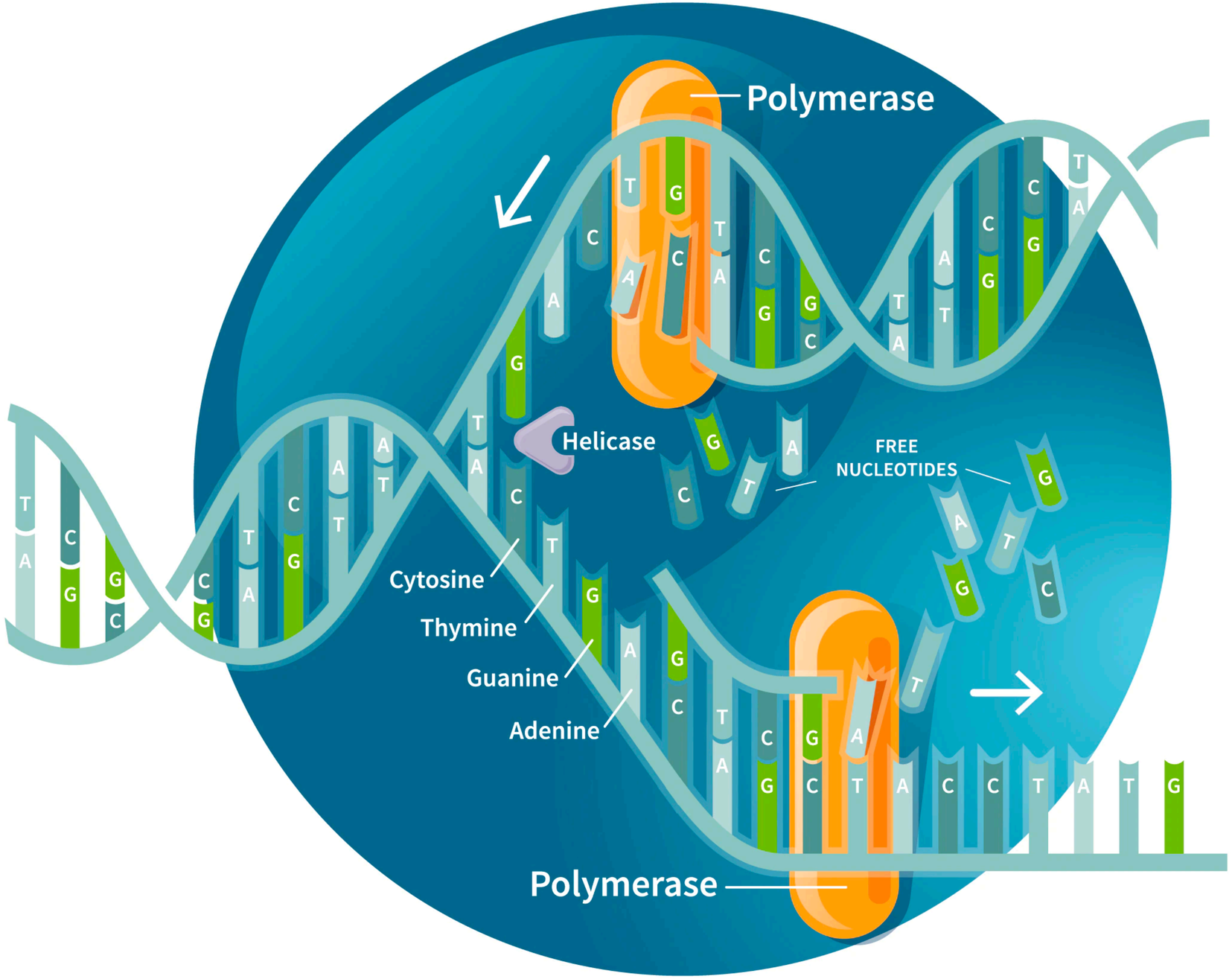
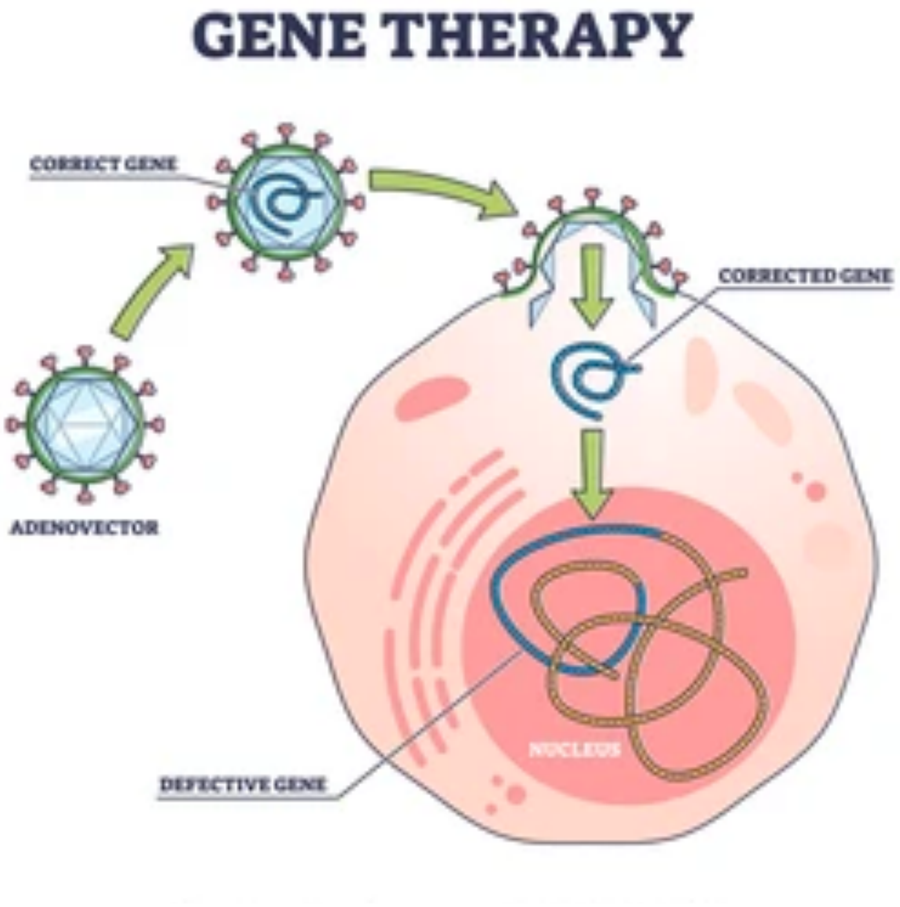
Origin of Replication

Genomic Big Data



Origin of Replication

- Replication begins in a genomic region called the replication origin (denoted oriC) molecular copy machines called DNA polymerases.
- Gene therapy methods use genetically engineered mini-genomes, which are called viral vectors because they are able to penetrate cell walls (just like real viruses).
- To engineer frost-resistant tomatoes and pesticide-resistant corn.
- In 1990, gene therapy was first successfully performed on humans when it saved the life of a four-year-old girl suffering from Severe Combined Immunodeficiency Disorder.
- The idea of gene therapy is to intentionally infect a patient who lacks a crucial gene with a viral vector containing an artificial gene that encodes a therapeutic protein. Once inside the cell, the vector replicates and eventually produces many copies of the therapeutic protein, which in turn treats the patient's disease.



Origin of Replication

- A biologist
  - Delete various short segments from the genome in an effort to find a segment whose deletion
- A computer scientist?

Finding Origin of Replication Problem:

Input: A DNA string *Genome*.
Output: The location of *oriC* in *Genome*.

Bacterial Genome

- A single circular chromosome
- Bacterial genome encoding oriC is typically a few hundred nucleotides long.
- Begin with a bacterium in which oriC is known, and then determine what makes this genomic region special
- Vibrio cholerae chromosome, which consists of 1,108,250 nucleotides

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

DnaA Box

- The initiation of replication is mediated by DnaA, a protein that binds to a short segment within the oriC known as a DnaA box.

- DnaA box as a message within the DNA sequence

Hidden Message Problem:
Find a "hidden message" in the replication origin.

Input: A string *Text* (representing the replication origin of a genome).
Output: A hidden message in *Text*.

A hidden message

```
53++!305))6*;4826)4+.)4+);806*;48!8‘60))85;1+(;:+*8
!83(88)5*!;46(;88*96*?;8)*+(;485);5*!2:*+(;4956*2(5
*-4)8‘8*;4069285);)6!8)4++;1(+9;48081;8:8+1;48!85:4
)485!528806*81(+9;48;(88;4(+?34;48)4+;161;:188;+?;
```

A hidden message

```
53++!305))6*;48826)4+.)4+);806*;48!8‘60))85;1+(;:+*8
!83(88)5*!;46(;88*96*?;8)*+(;485);5*!2:*+(;4956*2(5
*-4)8‘8*;4069285);)6!8)4++;1(+9;48081;8:8+1;48!85;4
)485!528806*81(+9;48;(88;4(+?34;48)4+;161;:188;+?;
```

A hidden message

```
53++!305))6*THE26)H+.)H+)TE06*THE!E‘60))E5T1+(T:+*E
!E3(EE)5*!TH6(TEE*96*?TE)*+(THE5)T5*!2:*+(TH956*2(5
*-H)E‘E*TH0692E5)T)6!E)H++T1(+9THE0E1TE:E+1THE!E5TH
)HE5!52EE06*E1(+9THET(EETH(+?3HTHE)H+T161T:1EET+?T
```

Frequent Words

```
FrequentWords(Text, k)
    FrequentPatterns ← an empty set
    for i ← 0 to |Text| − k
        Pattern ← the k-mer Text(i, k)
        Count(i) ← PatternCount(Text, Pattern)
    maxCount ← maximum value in array Count
    for i ← 0 to |Text| − k
        if Count(i) = maxCount
            add Text(i, k) to FrequentPatterns
    remove duplicates from FrequentPatterns
    return FrequentPatterns
```

Frequent Words

k	3	4	5	6	7	8	9
count	25	12	8	8	5	4	3
***k*-mers**	tga	atga	gatca	tgatca	atgatca	atgatcaa	atgatcaag
			tgatc				cttgatcat
							tcttgatca
							ctcttgatc

Frequent Words

Frequent Words

Pattern Matching Problem

Pattern Matching Problem:
Find all occurrences of a pattern in a string.

Input: Strings *Pattern* and *Genome*.
Output: All starting positions in *Genome* where *Pattern* appears as a substring.

ATGATCAAG

116556, 149355, **151913**, **152013**, **152394**, 186189, 194276, 200076, 224527, 307692, 479770, 610980, 653338, 679985, 768828, 878903, 985368

Thermotoga petrophila

```
aactctatacctccttttttgtcgaatttgtgtgatttatagagaaaatcttattaactga
aactaaaatggtaggtttggtggtaggttttgtgtacattttgtagtatctgatttttaa
ttacataccgtatattgtattaaattgacgaacaattgcatggaattgaatatatgcaaa
acaaacctaccaccaaactctgtattgaccattttaggacaacttcagggtggtaggttt
ctgaagctctcatcaatagactattttagtctttacaaacaatattaccgttcagattca
agattctacaacgctgttttaatgggcgttgcagaaaacttaccacctaaaatccagtat
ccaagccgatttcagagaaacctaccacttacctaccacttacctaccacccgggtggta
agttgcagacattattaaaaacctcatcagaagcttgttcaaaaatttcaatactcgaaa
cctaccacctgcgtcccctattatttactactactaataatagcagtataattgatctga
```

Thermotoga petrophila

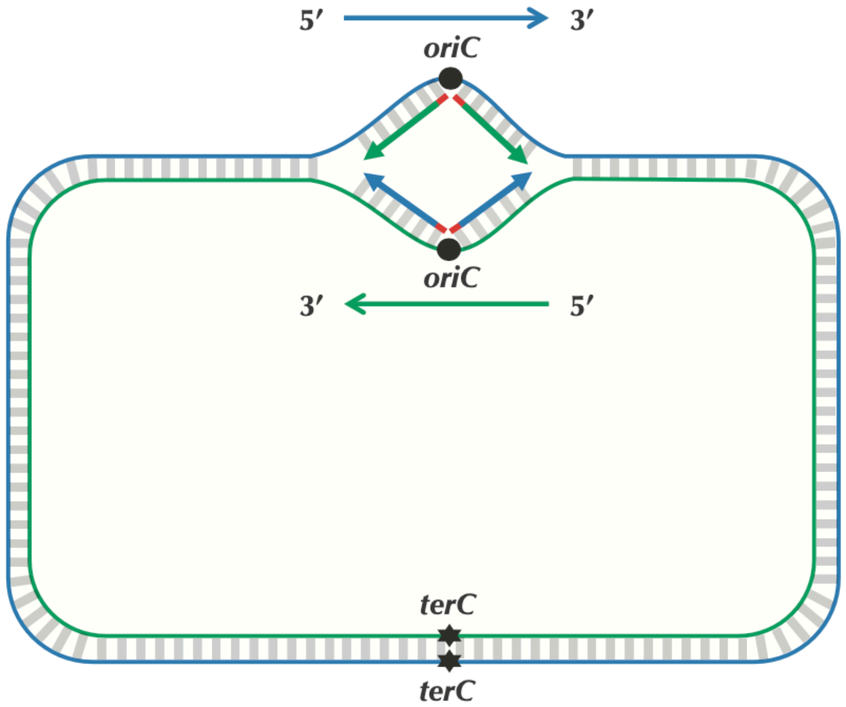
aactctatacctccttttgtcgaatttgtgtgatttatagagaaaatcttattaactga
aactaaaatggtaggtttGGTGGTAGGttttgtgtacattttgtagtatctgatttttaa
ttacataccgtatattgtattaaattgacgaacaattgcatggaattgaatatatgcaaa
acaaaCCTACCACCaaactctgtattgaccattttaggacaacttcagGGTGGTAGGttt
ctgaagctctcatcaatagactattttagtctttacaaacaatattaccgttcagattca
agattctacaacgctgttttaatgggcgttgcagaaaacttaccacctaaaatccagtat
ccaagccgatttcagagaaacctaccacttacctaccactaCCTACCACCcgggtggta
agttgcagacattattaaaaacctcatcagaagcttgttcaaaaatttcaatactcgaaa
CCTACCACCtgcgtcccctattatttactactactaataatagcagtataattgatctga

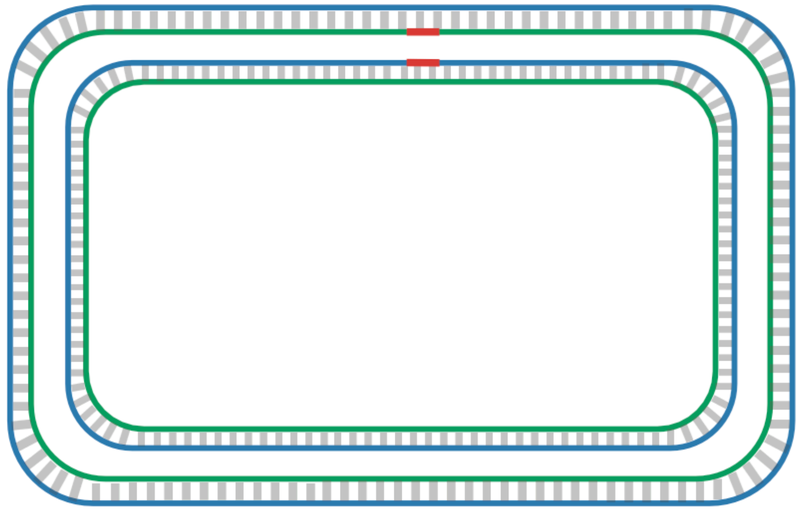
Clump Finding Problem:
Find patterns forming clumps in a string.

Input: A string *Genome*, and integers k, L, and t.
Output: All distinct k-mers forming (L, t)-clumps in *Genome*.

E. Coli

- Let's look for clumps in the Escherichia coli (E. coli) genome, the workhorse of bacterial genomics. We find hundreds of different 9-mers forming (500, 3)-clumps in the E. coli





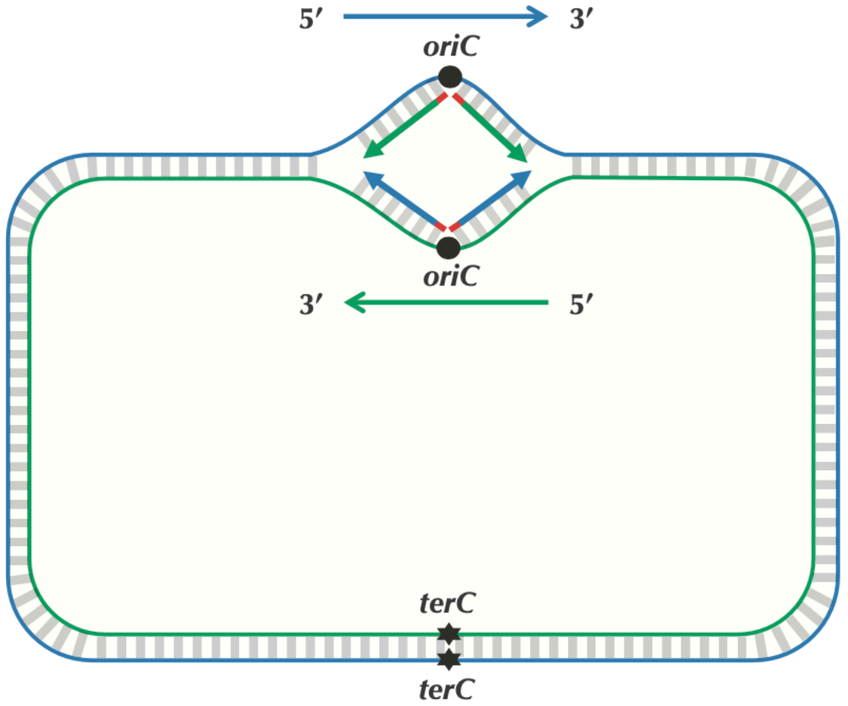
Replication

- As the strands unwind, they create two replication forks, which expand in both directions around the chromosome until the strands completely separate at the replication terminus (denoted terC). The replication terminus is located roughly opposite to oriC in the chromosome.

- An important thing to know about replication is that a DNA polymerase does not wait for the two parent strands to completely separate before initiating replication; instead, it starts copying while the strands are unraveling.

- Thus, just four DNA poly- merases, each responsible for one half-strand, can all start at oriC and replicate the entire chromosome.

- To start replication, a DNA polymerase needs a primer, a short complementary segment (shown in red in Figure 1.5) that binds to the parent strand and jump starts the DNA polymerase.

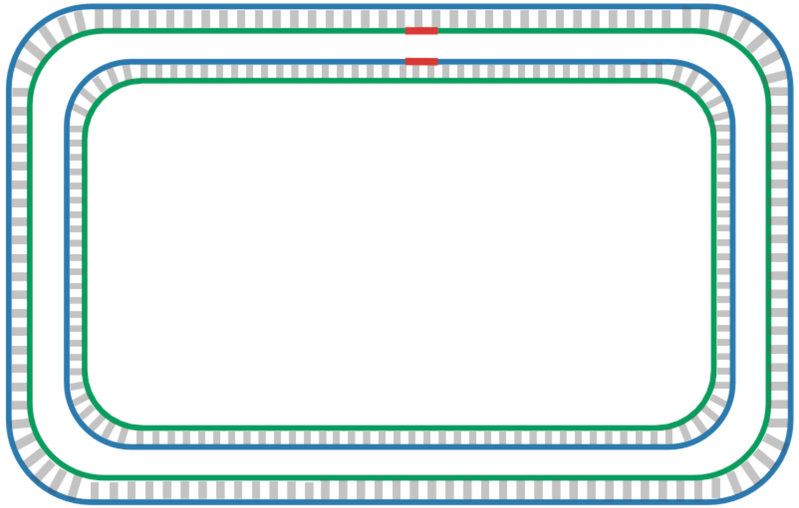


Replication

- After the strands start separating, each of the four DNA polymerases starts replication by adding nucleotides, beginning with the primer and proceeding around the chromosome from oriC to terC in either the clockwise or counterclockwise direction.

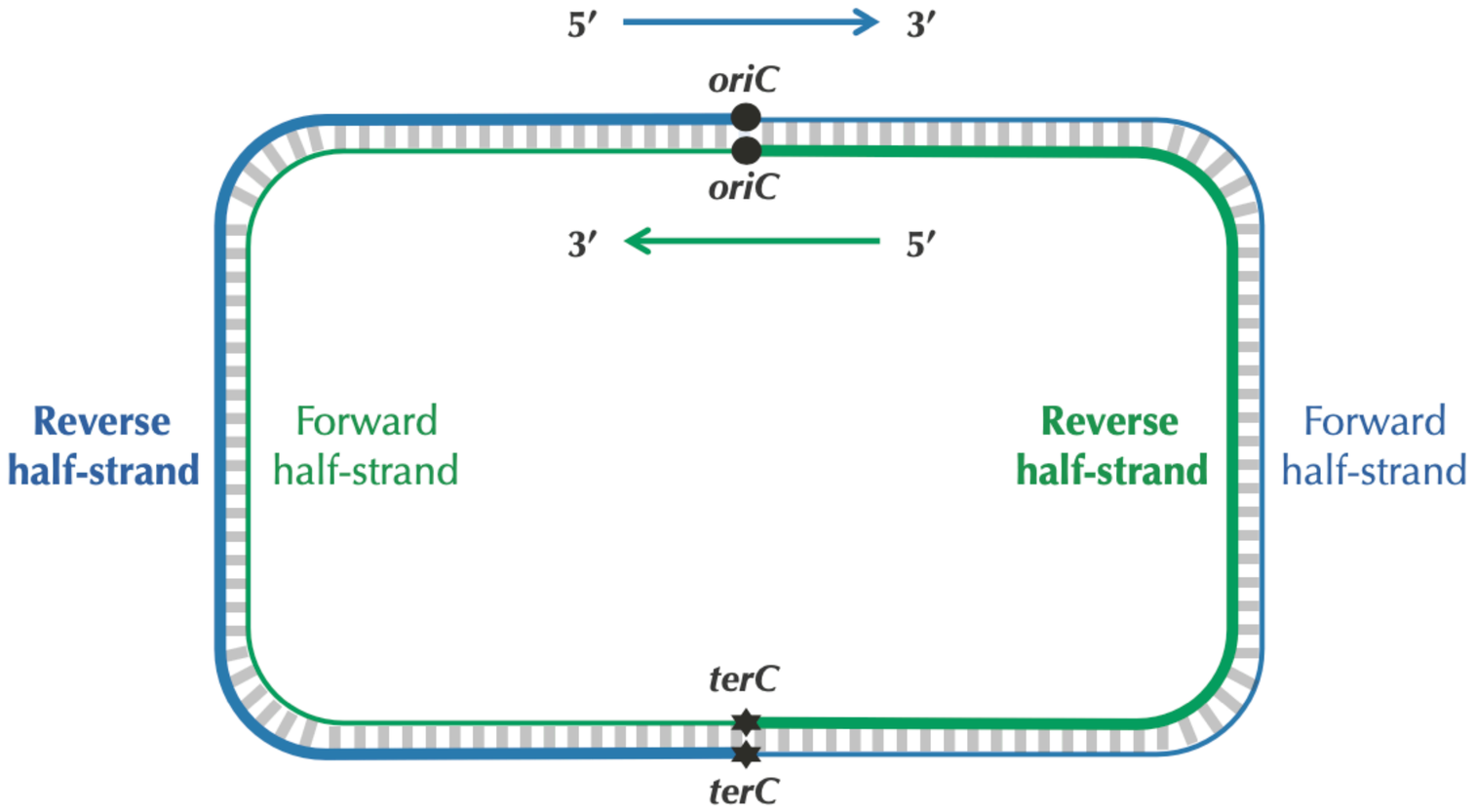
- When all four DNA polymerases have reached terC, the chromosome's DNA will have been completely replicated, resulting in two pairs of complementary strands, and the cell is ready to divide.

- DNA polymerases are unidirectional, meaning that they can only traverse a template strand of DNA in the 3' -> 5' direction.



Forward and Reverse

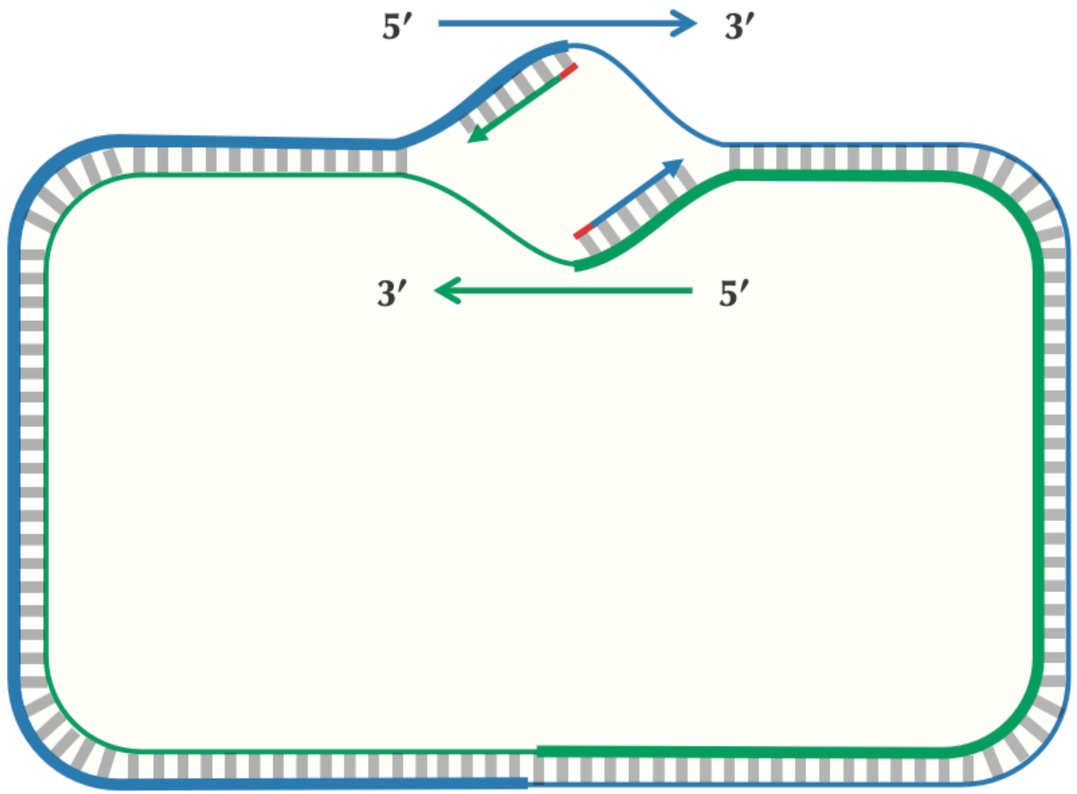
- Imagine that you decided to walk along DNA from oriC to terC.
- There are four different half-strands of parent DNA connecting oriC to terC,



Asymmetric

- On a forward half-strand, in order to replicate DNA, a DNA polymerase must wait for the replication fork to open a little (approximately 2,000 nucleotides) until a new primer is formed at the end of the replication fork;

- Afterwards, the DNA polymerase starts replicating a small chunk of DNA starting from this primer and moving backward in the direction of oriC.

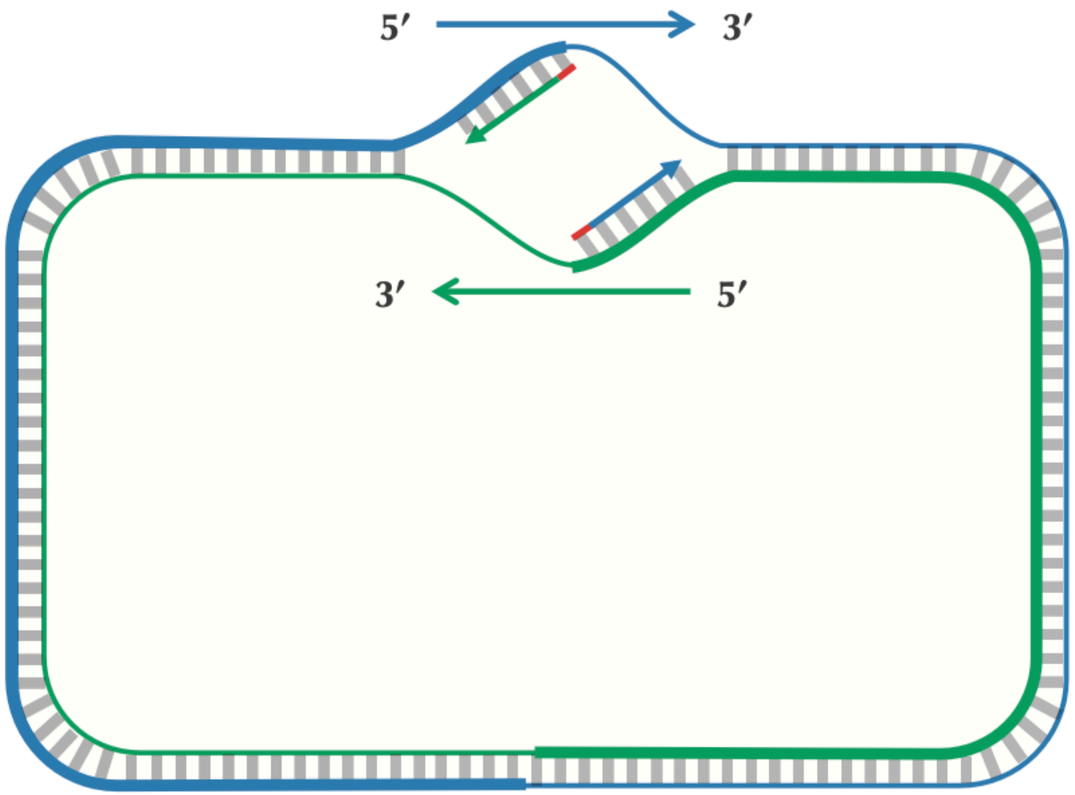


Stopping and Starting

- After this point, replication on each reverse half-strand progresses continuously; however, a DNA polymerase on a forward half-strand has no choice but to wait again until the replication fork has opened another 2,000 nucleotides or so.

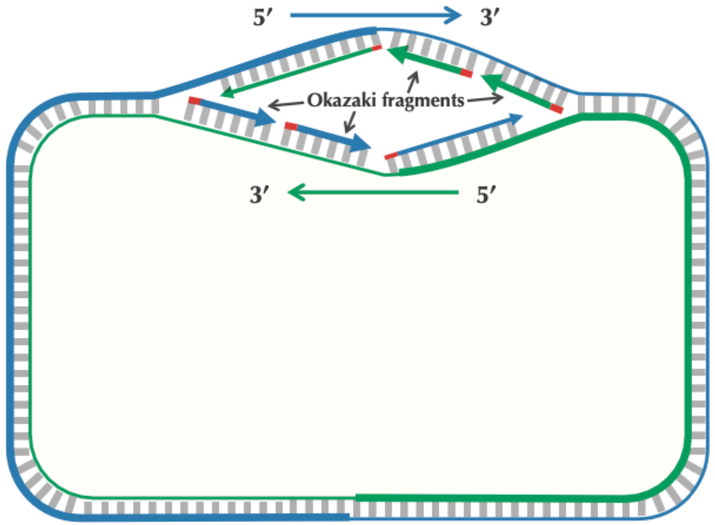
- It then requires a new primer to begin synthesizing another fragment back toward oriC.

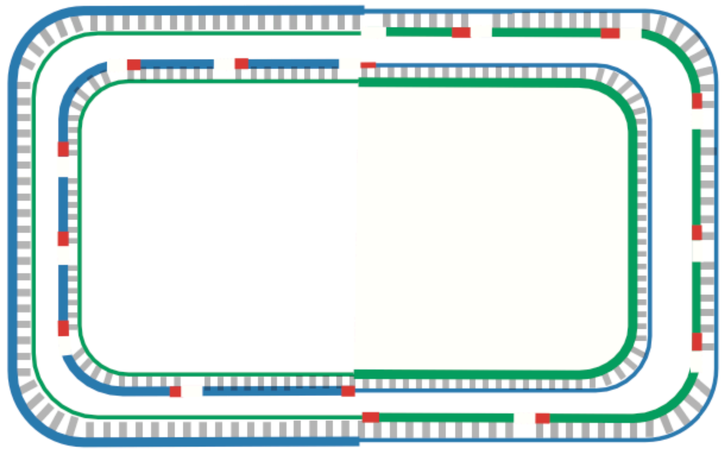
- On the whole, replication on a forward half-strand requires occasional stopping and restarting results in the synthesis of short Okazaki fragments that are complementary to intervals on the forward half-strand.

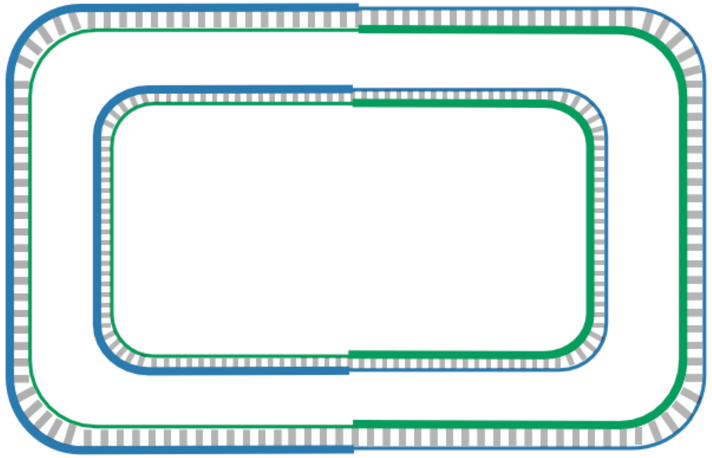


Okazaki Fragments

- Finally, consecutive Okazaki fragments are sewn together by an enzyme called DNA
- ligase, resulting in two intact daughter chromosomes, each consisting of one parent strand and one newly synthesized daughter strand



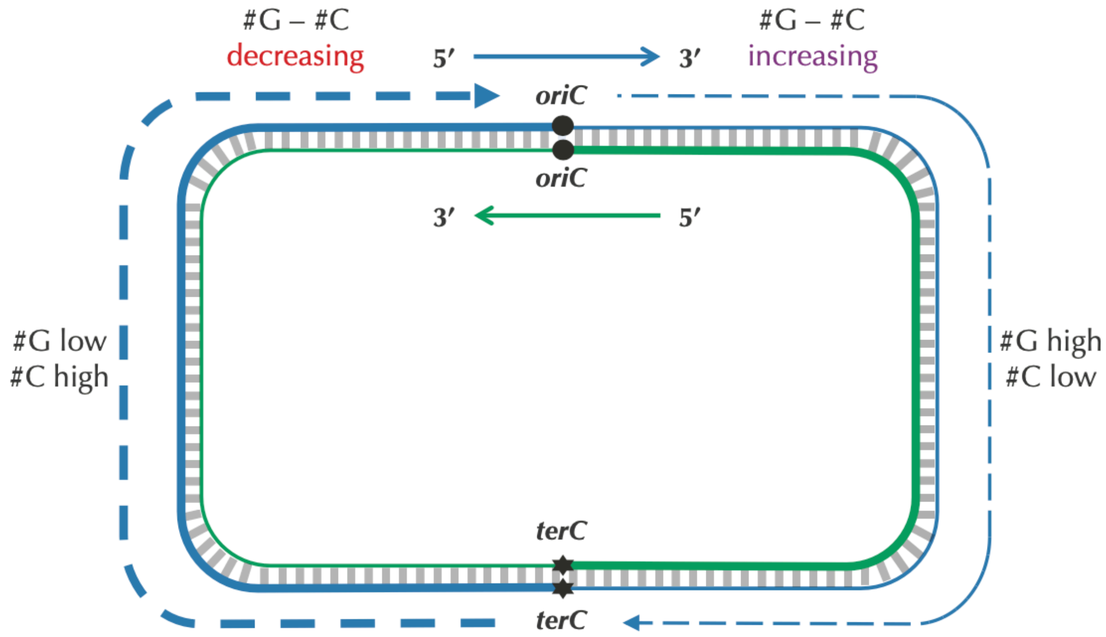




Deamination

- Cytosine (C) has a tendency to mutate into thymine (T) through a process called deamination.
- Thermotoga petrophila

	#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



The skew diagram

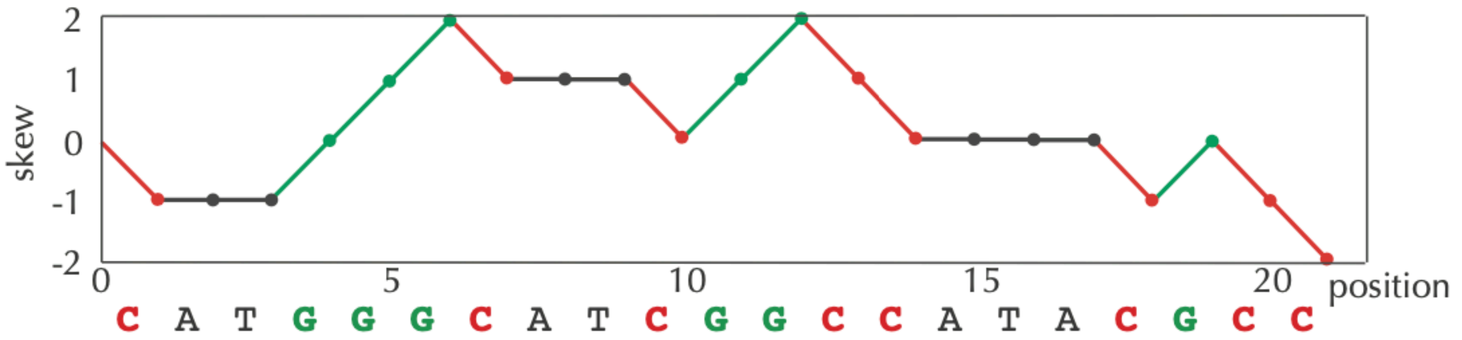


FIGURE 1.12 The skew diagram for *Genome* = CATGGGCATCGGCCATACGCC.

The skew diagram

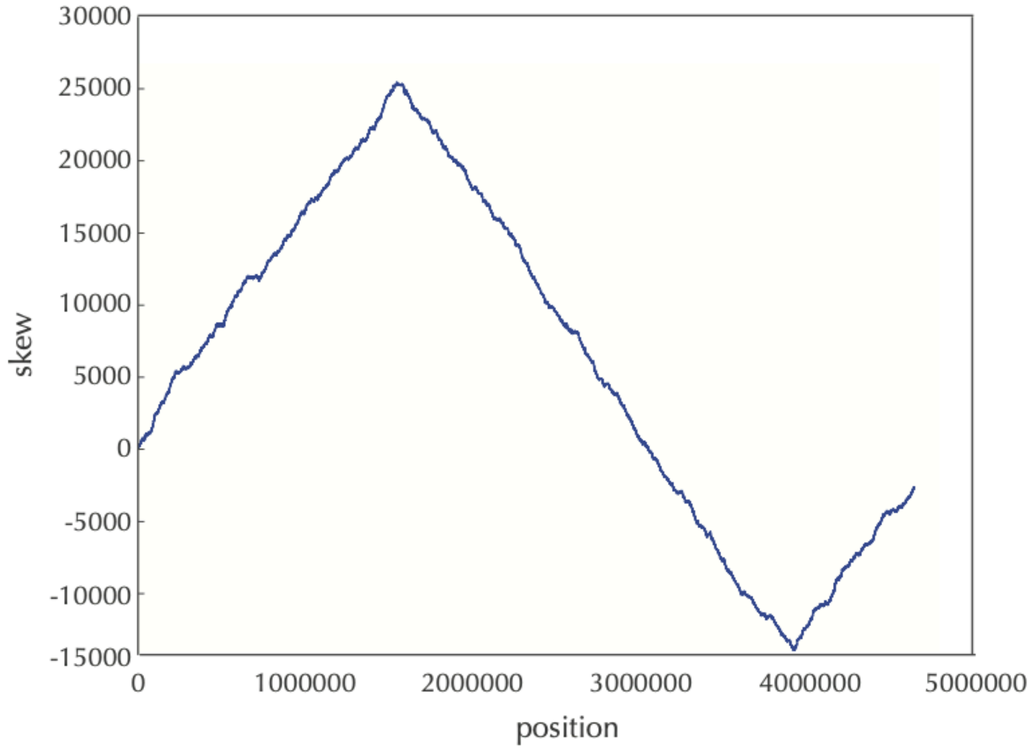


FIGURE 1.13 The skew diagram for *E. coli* achieves a maximum and minimum at positions 1550413 and 3923620, respectively.

E. Coli

- No string matches!

```
aatgatgatgacgtcaaaaggatccggataaaacatggtgattgcctcgcataacgcggt
atgaaaatggattgaagcccgggccgtggattctactcaactttgtcggcttgagaaaga
cctgggatcctgggtattaaaaagaagatctatttatttagagatctgttctattgtgat
ctcttattaggatcgcactgccctgtggataacaaggatccggcttttaagatcaacaac
ctggaaaggatcattaactgtgaatgatcggtgatcctggaccgtataagctgggatcag
aatgaggggttatacacaactcaaaaactgaacaacagttgttctttggataactaccgg
ttgatccaagcttcctgacagagttatccacagtagatcgcacgatctgtatacttattt
gagtaaattaacccacgatcccagccattcttctgccggatcttccggaatgtcgtgatc
aagaatgttgatcttcagtg
```

Approximate matches

- No string matches!