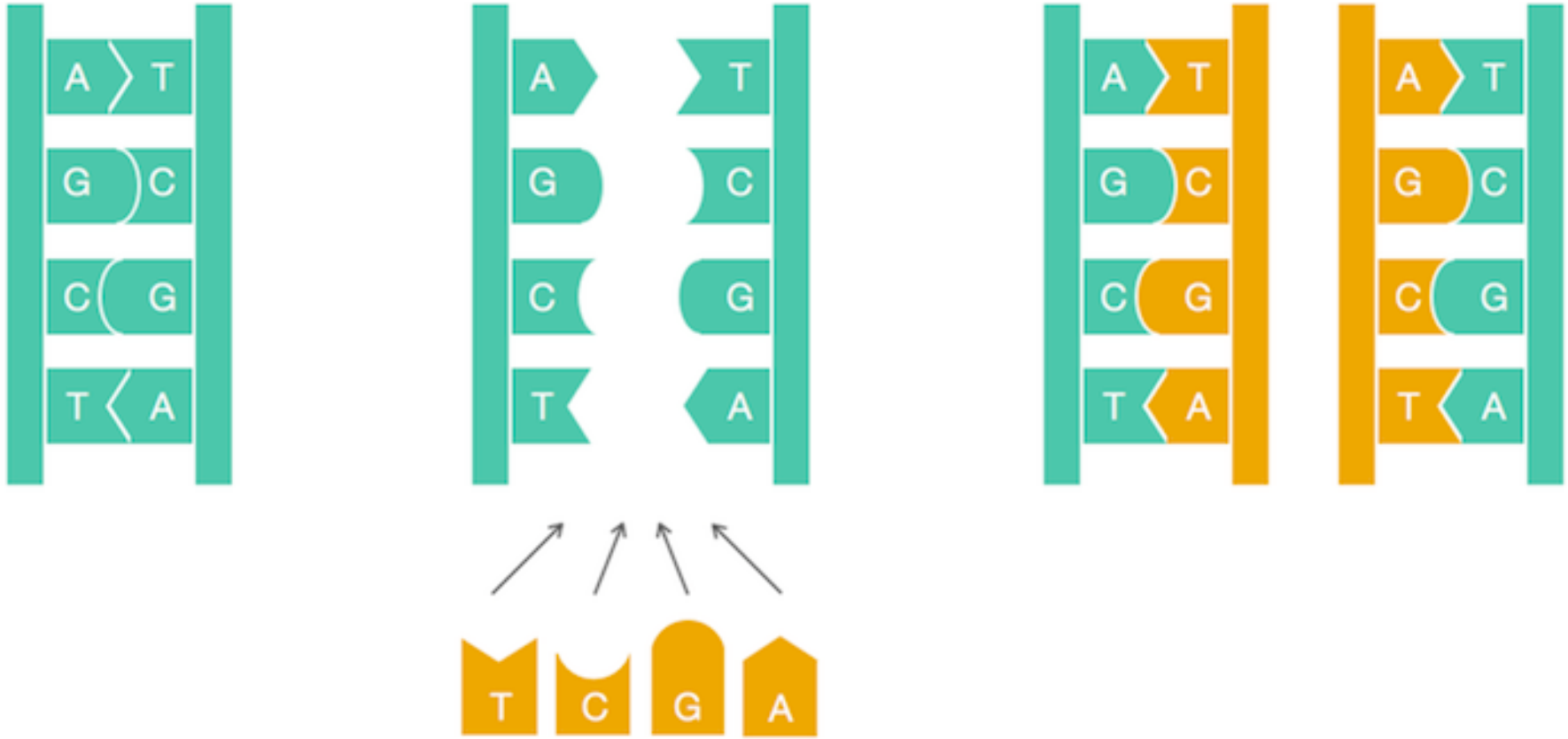
DNA- the code of life



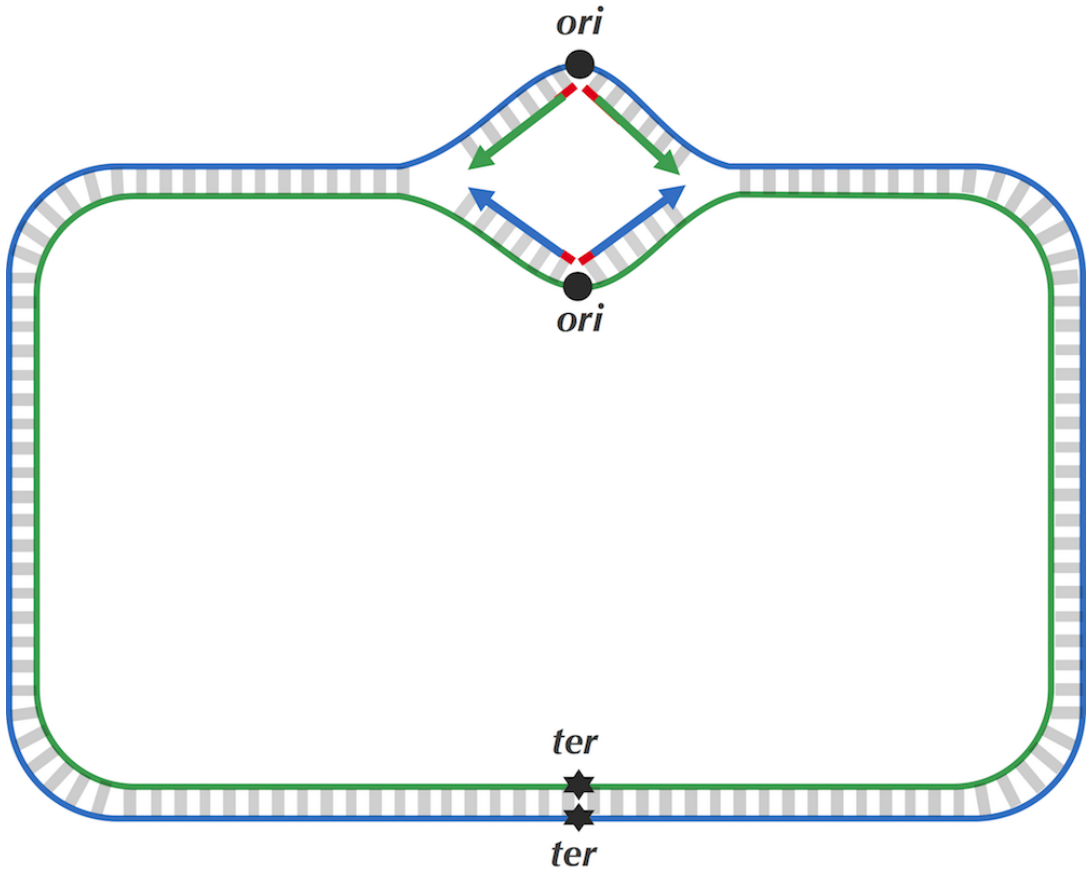
- Stores genetic information
- Consists of four types of bases (A, T, C, G)
- Nucleotide=base + sugar + phosphate
- A binds to T, C binds to G
- Double helix

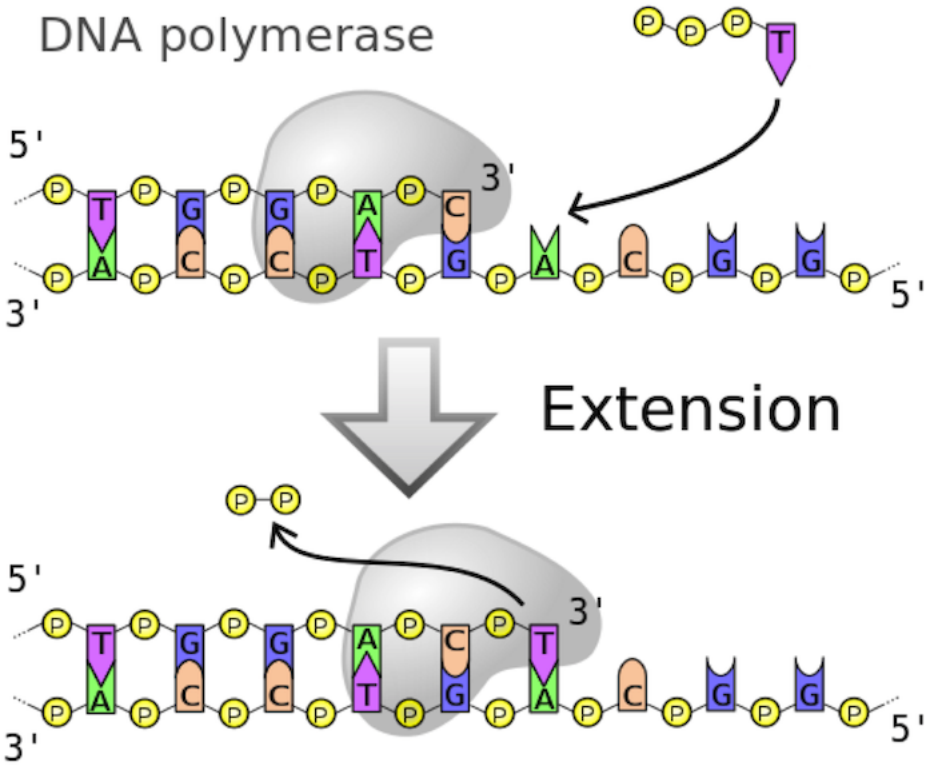
https://ghr.nlm.nih.gov/primer/basics/dna

DNA Replication



DNA Replication





Problem: Finding Origin of Replication

- **Input**:

Genome, a DNA string of nucleotides from the four-letter alphabet {A, C, G, T}.

- **Output**:

The location of *ori* in *Genome*.

- Focusing on bacterial genomes, which consist of a single circular chromosome

- NOT A CLEARLY STATED COMPUTATIONAL PROBLEM!

DnaA mediates initiation of replication

ori of *Vibrio cholerae*:

atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc

DnaA: protein that binds to a short segment within the *ori* to begin replication

***DnaA* box**: where *DnaA* binds, the “hidden” message within the *ori*

Problem: Finding Hidden Messages in the Replication of Origin

- **Input**:

A string *Text* (representing the replication origin of a genome).

- **Output**:

A hidden message in *Text*.

- NOT A CLEARLY STATED COMPUTATIONAL PROBLEM!

Finding Hidden Messages

- Two paradigms
  - Look for surprising events
  - Leverage biological knowledge

Finding Hidden Messages

- Look for deviations from what is expected
- Random DNA strings do not have long “parts” that repeat nearby each other
- Key idea: Find k-mers that are more frequent than expected

Finding Hidden Messages

- Two paradigms
  - Look for surprising events
  - Leverage biological knowledge

- Computational solution:
  Counting letters and words

https://en.wikipedia.org/wiki/Count_von_Count