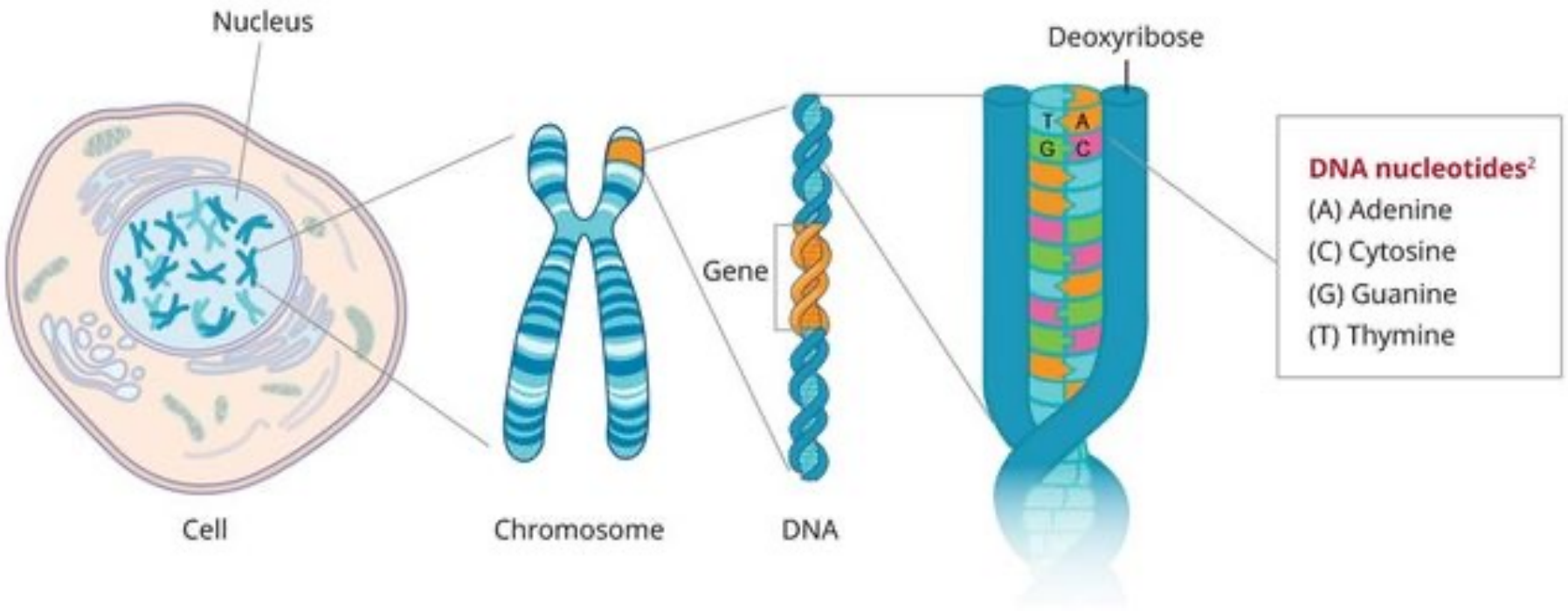
CS-4049
Bioinformatics

Spring 2025

Rushda Muneer

What is DNA?

- **Deoxyribonucleic Acid (DNA)** is the genetic material that carries instructions for the growth, development, and functioning of all living organisms.
- **Key Features:**
- **Double Helix Structure** (discovered by Watson & Crick)
- **Made of Nucleotides** (Adenine, Thymine, Cytosine, Guanine)
- **Carries Genetic Information** (codes for proteins)
- **Passed from Parents to Offspring**
- **In Short DNA is the blueprint of life!**

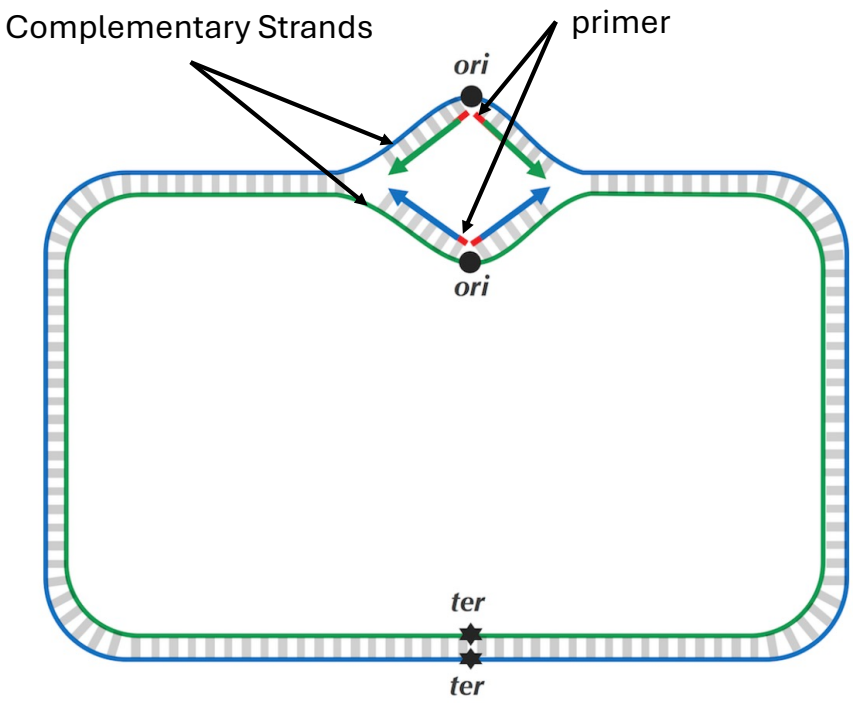

Nucleus
Cell
Chromosome
Gene
DNA
Deoxyribose
T
A
G
C
DNA nucleotides[2]
(A) Adenine
(C) Cytosine
(G) Guanine
(T) Thymine

What is genome?

- A **genome** is the complete set of genetic material (DNA) in an organism. Total amount of DNA (nucleus and mitochondria)
- **Think of the genome as the biological instruction book of life in which DNA is a blueprint**

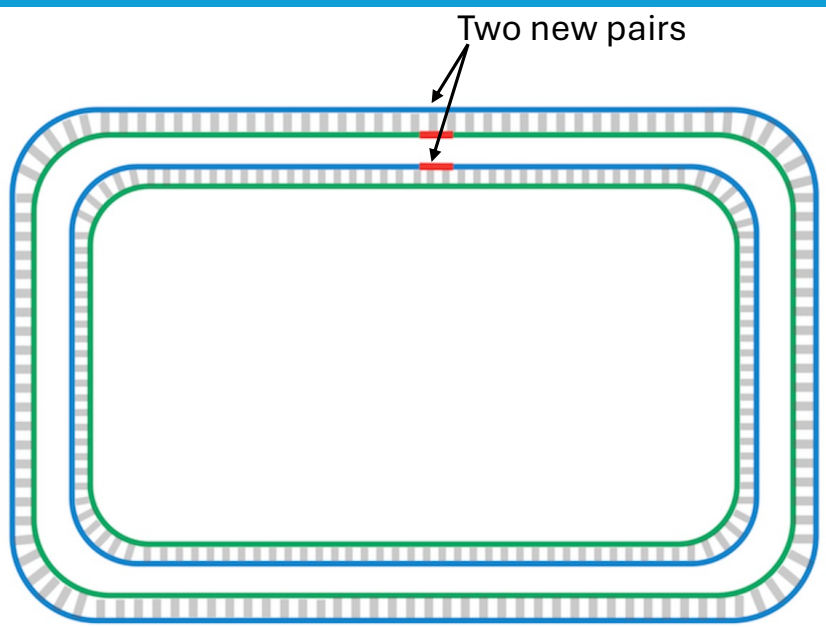
DNA Replication Process

- DNA replication begins at the **origin of replication (ori)** and proceeds **bidirectionally** (assumption).
- The **two complementary DNA strands unwind**, forming **replication forks** that expand around the chromosome.
- Replication continues until the **replication terminus (ter)**, located **opposite to ori**, is reached.
- **DNA polymerase does not wait for full strand separation**; it starts copying while the strands unravel.
- **Four DNA polymerases**, each responsible for a **half-strand**, initiate replication at ori.
- A **primer** (short complementary segment) is needed to **start replication**.



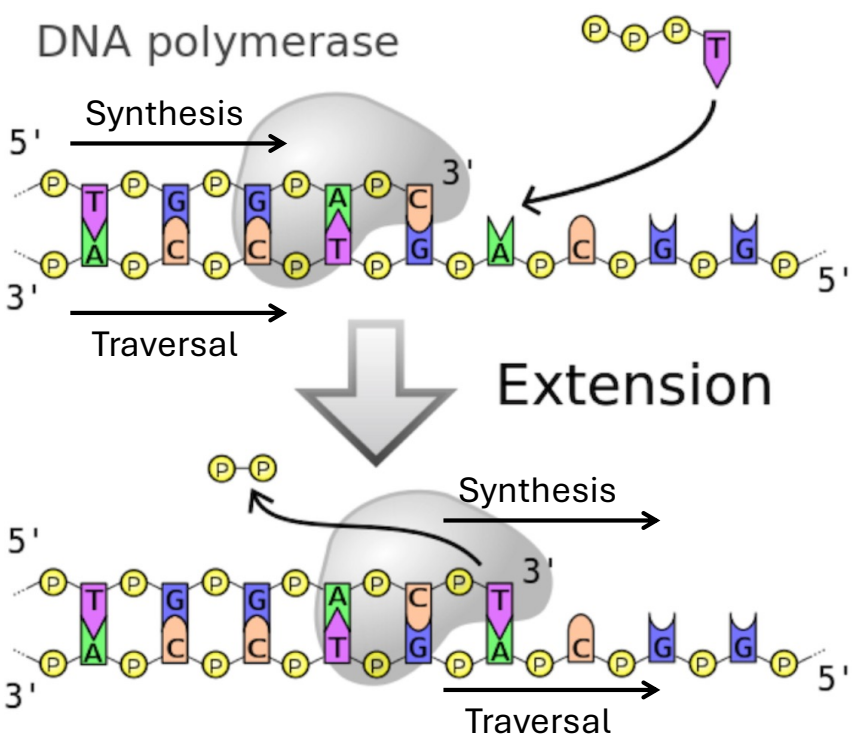
Replication Mechanism

- Nucleotides are added in either **clockwise** or **counterclockwise** direction until ter is reached.
- Complete chromosome replication results in **two pairs of complementary DNA strands**.
- The cell is now ready to divide.



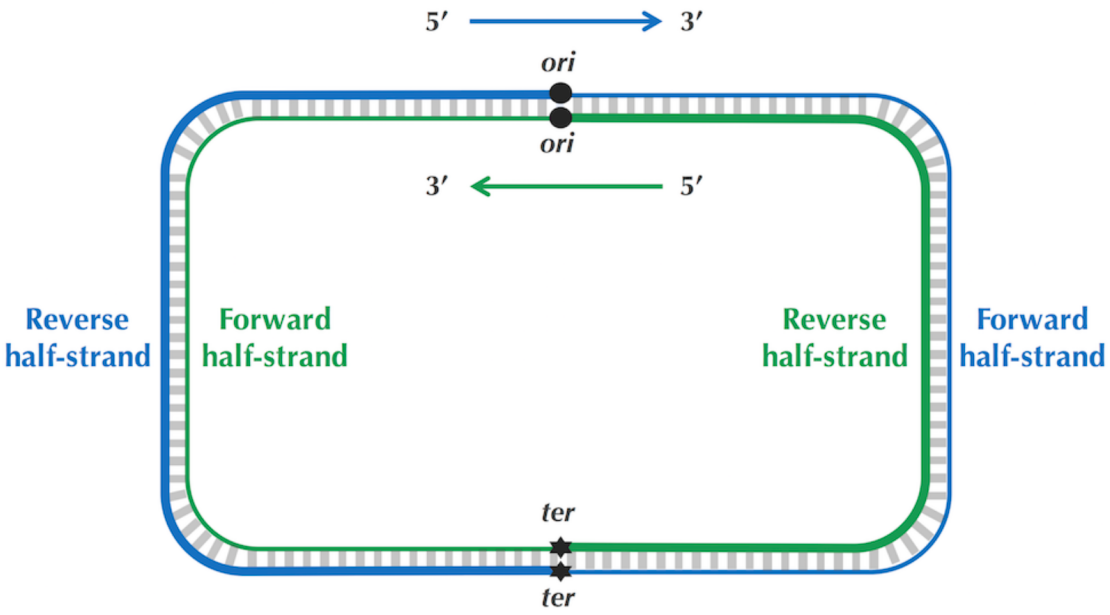
Directionality of DNA Polymerase

- Our previous description incorrectly assumed that DNA polymerases can copy DNA in both directions along a strand.
- In reality, DNA polymerases are **unidirectional** and can only traverse a template strand in the **3' → 5'** direction.
- DNA synthesis occurs only in the **5' → 3'** direction.
- This unidirectionality influences the mechanisms of **leading** and **lagging strand** synthesis.



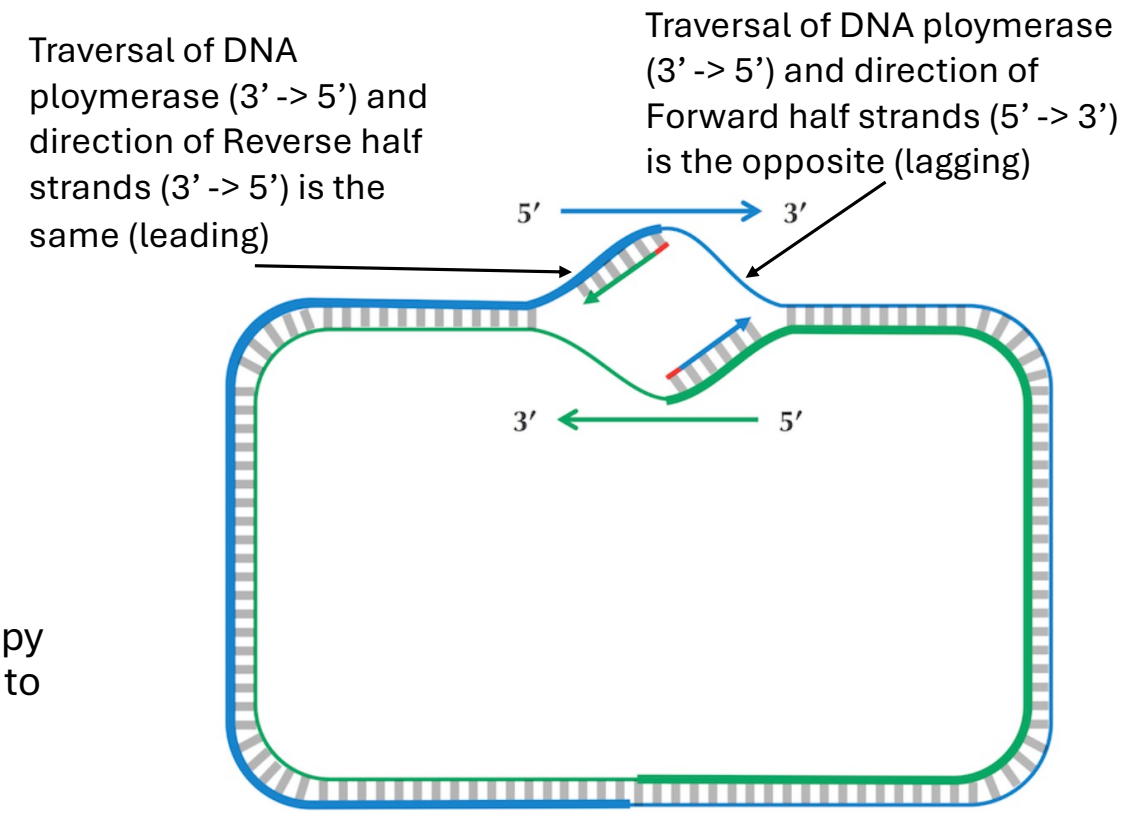
Understanding DNA Strands

- The unidirectionality of DNA polymerase necessitates a major revision of our naive replication model.
- If you walk along DNA from ori to ter, you will encounter four different half-strands of parent DNA.
- These half-strands are categorized based on their directionality:
  - **Forward Half-Strands** (thin blue and green lines): **5’ → 3’** direction.
  - **Reverse Half-Strands** (thick blue and green lines): **3’ → 5’** direction.



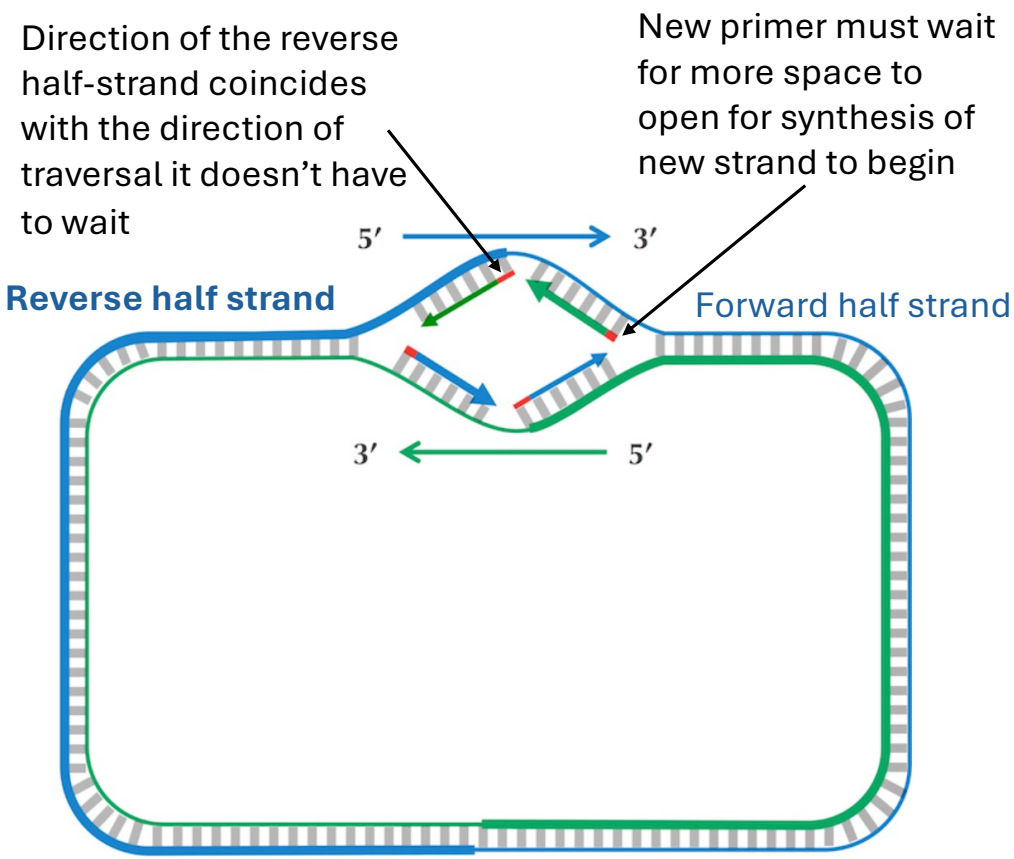
Asymmetry in DNA Replication

- DNA replication is **asymmetric**, meaning forward and reverse half-strands undergo very different replication processes.
- Reverse half-strands (**3' → 5'** direction): DNA polymerase can copy nucleotides **continuously** from ori to ter (**leading**).
- Forward half-strands (**5' → 3'** direction): DNA polymerase must replicate **backwards** toward ori because it cannot move in the 5' → 3' direction (**lagging**).



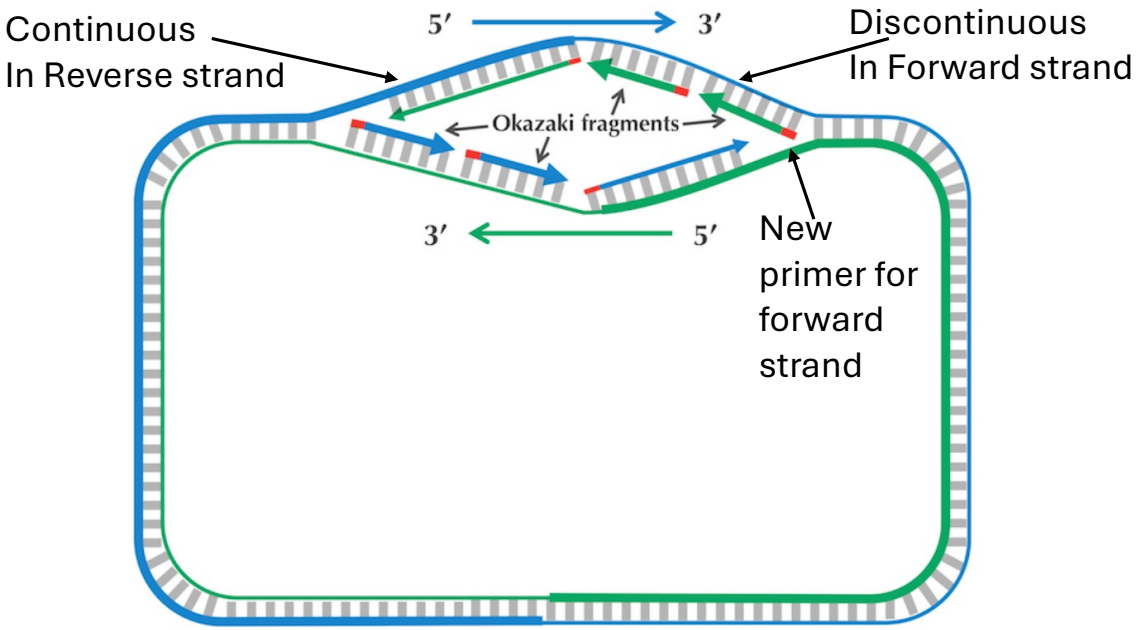
Replication on Forward Half-Strand

- DNA polymerase on a forward half-strand must wait for the replication fork to open (~2,000 nucleotides).
- A **new primer** is formed at the end of the replication fork.
- DNA polymerase then starts replicating a **small DNA fragment** from the primer **backward** toward ori.



Okazaki Fragments on Forward Half-Strands

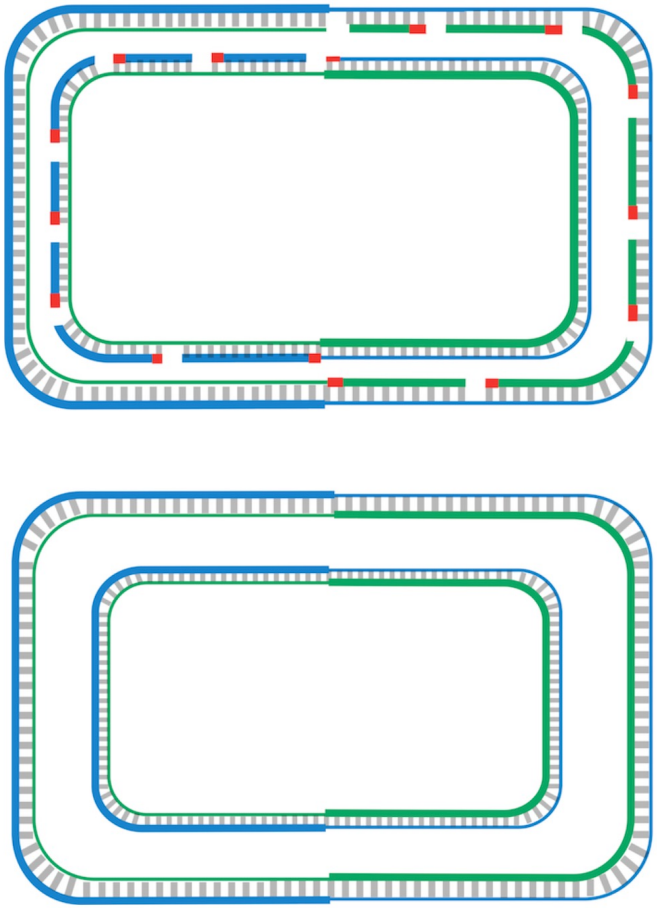
- Replication on reverse half-strands progresses **continuously** with a single primer.
- Replication on forward half-strands is **discontinuous** and requires multiple primers.
- DNA polymerase must **pause** after replicating a fragment until the replication fork opens another ~2,000 nucleotides.
- A **new primer** is required for each new fragment.
- This results in the formation of **Okazaki fragments**, which are short DNA segments synthesized from multiple primers.



- The replication fork continues to expand.
- Reverse half-strands (thick lines) require only **one primer**.
- Forward half-strands (thin lines) require **multiple primers** (shown in red) to synthesize Okazaki fragments.

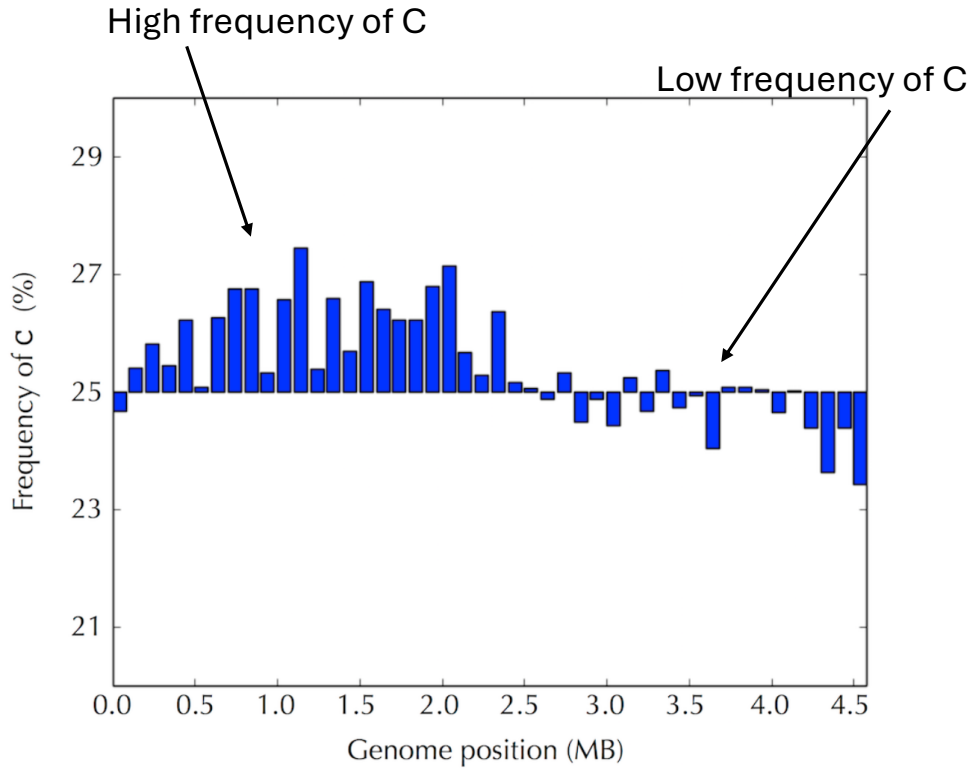
Finalizing DNA Replication

- When the replication fork reaches **ter**, most of the DNA has been synthesized, but **gaps remain** between Okazaki fragments.
- DNA ligase **sews together** consecutive Okazaki fragments.
- This process results in **two intact daughter chromosomes**, each with **one parent strand and one newly synthesized strand**.
- In reality, DNA ligase **works continuously**, sealing Okazaki fragments as they are formed rather than waiting until the end.



Peculiar Statistics of the Forward and Reverse Half-Strands

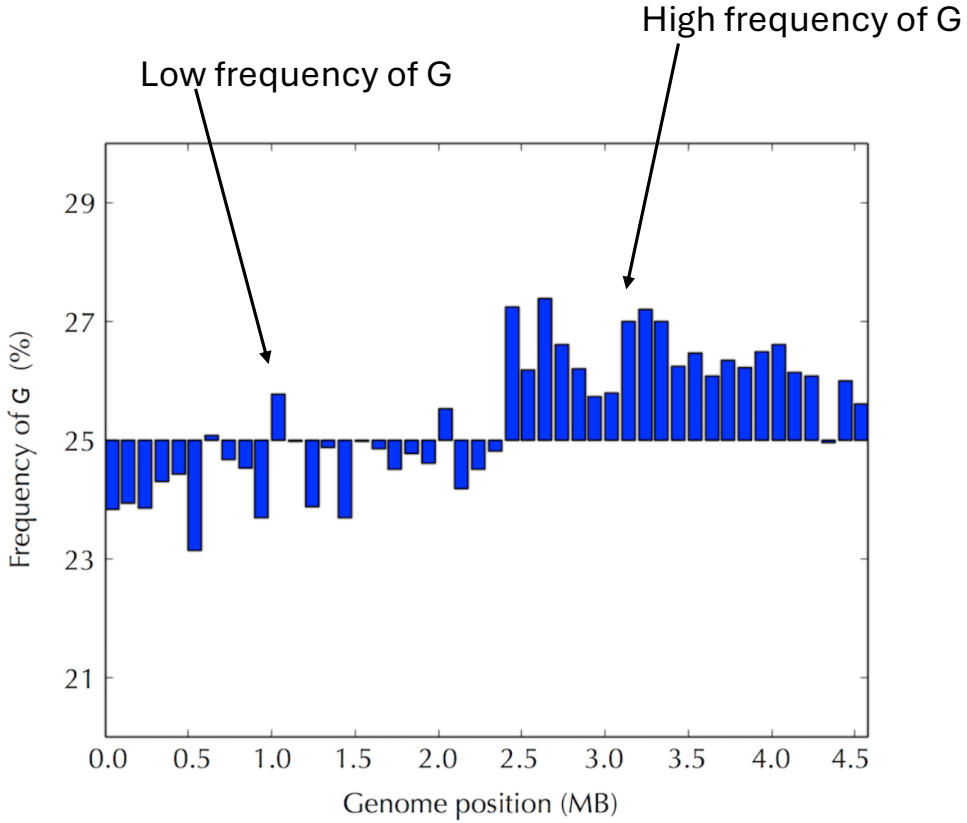
- **Key Observation:**
  - A **surprising pattern** emerges when analyzing cytosine frequency across the E. coli genome.
  - The genome is partitioned into **46 equal-sized fragments** (~100,000 nucleotides each), starting at **ter**.
- **Findings:**
  - The **first 23 fragments** (reverse half-strand) have a **high cytosine frequency** (above 25%).
  - The **last 23 fragments** (forward half-strand) have a **low cytosine frequency** (below 25%).



- Histogram shows cytosine frequency across the genome.
- **ter** is at position **0**, and **ori** is ~2.3 million nucleotides away.
- Reverse half-strand spans **first half**; forward half-strand spans **second half**.

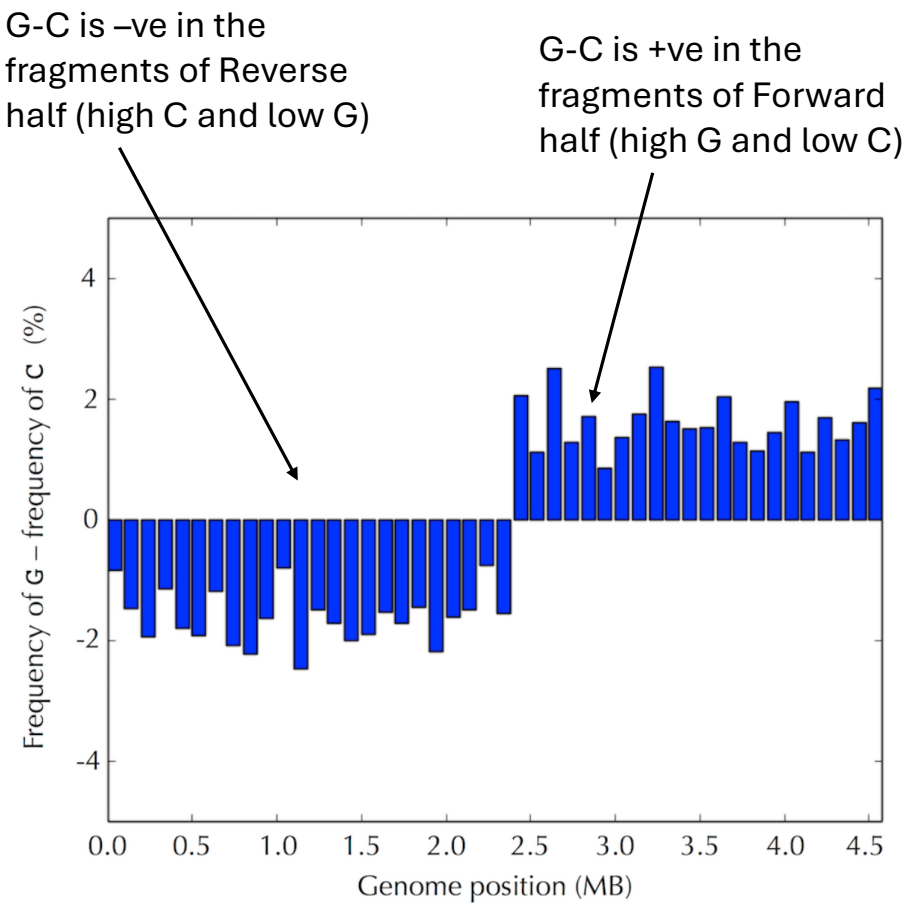
Peculiar Statistics of the Forward and Reverse Half-Strands

- **Key Observation:**
  - A **contrasting pattern** appears when analyzing guanine frequency across the E. coli genome.
- **Findings:**
  - **Reverse half-strand**: Most fragments have a **low guanine frequency** (below 25%).
  - **Forward half-strand**: Most fragments have a **high guanine frequency** (above 25%).
- **Implication:**
  - This pattern suggests a **strand-specific nucleotide composition bias** that may have biological significance.



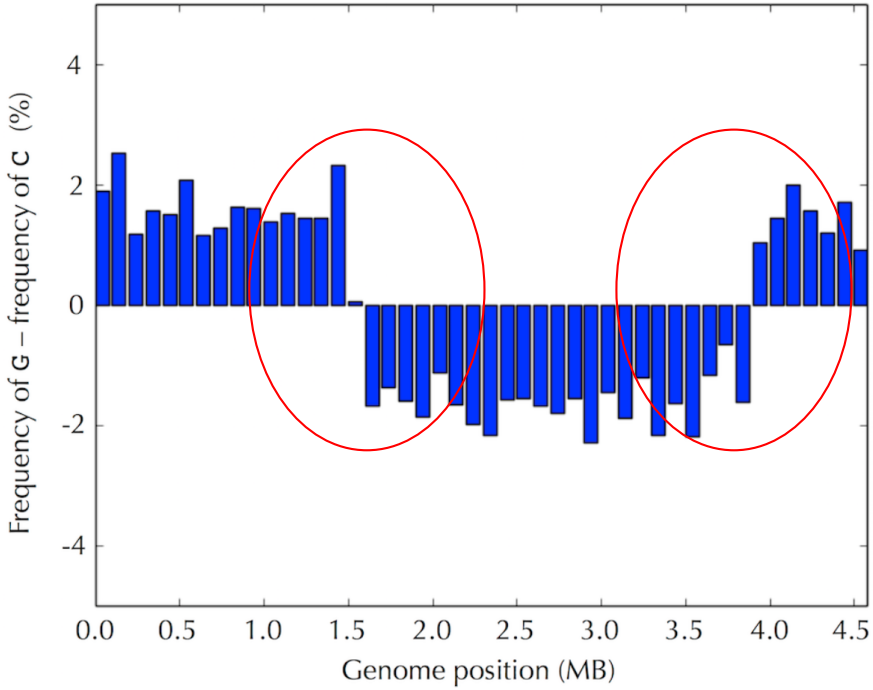
G-C Frequency Difference Analysis

- A **striking visualization** emerges when comparing guanine (G) and cytosine (C) frequency differences across **genome fragments.**
- The difference in **G and C frequencies** highlights a **peculiar statistical bias** between the **reverse and forward half-strands**.
- Forward and reverse half strands unite at **ori** or **ter**.
- The striking frequency difference can assist in finding the origin of replication



Using GC Frequency to Identify ori

- The transition point where **G - C frequency shifts from negative to positive** provides a clue about the location of **ori**.
- **Implication:**
  - If this pattern is **not a statistical fluke**, it suggests a **simple test** to identify ori.
  - By scanning the genome for the **G - C frequency transition**, we can estimate the replication origin.



Deamination

- **Replication Fork Asymmetry:**
  - DNA polymerase synthesizes DNA quickly on the reverse half-strand but faces delays on the forward half-strand.
- **Single-Stranded vs Double-Stranded DNA:**
  - Reverse half-strand remains double-stranded most of the time.
  - Forward half-strand spends more time single-stranded, which increases mutation rates.
- **Mutation Rate Discrepancy:**
  - Single-stranded DNA has a higher mutation rate.
  - A nucleotide with a higher mutation tendency in single-stranded DNA will be underrepresented on the forward half-strand.

Example : Thermotoga petrophila Genome

- Compare the nucleotide counts of the reverse and forward half-strands to detect substantial differences.
- This will help design an algorithm for locating ori in genomes where it's unknown.
- Nucleotide counts for forward and reverse half-strands are shown in the table.
- **Key Question:**
  - **STOP and Think:** Do you notice anything about the nucleotide counts in this table? What differences can be observed between the two strands?

	#C	#G	#A	#T
Entire strand	427419	413241	491488	491363
Reverse half-strand	219518	201634	243963	246641
Forward half-strand	207901	211607	247525	244722
Difference	+11617	-9973	-3562	+1919

Deamination and Nucleotide Frequency Discrepancies

- **A & T:** Frequencies are nearly identical between the forward and reverse half-strands.
- **C & G:** Noticeable discrepancies:
  - C is more frequent on the reverse half-strand (+11617).
  - G is more frequent on the forward half-strand (-9973).

	#C	#G	#A	#T
Entire strand	427419	413241	491488	491363
Reverse half-strand	219518	201634	243963	246641
Forward half-strand	207901	211607	247525	244722
Difference	+11617	-9973	-3562	+1919

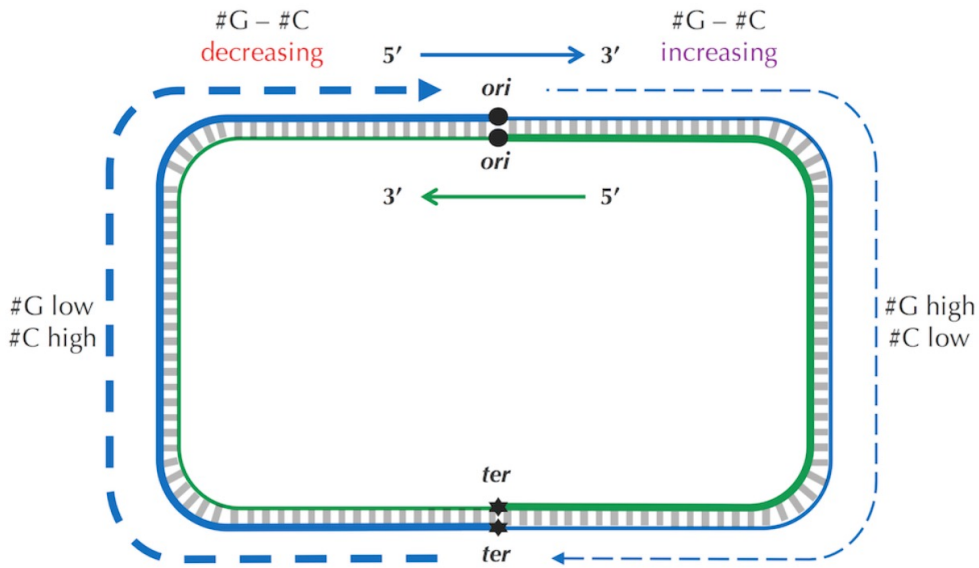
Deamination and Nucleotide Frequency Discrepancies

- **Deamination Process:**
  - **C → T Mutation:** Cytosine (C) deaminates to thymine (T), especially in single-stranded DNA.
- **Impact on G:** Deamination of C on the forward strand leads to a decrease in guanine (G) on the reverse strand.

Locating Ori Using Deamination Statistics

- The difference between guanine (G) and cytosine (C) is used to track strand direction.
  - The **reverse half-strand** shows a **negative** G-C difference: 201634−219518=−17884
  - The **forward half-strand** shows a **positive** G-C difference: 211607−207901=+3706
- **Running Total of G - C Difference:**
  - Traverse the genome and calculate the difference between the counts of G and C.
  - **Increasing Difference:** Suggests we are on the **forward half-strand**.
  - **Decreasing Difference:** Suggests we are on the **reverse half-strand**.

	#C	#G
Entire strand	427419	413241
Reverse half-strand	219518	201634
Forward half-strand	207901	211607



Skew Diagram Approach:

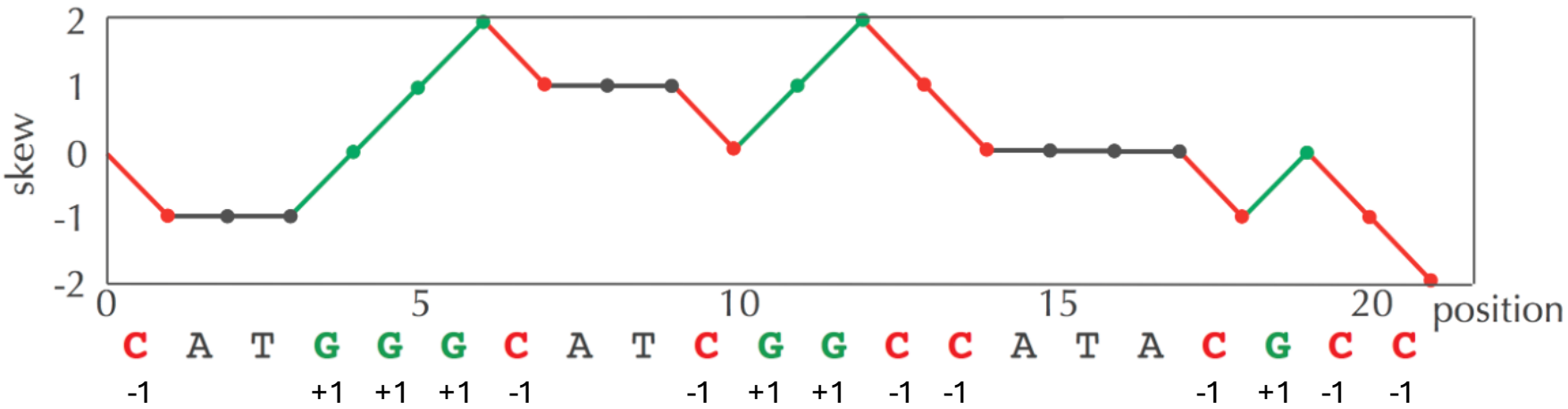
The difference between G and C provides a way to visualize strand direction, helping to identify ori.

Skew Diagram

- **Skew_i (Genome):** Difference between the total occurrences of G and C in the first *i* nucleotides of the genome.
- **Skew Diagram:** Plots **Skew_i(Genome)** as *i* ranges from 0 to the length of the genome (|Genome|), with **Skew_0(Genome) = 0**.
- **Skew Calculation:**
- For each nucleotide position *i* in the genome:
  - **If G:** $\text{Skew}_{i+1} = \text{Skew}_i + 1$
  - **If C:** $\text{Skew}_{i+1} = \text{Skew}_i - 1$
  - **Otherwise (A or T):** $\text{Skew}_{i+1} = \text{Skew}_i$

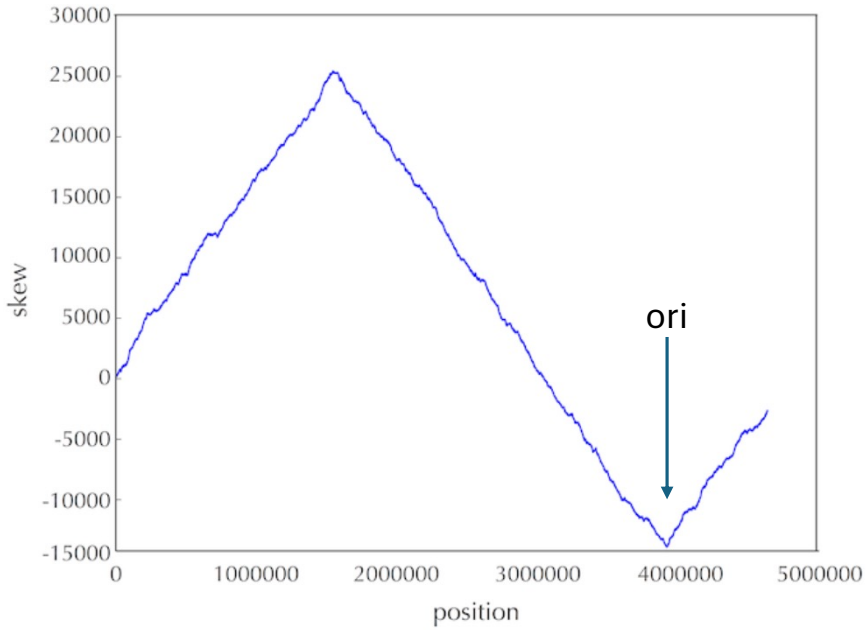
Skew Diagram

- **Example:**
  - DNA sequence **CATGGGCATCGGCCATACGCC**.



Skew Diagram for E. coli Genome

- The skew diagram of the linearized *E. coli* genome shows a clear pattern.
- The shape of the skew diagram is often similar in many bacterial genomes.
- **STOP and Think:** Based on the skew diagram, where do you think the **origin of replication (ori)** is located in *E. coli*?



Solving the Minimum Skew Problem and Locating Ori in E. coli

```
aatgatgatgacgtcaaaaggatccggataaaacatggtgattgcctcgc
ataacgcggtatgaaaatggattgaagcccgggccgtggattctactcaa
ctttgtcggcttgagaaagacctgggatcctgggtattaaaaagaagatc
tatttatttagagatctgttctattgtgatctcttattaggatcgcactg
ccctgtggataacaaggatccggcttttaagatcaacaacctggaaagga
tcattaactgtgaatgatcggtgatcctggaccgtataagctgggatcag
aatgaggggttatacacaactcaaaaactgaacaacagttgttctttgga
taactaccggttgatccaagcttcctgacagagttatccacagtagatcg
cacgatctgtatacttatttgagtaaattaacccacgatcccagccattc
ttctgccggatcttccggaatgtcgtgatcaagaatgttgatcttcagtg
```

- **Minimum Skew Result:**
  - Approximate location of ori in *E. coli* is at **position 3923620** based on the skew diagram.
- **Testing the Hypothesis:**
  - Solving the **Frequent Words Problem** in a window of length 500 starting at position 3923620 reveals no 9-mers (including reverse complements) that appear 3 or more times.
- **Results:**
  - Despite locating ori at position 3923620, **no ori** was found in this region, suggesting further exploration is needed to find DnaA boxes in *E. coli*.

Observations from Vibrio cholerae ori

```
atcaATGATCAACgtaagcttctaagcATGATCAAGgtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagATGATCAAGagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagCATGATCATggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctCTTGATCATcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctCTTGATCATgtt
tccttaaccctctattttttacggaagaATGATCAAGctgctgctCTTGATCATcgtttc
```

- In addition to the three occurrences of ATGATCAAG and its reverse complement CTTGATCAT, the ori contains:
  - ATGATCAAC
  - CATGATCAT
- These sequences differ from ATGATCAAG and CTTGATCAT by only a single nucleotide.
- This suggests that **subtle variations in the DnaA box sequences** might be the solution.
- This observation could guide adjustments to the algorithm to better identify DnaA boxes in *E. coli* and other bacterial genomes.

Frequent Words Problem with Mismatches

- Modify the algorithm for the **Frequent Words Problem** to find DnaA boxes by identifying frequent k-mers with possible mismatches.
- **Count(Text, Pattern, d):** The total number of occurrences of *Pattern* in *Text* with at most *d* mismatches.
- **Example:**
  - **Count(AACAAGCTGATAAACATTTAAAGAG, AAAAA, 1) = 4**
  - AAAAA appears four times with at most one mismatch: **AACAA, ATAAA, AAACA, AAAGA** (two occurrences overlap).
- **Exercise:** Compute *Count* (AACAAGCTGATAAACATTTAAAGAG, AAAAA,2).

Answer:11

Summary

- Genome Replication
- Decyphering DNA language
- Origin of Replication Finding Problem
- Frequent Word Problem
- Replication Process
- Asymetry in DNA Replication
- G-C Frequency Analysis
- Skew Diagram
- Frequent Word Problem with Mismatches

Assignment 1

- Details will be uploaded to google classroom by Friday (January 24)
- Submission on google classroom Friday(January 31)
- Design algorithms for topics discussed in class
- Check the uploaded assignment questions before the next class
- If you have any questions dicuss in the next class