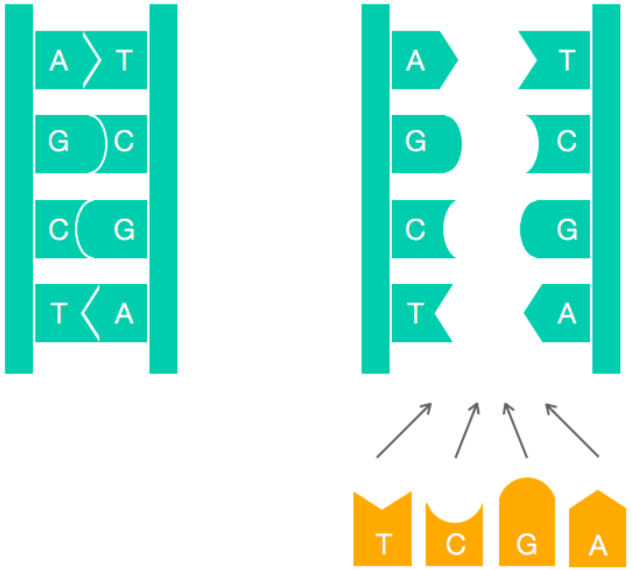
**Qualitative biological question**

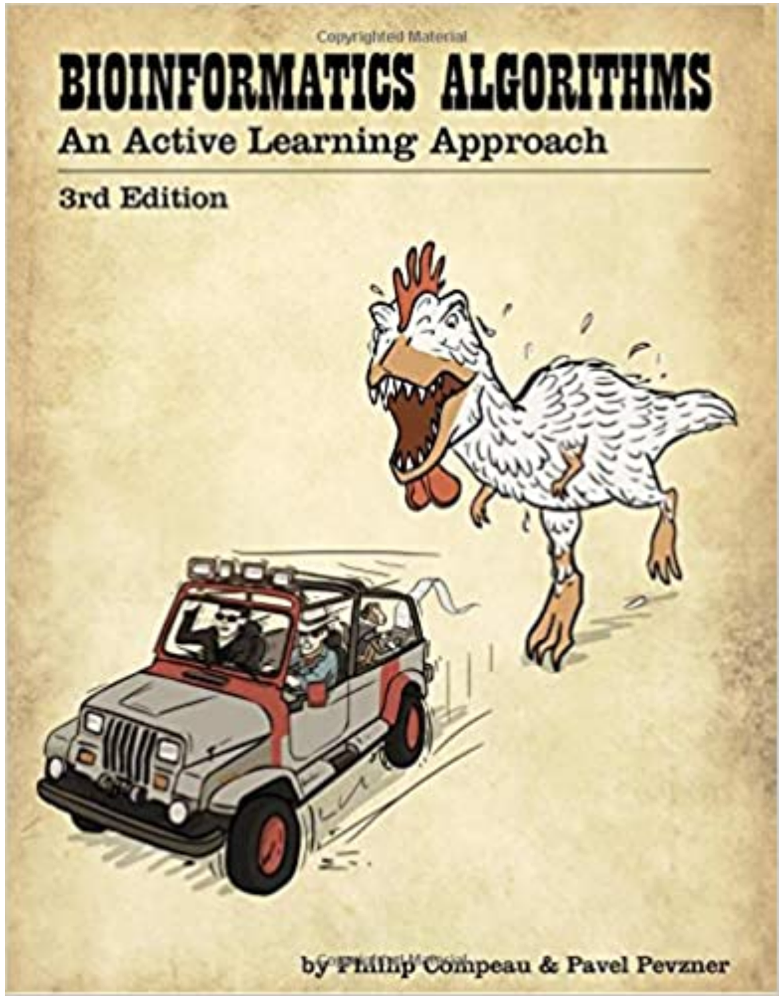
Not well-defined
Hard to answer

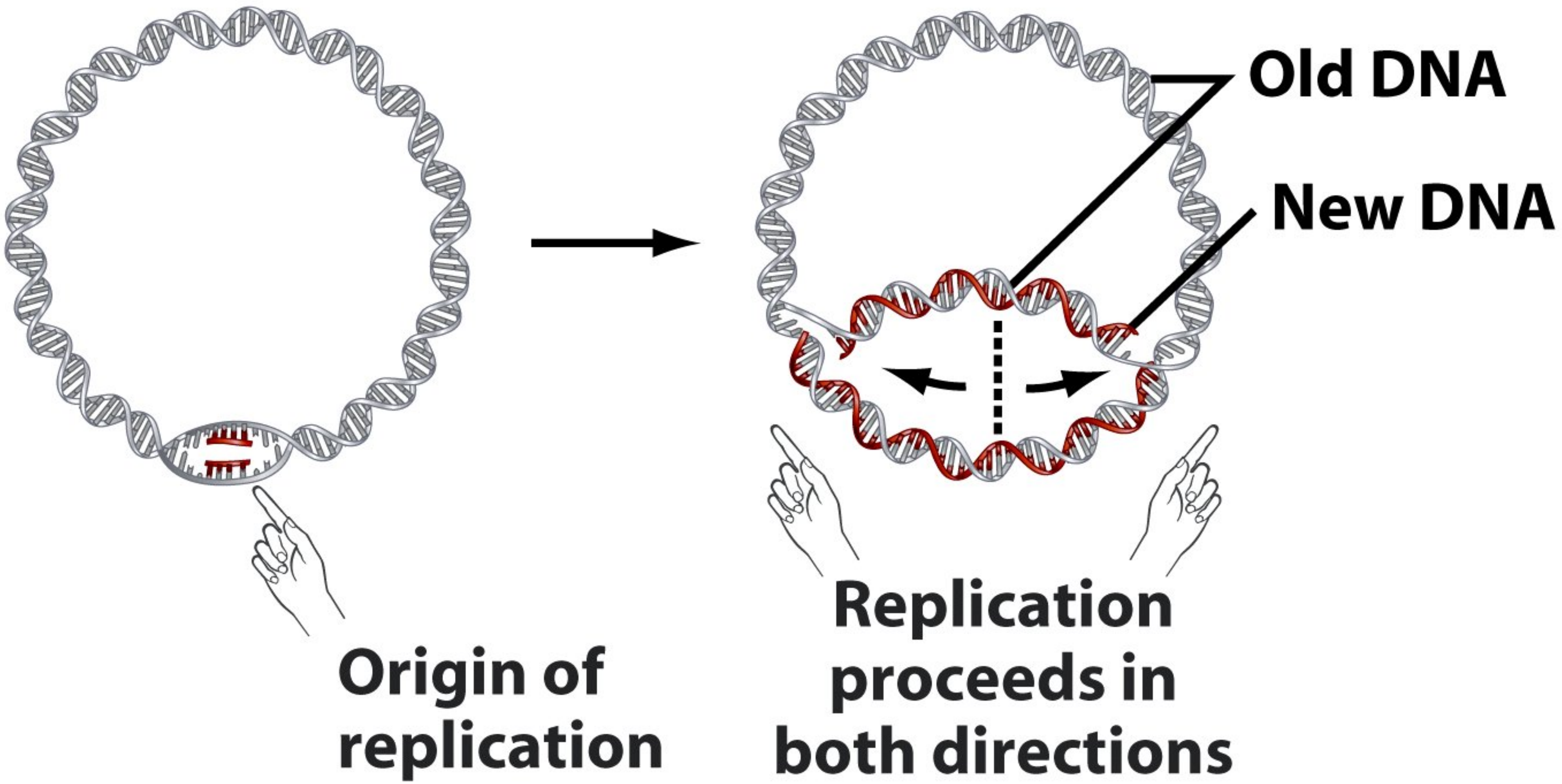
↓

**Quantitative question**

# Where in the genome does the DNA replication begin?




Copyrighted Material
BIOINFORMATICS ALGORITHMS
An Active Learning Approach
3rd Edition
by Phillip Compeau & Pavel Pevzner



Old DNA

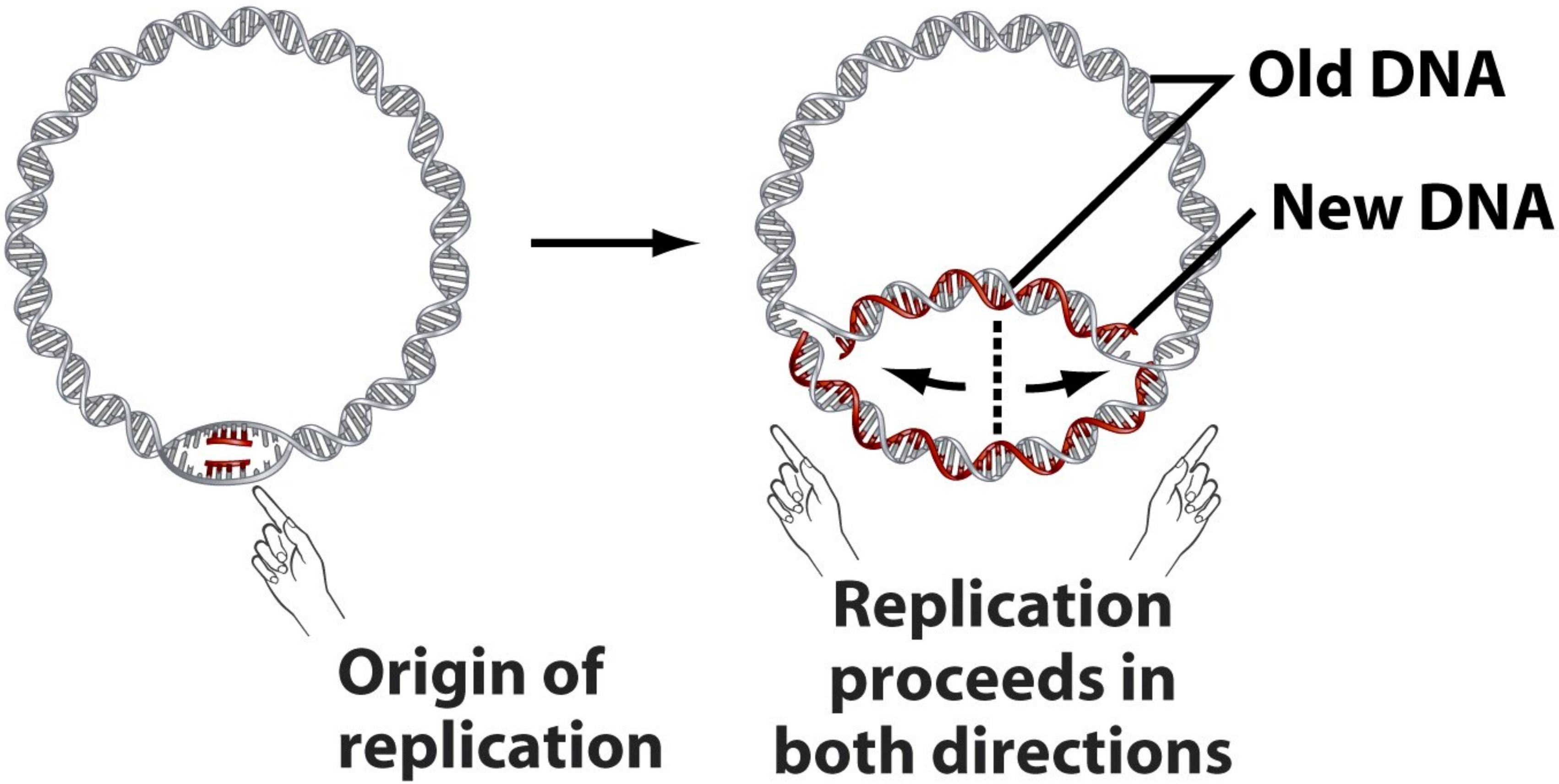
New DNA

Origin of replication

Replication proceeds in both directions

oriC

GATAATCGTATAGTTATCCACATGAGATTTGATTGAAAAAGCATCAATCAATTTTTTCACTACCGTTAAATTTATCCACA
ATCCAAAAAAAAGAGCGGCATTAAGCCGCTCTGCATGGAATAGGTCATTATTTAGAAGCGATTGATGACGCGTTTGAGCC
AAGCTTCAGCGGCATCTTCAGGCACTGGGTGCTCTTGTACATCGATGGTAAAGCAGTTGGCCAGAGGTTTAGCACCAATA
TCCCCCAGCAGCTGATAGGCATGTTTACCTGCCGCGCAGAAAGTATCGTAGCTTGAATCACCAATCGCGACCACGGCATA
ACGTAGTGCAGAGGTATTCGGTGGTGTATTCTGCAGAGCCTGAATAAAGGGCTGGATATTATCCGGGTACTCACCAGCCC
CGTGGGTTGAGGTGATGATCAGCCAAGTCCCTTTAGCAGGGATCTCACTCATGTTGGGCTGGTTATGAATTTTGGTGTCA
AAGCCTTGTTCTTGCAGTAAATCACTCAGGTGGTCACCCACATATTCCGCACCGCCTAGGGTGCTGCCAGTAATGATATG
AATCATAGCGTTACTCTATTTCCCAATACAGAATGATGAAAAAATGCGGCCAAGCAGATCATCGGAGCTGAACTCGCCCG
TAATTTCGTTAAGGTGTTGCTGGGCTATACGCAGCTCTTCGGCGAGGATTTCTCCGGCCATATAGCCTTCAAGTTGTTGC
TGGCCAATCGCTAAGTGCTCTGCGGCTCGCTCTAGGGCATCGAGATGACGGCGGTGCCATAAAGCCACCTTCCTGATT
GCCTGAAAAACCCATGCACTCTTTGAGGTGCTGACGCAAGGCATCGACCCCTTGGCCTGTTTTGGCTGATAGGCGGATCA
AGGTGGGTTGATTAACATGGCAGATCCCAAGGGGCTCACCAGTTTGATCGGCTTTATTACGGATCACAGTGATCCCAATA
TTCTCTGGCAGTTTGTCAACAAAATCAGGCCAGATGTCCTGTGGATCGGTGGCCTCTGTGGTGGTGCCATCGACCATAAA
~~[illegible]~~
~~GTCCCGCACTATCGATGATGTGCAGCGGCATCCCATCAATATGGATATGCTCACGCAGAACATCACGGGTCCTACCGGCA~~
~~ATGTCGGTAACGATGGCAGACTCTTTACCTGAAAGCGCATTGAGTAGGCTCGATTTACCCGCATTAGGACGCCCAGCAAT~~
~~CACCACCTTCATCCCTTCGCGCATAATGGCGCCTTGGTTGGCTTCACGGCGCACTGCGGCAAGATTATCTATGATGGTTT~~
~~GCAGATCAGCGGAAACCTTACCATCGGCCAGAAATCGATCTCTTCTTTCTGGGAAATCAATTGCGGCTTCAACATAGATG~~
~~CGCAGGTGAATCAGCGATTCCACCAAGGTATGGATGCCTTTACAAAACTCGCCTTGCAGTGATTGCAGCGCGGATTTCGC~~
~~GGCTTGCTCAGAGCTGGCATCAATCAGGTCTGCGATGGCTTCCGCTTGGGTTAAATCCATCTTGTCATTGAGGAAAGCGC~~
~~GTTCTGAGAATTCACCGGGACGGGCTGGGCGCACTCCTTTAATCTGCAAAATACGGCGGATCAGCATATCCATGACGACC~~
~~GGGCCACCGTGACCTTGCAGCTCAAGCACATCTTCACCGGTAAATGAATGAGGATTGGGGAAAAACAGCGCAATGCCTTG~~
~~ATCCAGCTGTTGGCCATCTTCATCGGTGAAAGGCAAGTATTCGGCATAGCGGGGTCTGAGCGTGCGTCCAGTGACTCTCT~~
GCGCGACGTGGGCAGCCAGTGGGCCTGATACACGAATAATGCCGACACCACCACGGCCGGGTGCCGTAGCTTGCGCGACA
ATGGTATCTGTTGTCATAGTGTTACCTGAACAGGATTGAATTAGCGCCATTGTAATCAGCAGCCAACAAAAAGGCGACCT
TTTGGCCGCCTCTTTATTACTCAATCAAACTTACTTGGAGTGTAAGCCTTTTTTCTCTAGCGCTTTGTAGATCAGCGTTT
GCTGGATTAGCGTTACGATGTTCGACACCAACCAGTACAGAACCAGACCTGATGGGAACCACAGGAAGAAGAAAGTGAAC



oriC

***DnaA***
binds to
***DnaA box***
in oriC

# There may be multiple DnaA boxes in oriC.

Find repeating sequences of letters

Find the sequence of letters that repeats the most

# Algorithm

**k-mer:** A string of letters of length k

**string.count(pattern):** An efficient Python string function that returns the number of repetitions of k-mer **pattern** in **string.**

**Ex)** ‘ACAACTACGATTACTACAGGGACTACT‘**.count(**‘ACTAC’)

# Algorithm

A rigorously defined computational problem:

Find the most frequence ***k-mer*s in a string**

Input: A string **`Seq`**, and an integer **`k`**
Output: All most frequent ***k-mer*s in **`Seq`**

**Time to think!!!**

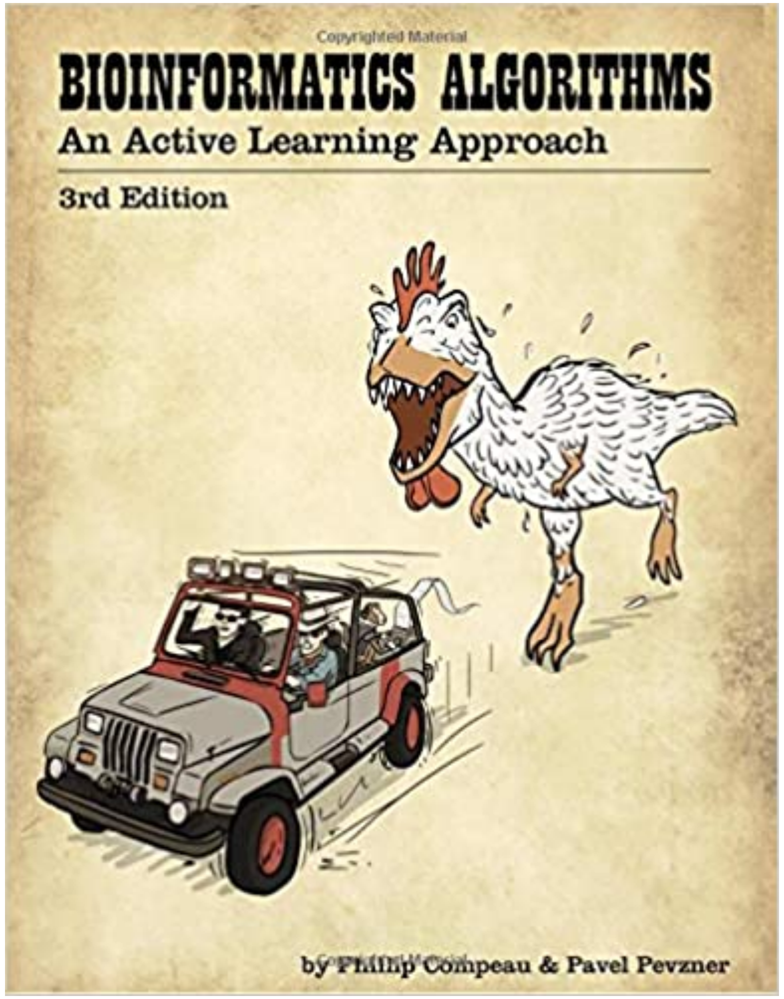
| k | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| count | 1022 | 762 | 302 | 45 | 30 | 9 | 5 |
| k-mers | TGA | ATGA | GATCA<br>TGATC | GTACTA | ATGATCA | ATGATCAA | ATGATCAAG<br>CTTGATCAT<br>TCTTGATCA<br>CTCTTGATC |

***DO NOT RUN THE SHOWN CODE IN THE SERVER!!!***

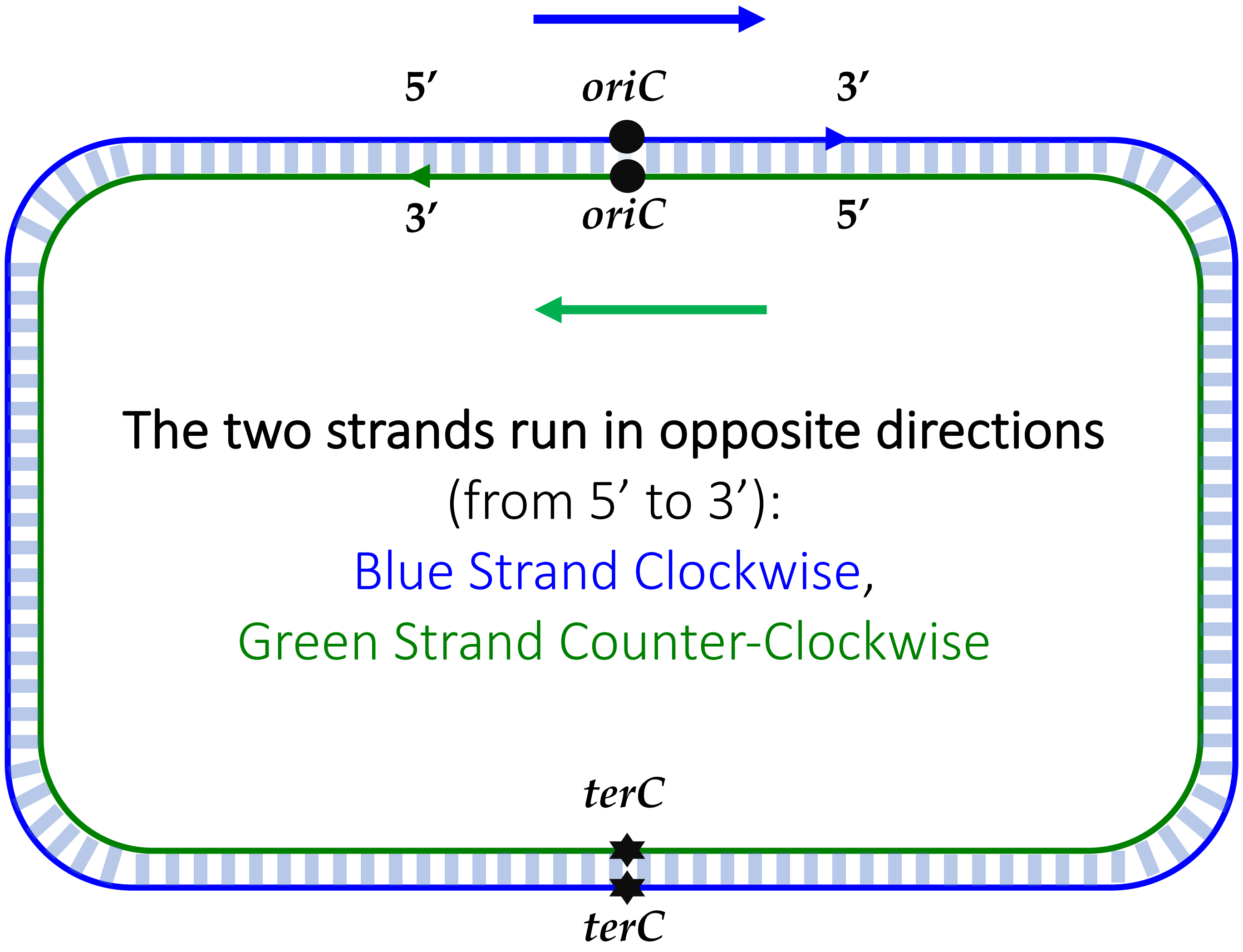
We have a ***limited*** computation time from Google.

You can install Anaconda and try this code in your local machine, ***BUT NOT IN THE SERVER!!!***

# Is there any other insight that we can use to simplify the computation of finding the oriC?


Copyrighted Material
BIOINFORMATICS ALGORITHMS
An Active Learning Approach
3rd Edition
by Phillip Compeau & Pavel Pevzner

# DNA Strands Have Directions!



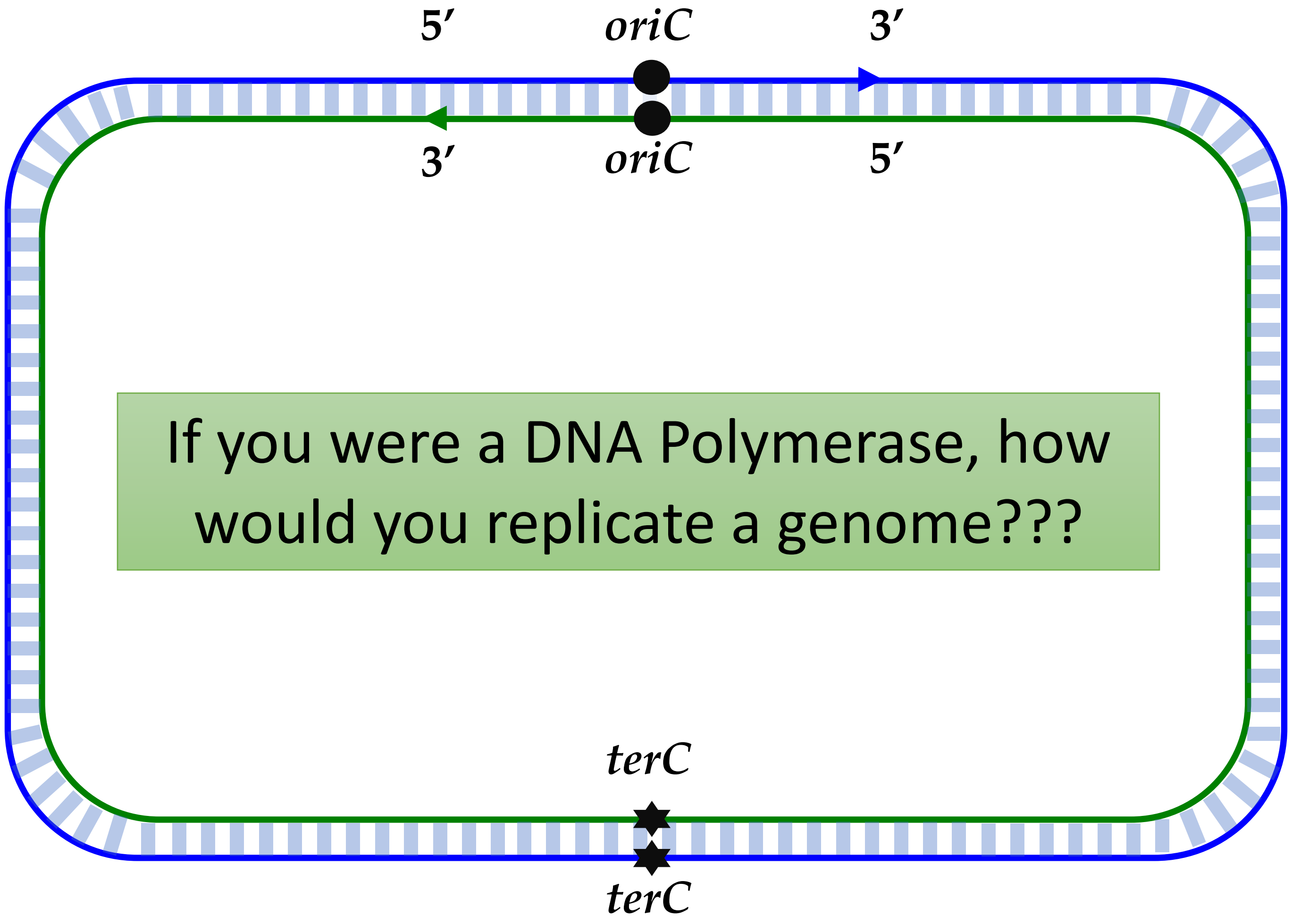
The two strands run in opposite directions

(from 5’ to 3’):

Blue Strand Clockwise,

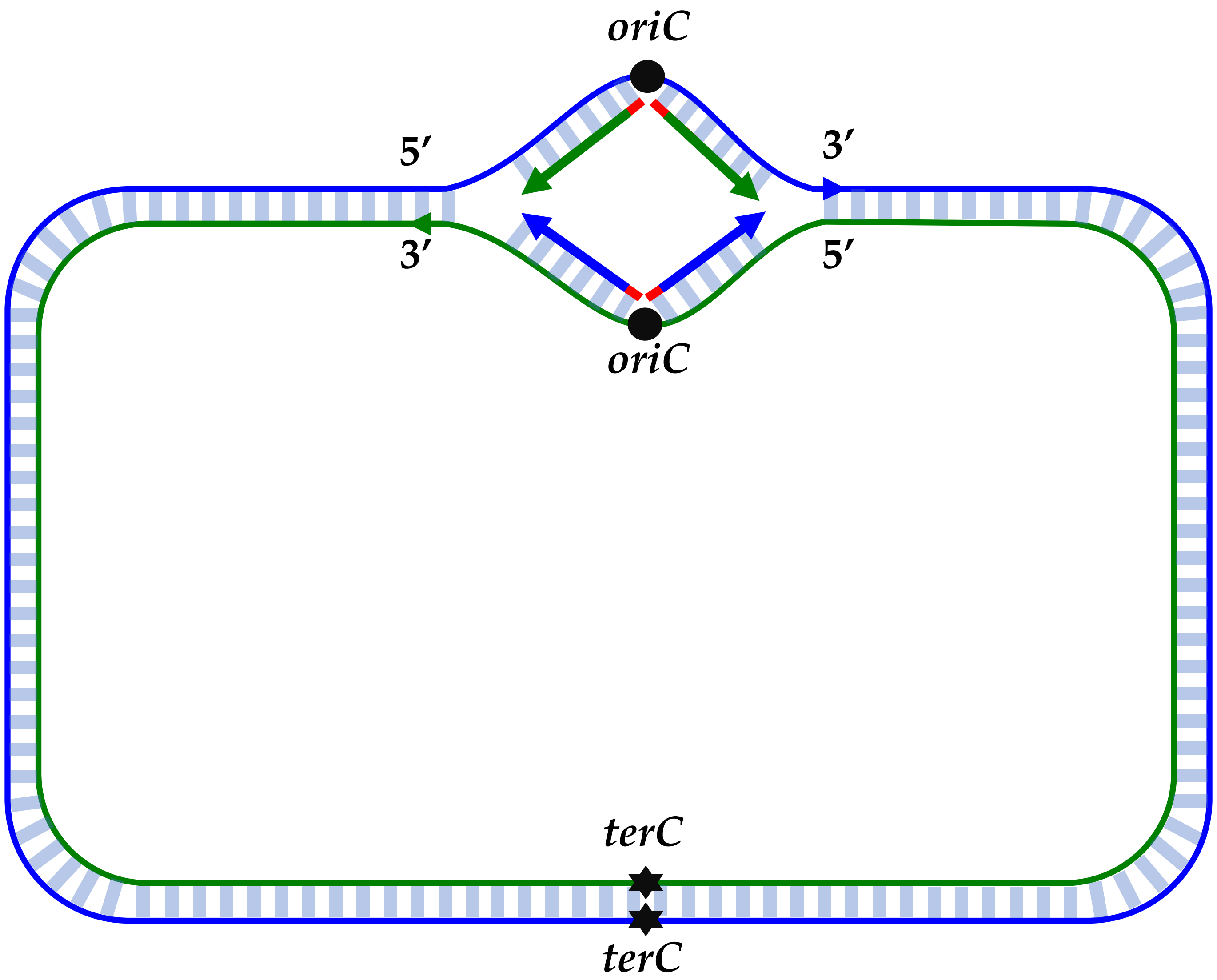
Green Strand Counter-Clockwise

# DNA Strands Have Directions

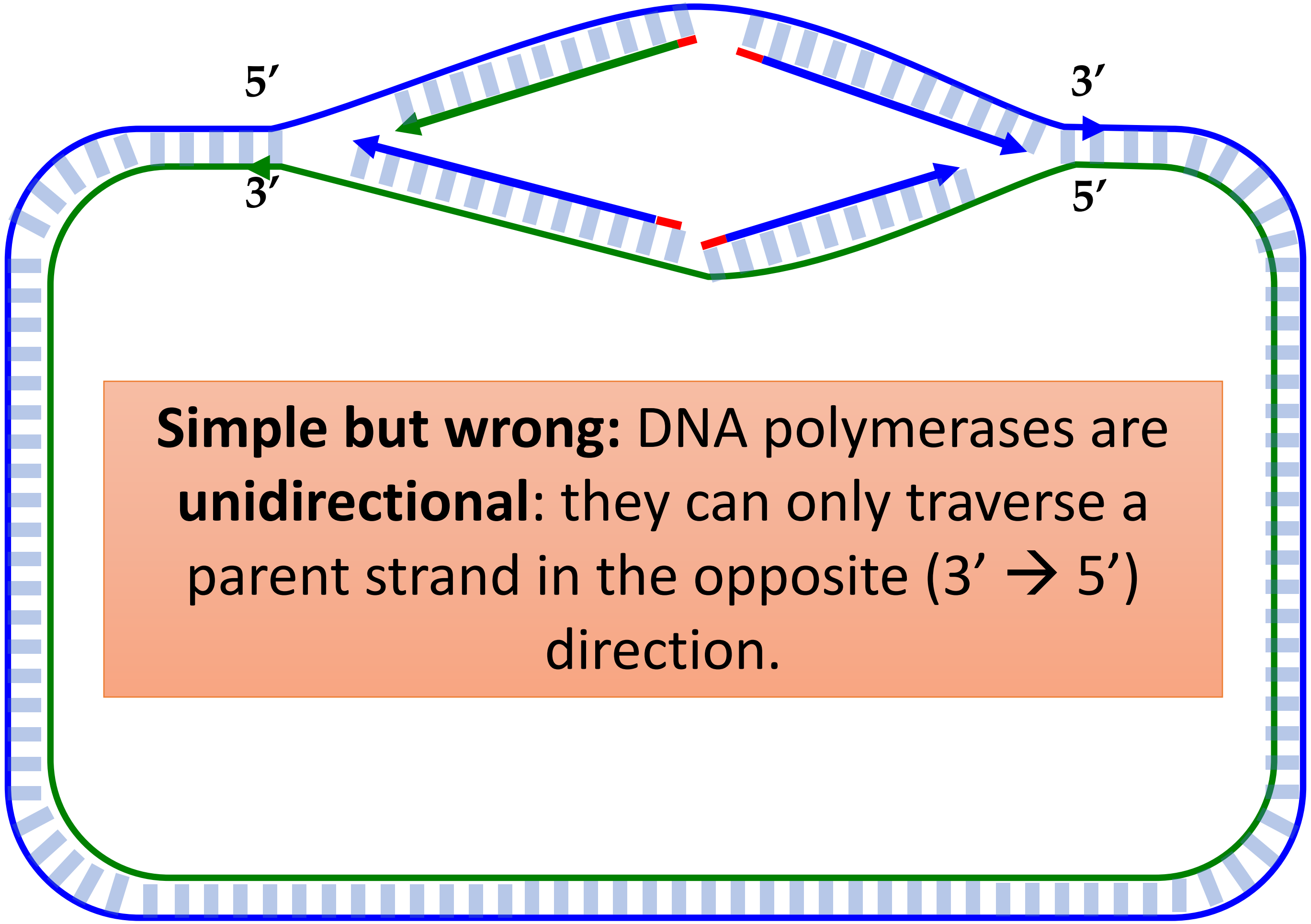


If you were a DNA Polymerase, how would you replicate a genome???

# Four DNA Polymerases Do the Job

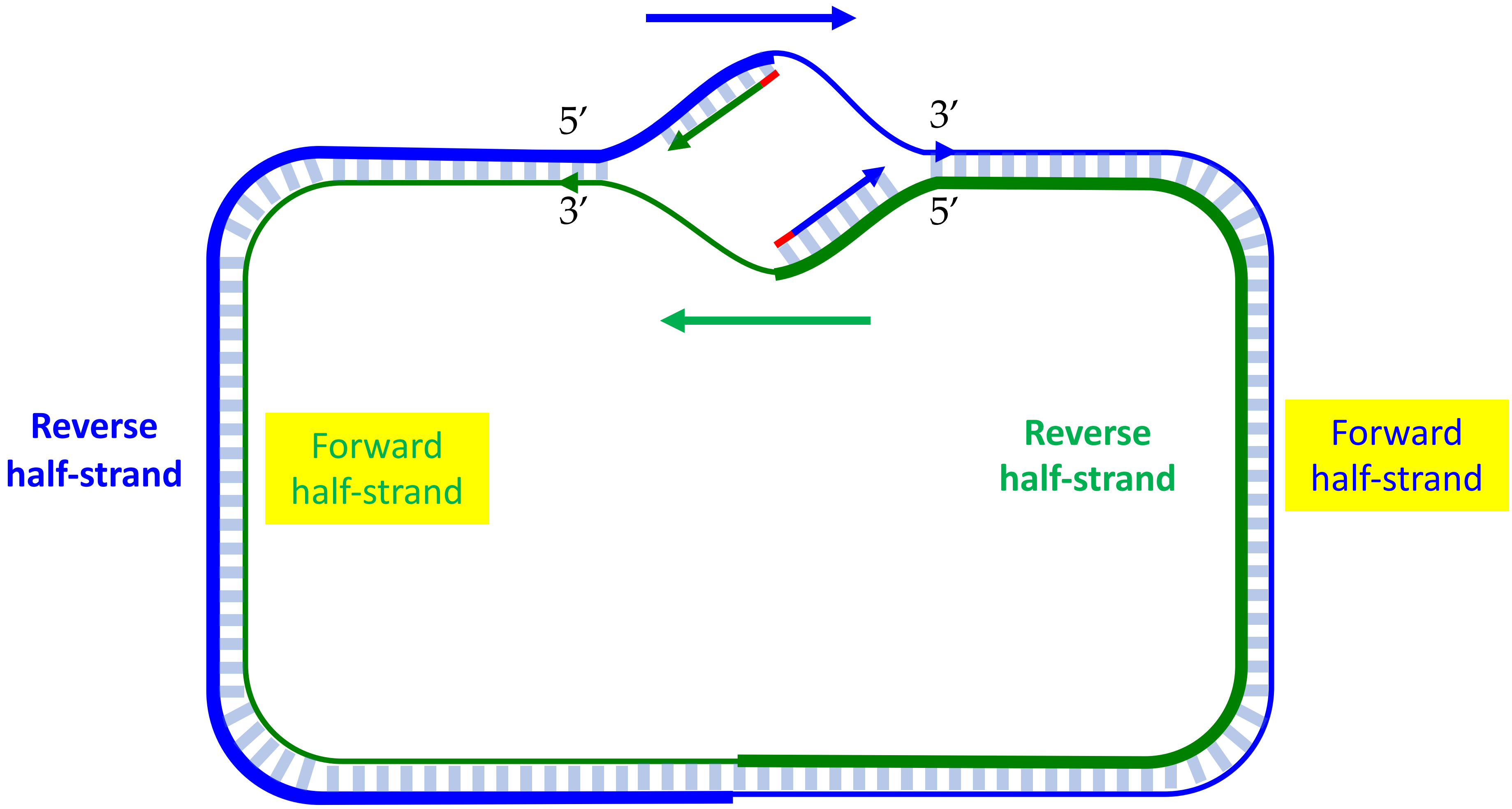


# Continue as Replication Fork Enlarges



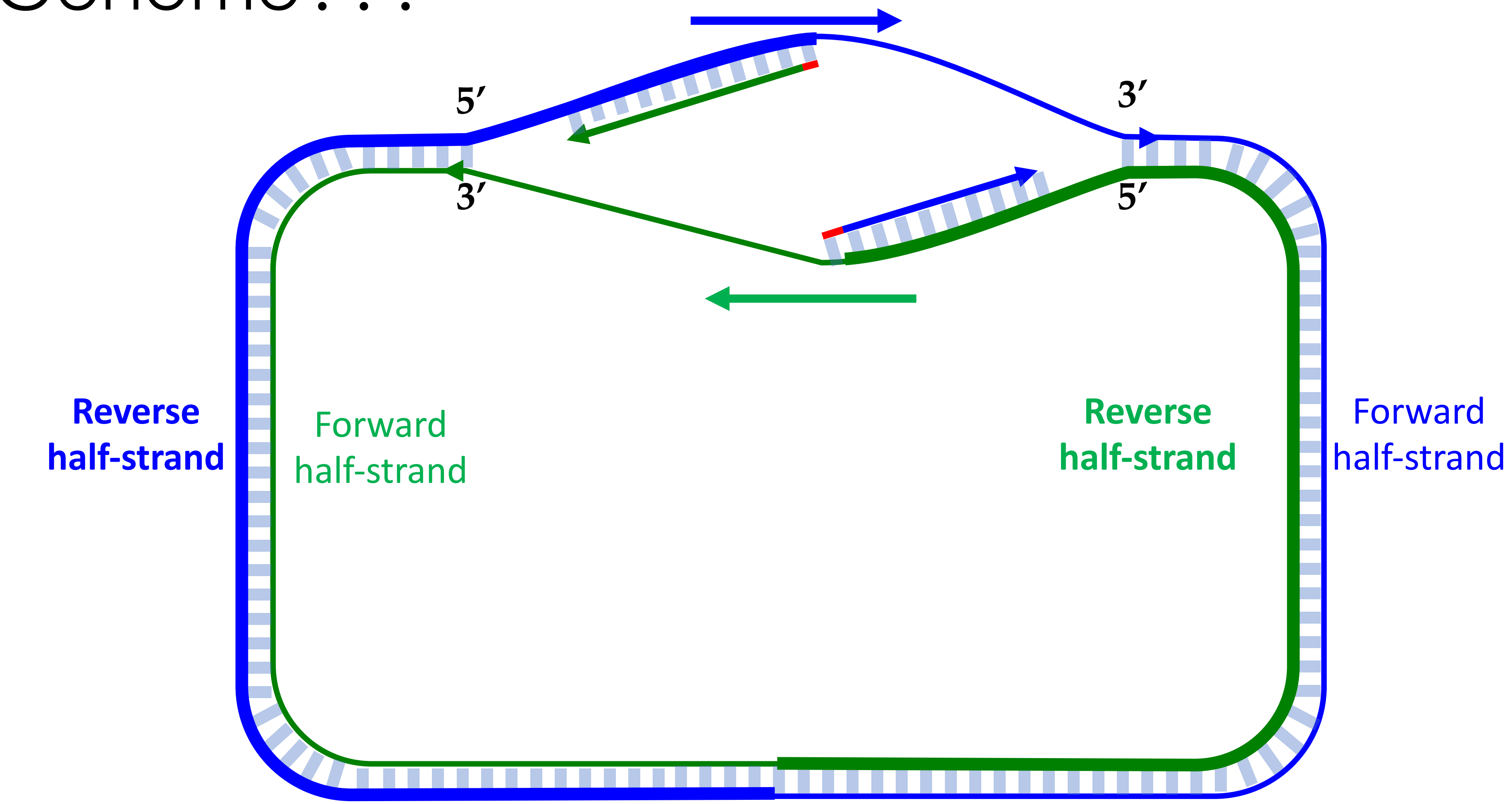
**Simple but wrong:** DNA polymerases are **unidirectional**: they can only traverse a parent strand in the opposite (3’ → 5’) direction.

# If you Were a **UNIDIRECTIONAL** DNA Polymerase, how Would you Replicate a Genome?

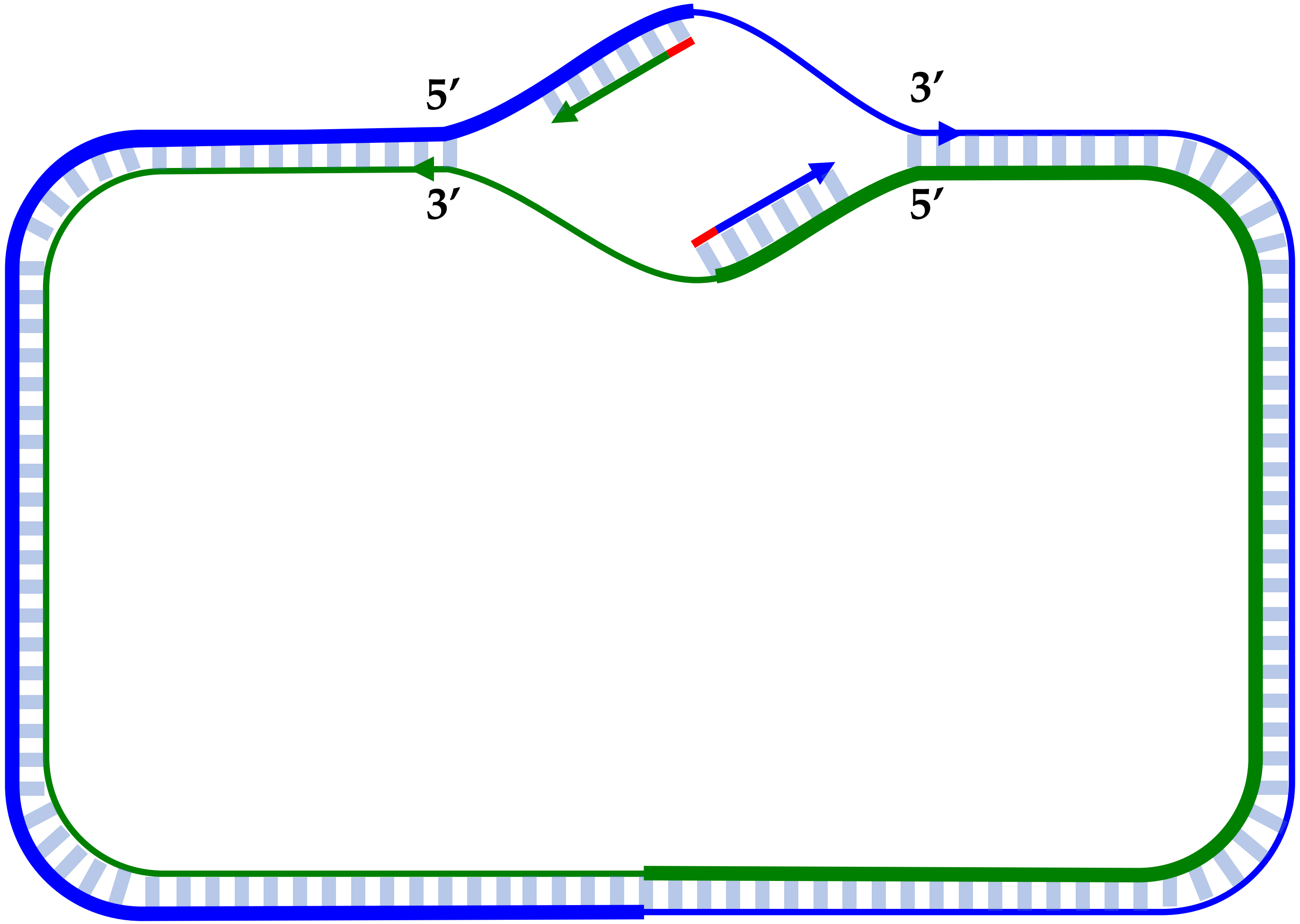


**Big problem replicating forward half-strands (thin lines).**

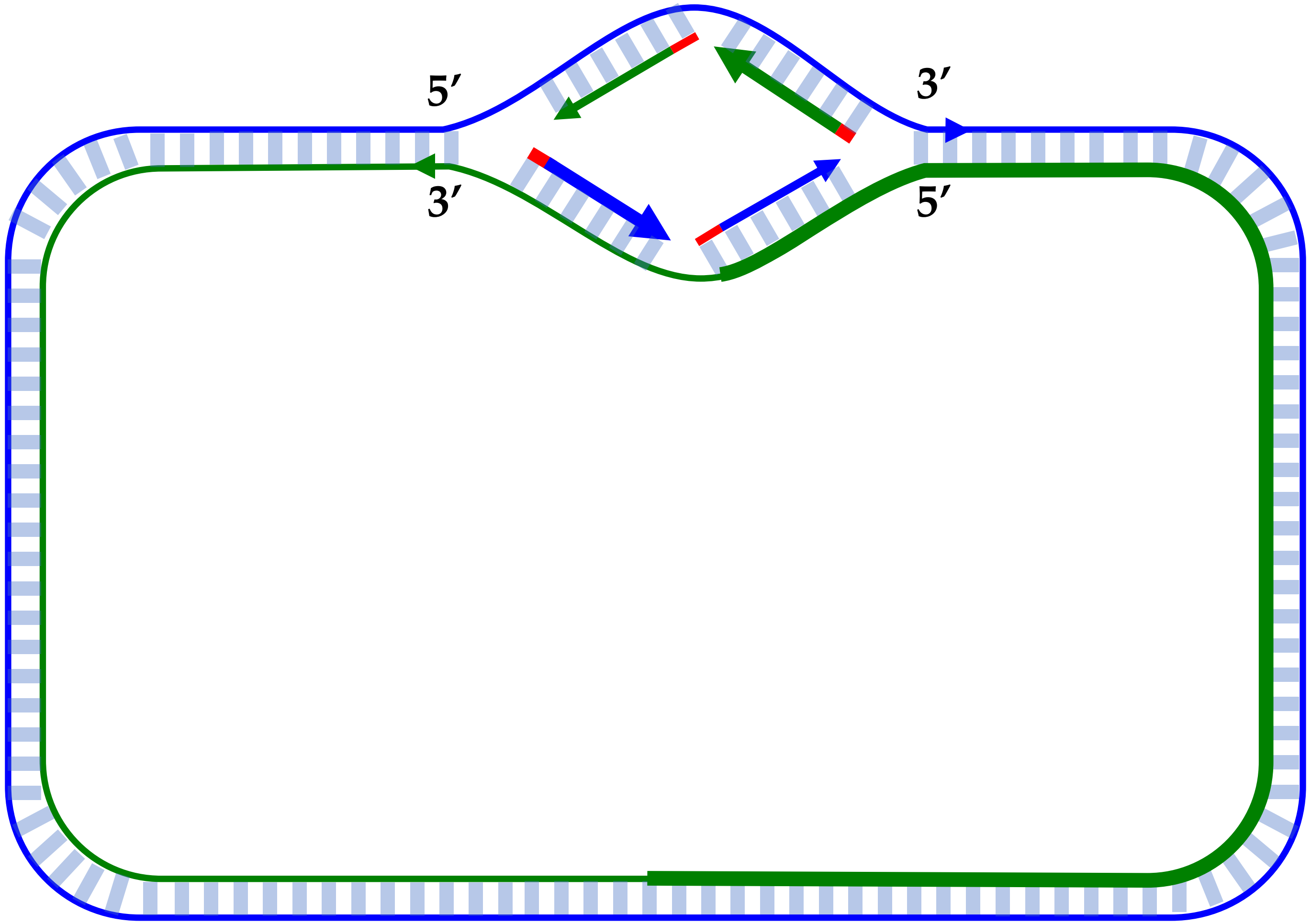
# If you Were a **UNIDIRECTIONAL** DNA Polymerase, How Would you Replicate a Genome???



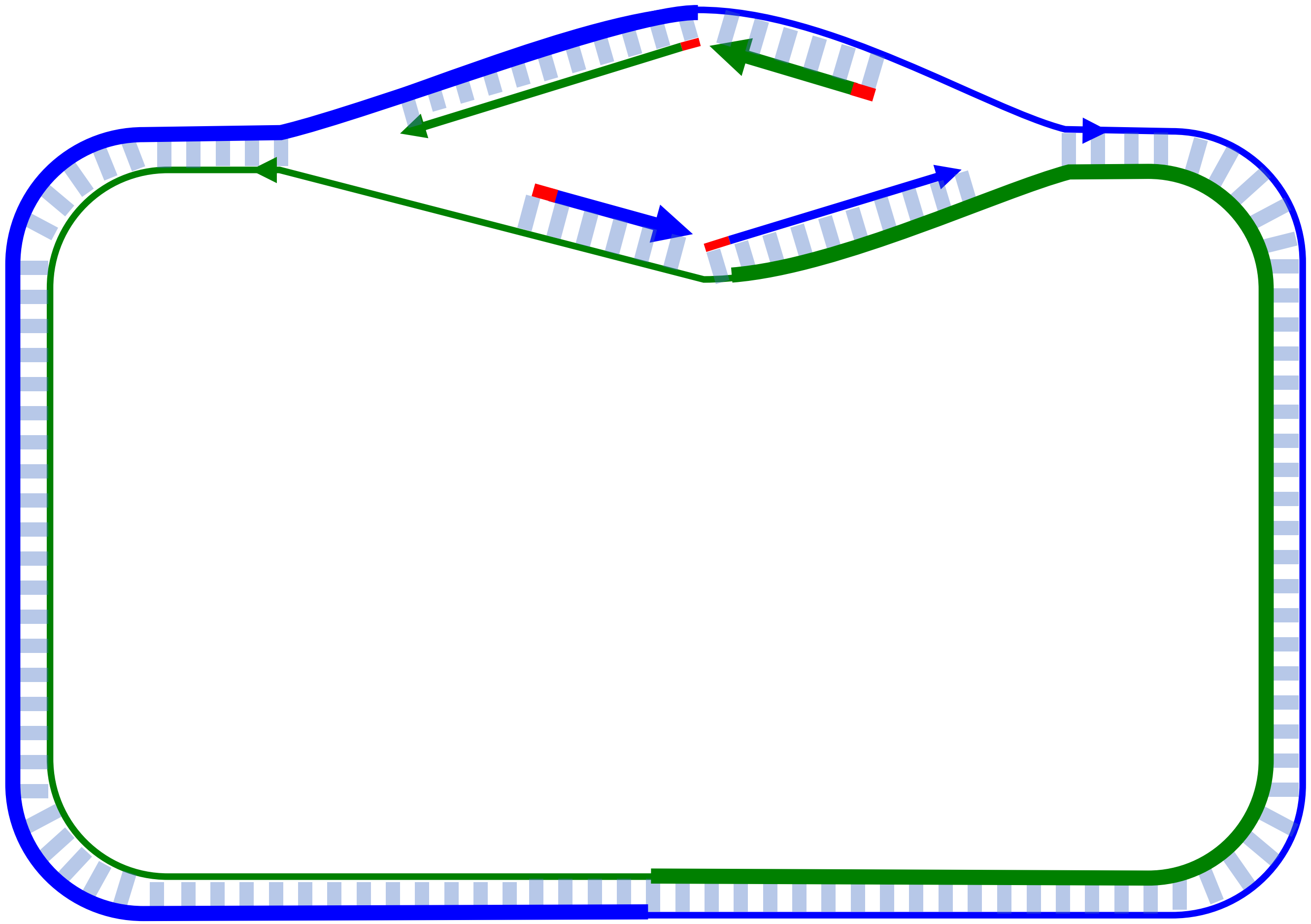
# Wait until the Fork Opens and…



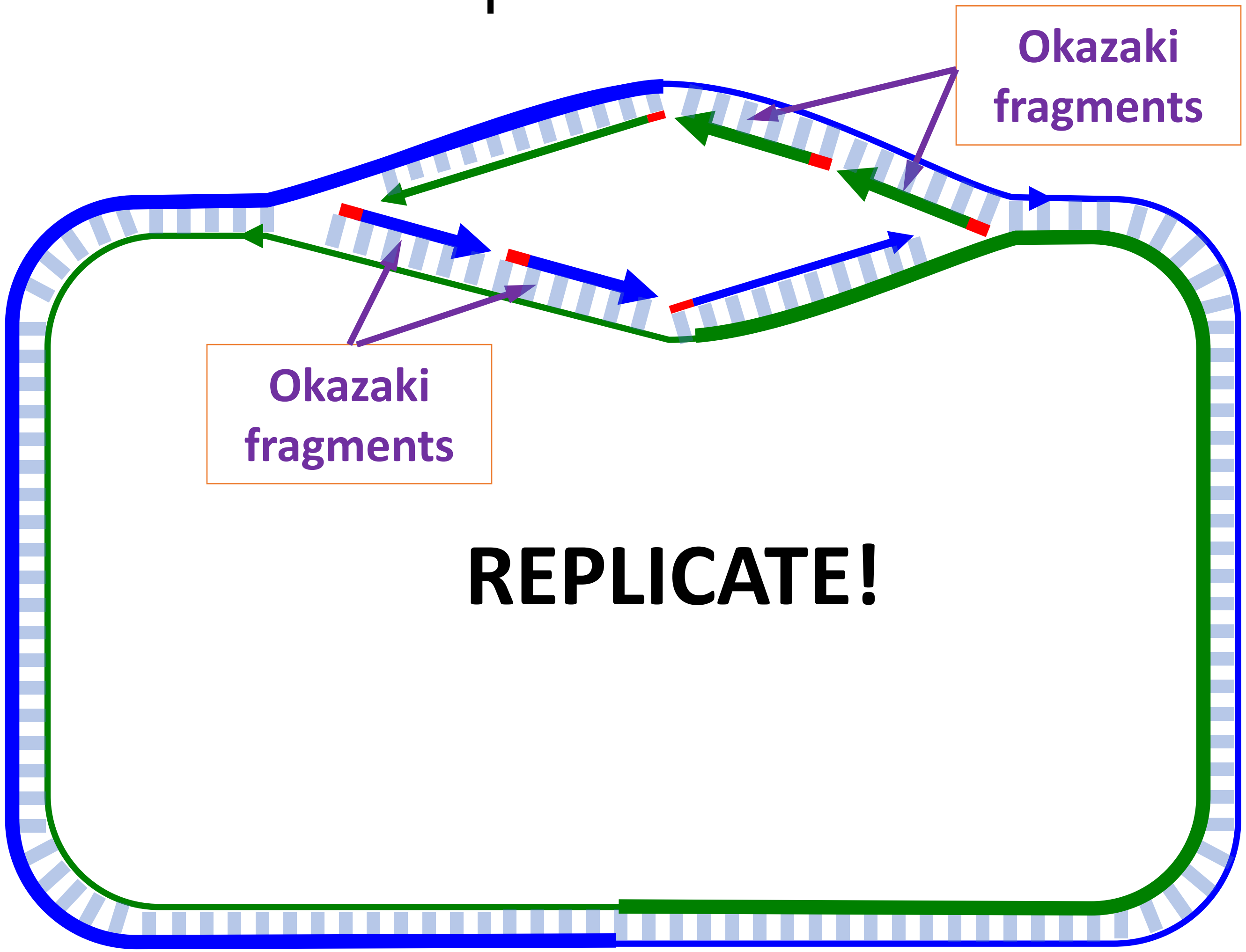
# Wait until the Fork Opens and Replicate



# Wait until the Fork Opens and Replicate Wait until the Fork Opens Even More and…



# Wait until the Fork Opens and Replicate
# Wait until the Fork Opens Even More and…



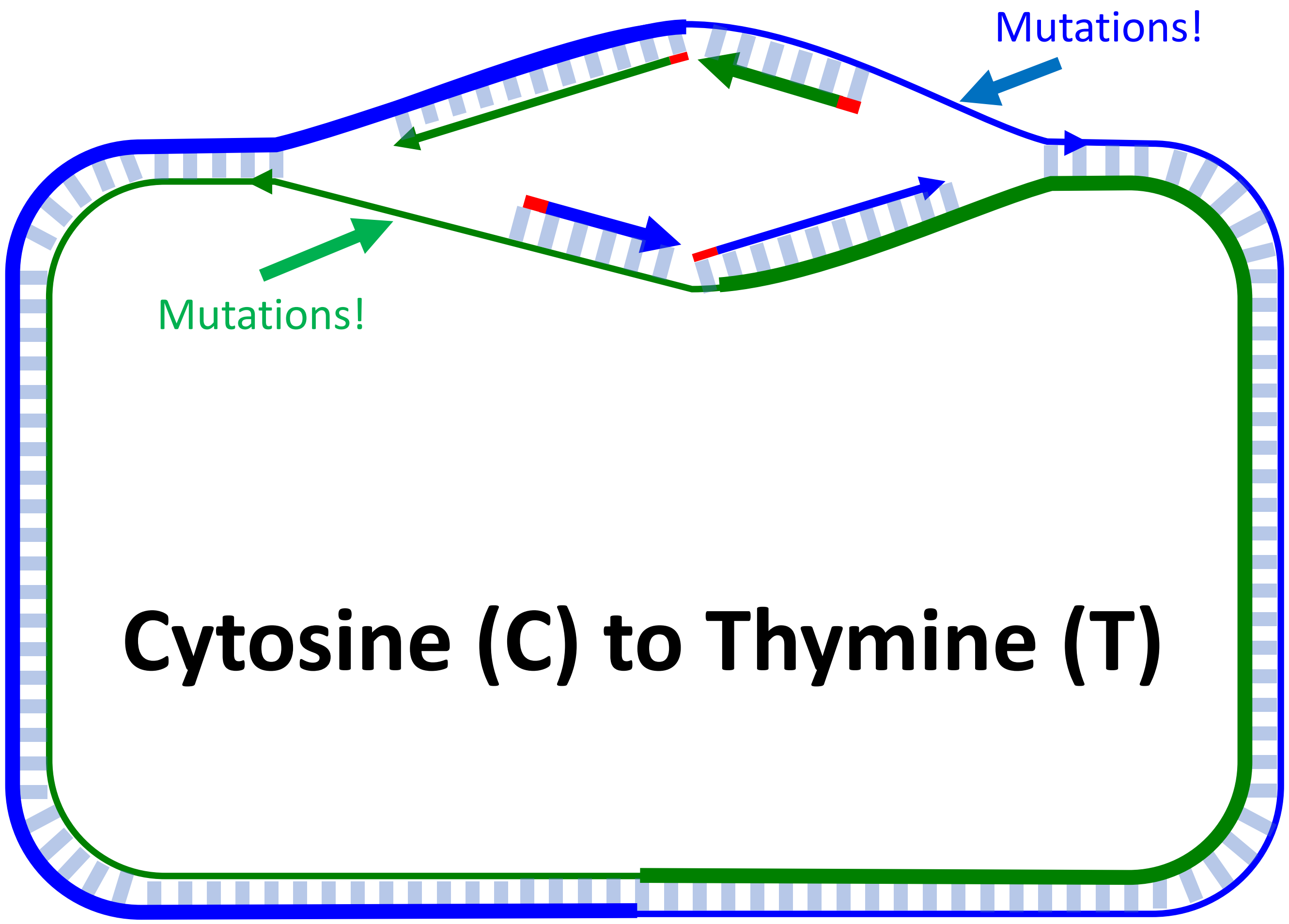
Okazaki fragments

Okazaki fragments

**REPLICATE!**

Instead of copying the entire half-strand, many **Okazaki fragments** are replicated.

# The reverse strand is very quickly replicated. The forward strand takes a longer time to replicate.



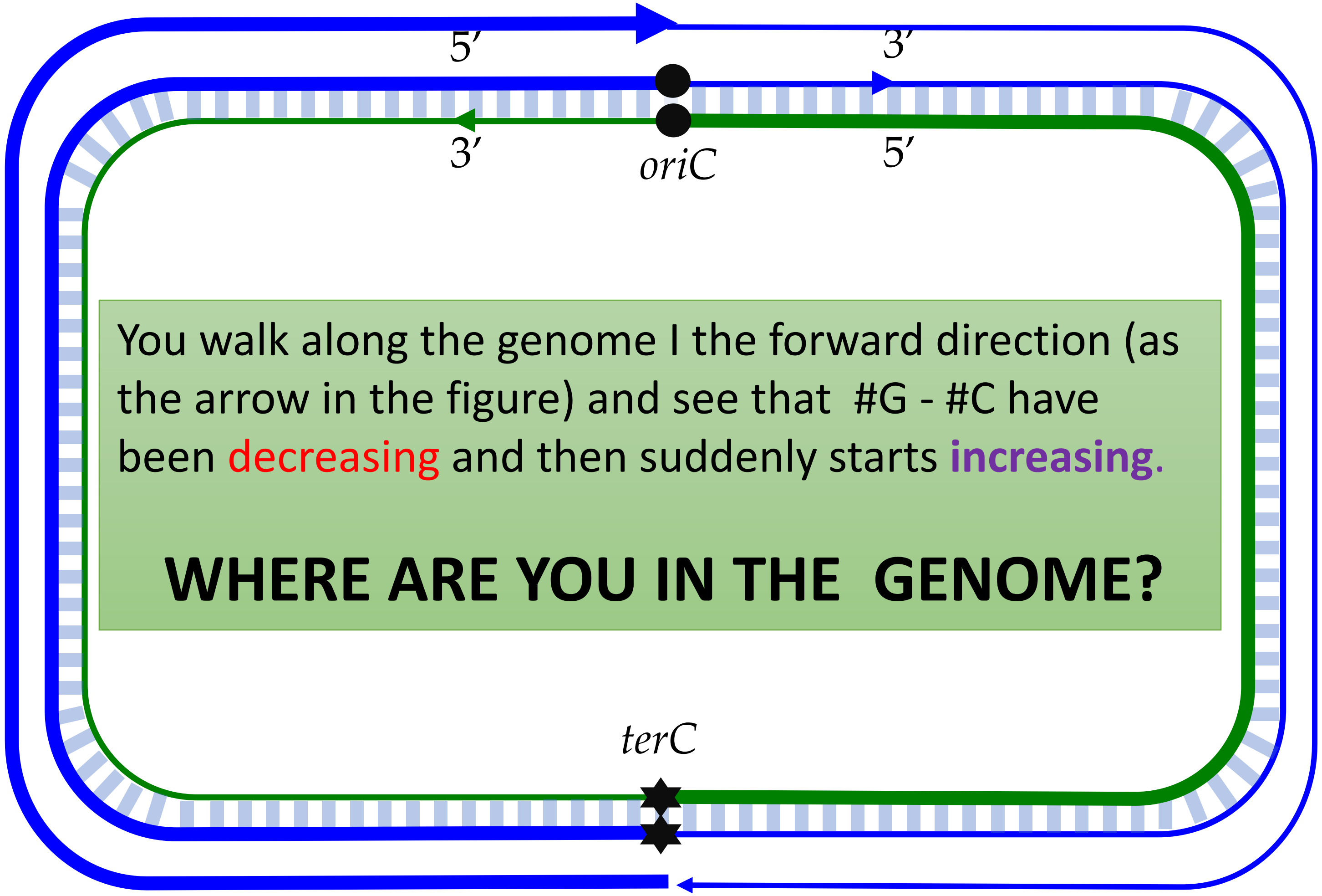
C – G

T – G

T – A

# Take a Walk Along the Genome

**#G - #C is DECREASING** #G - #C is INCREASING



C high
G low

C low
G high

You walk along the genome I the forward direction (as the arrow in the figure) and see that #G - #C have been decreasing and then suddenly starts **increasing**.

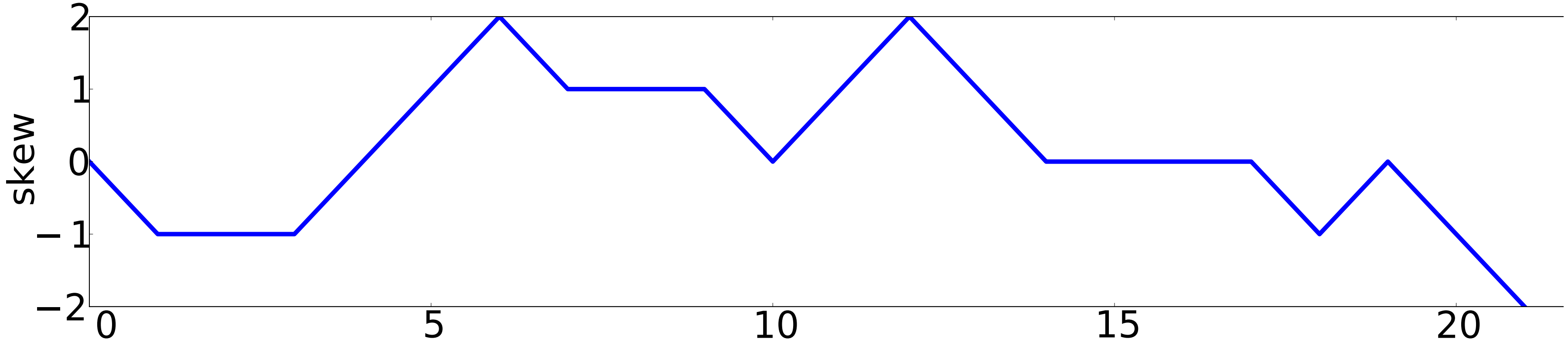
**WHERE ARE YOU IN THE GENOME?**

**C high/G low → #G - #C is DECREASING as we walk along the REVERSE half-strand**

C low/G high → #G - #C is INCREASING as we walk along the FORWARD half-strand

# Skew Diagram

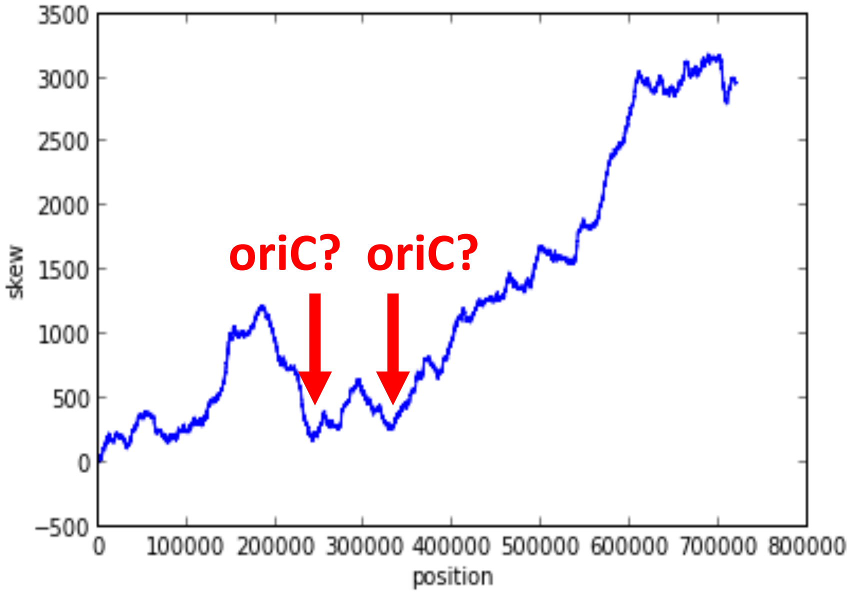
## #G - #C Along the genome



CATGGGCATCGGCCATACGCC

# Finding Multiple Origins of Replication in a Bacterial Genome

- Biologists long believed that each bacterial chromosome has a single replication origin.
- Xia (2012) argued that some bacteria may have multiple replication origins.

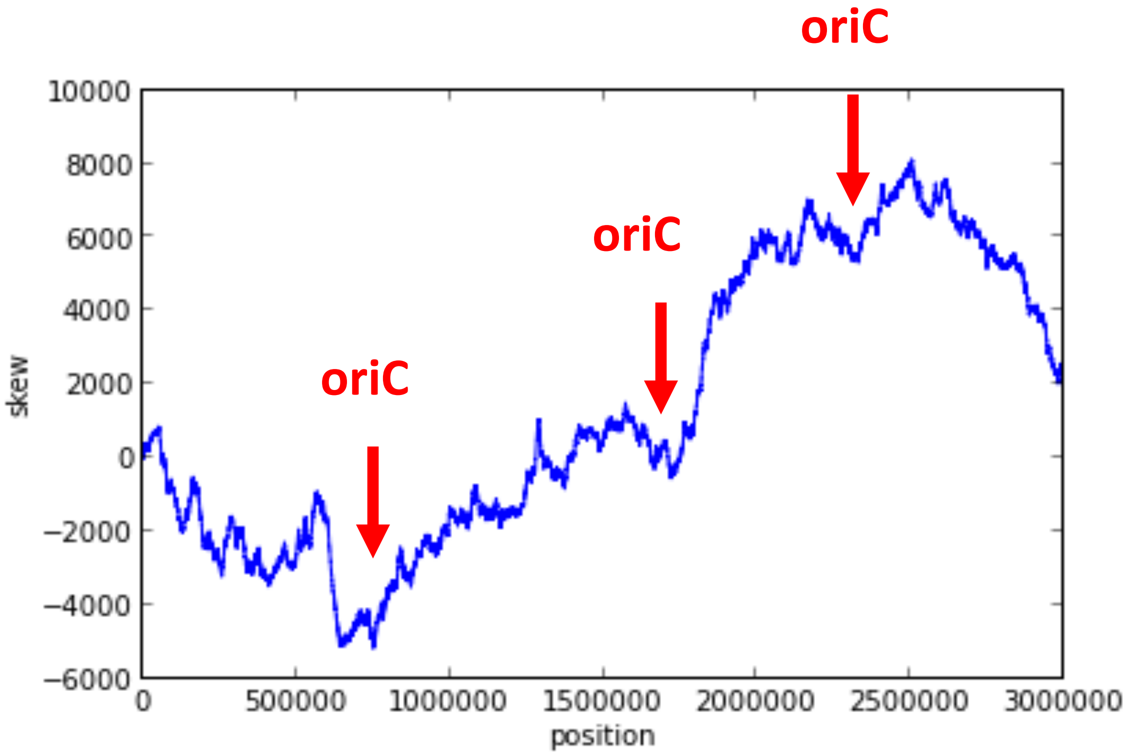


Skew diagram of *Wigglesworthia glossinidia*

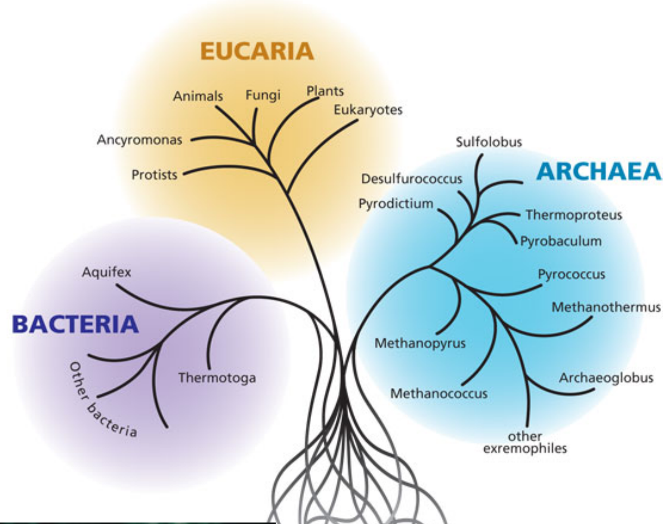
**Open Problem:** Can you confirm or refute the Xia conjecture that this bacterial genome indeed has multiple replication origins?

Project Director
**Mikhail Gelfand**

# Finding *oriC* in Archaea

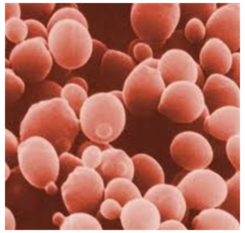


The skew diagram for *Sulfolocus salfataricus*



Project Director
**Mikhail Gelfand**

**Open Problem:** *Archaea* do have multiple origins of replication (3 in *Sulfolocus salfataricus)* but there is no algorithm and software tool yet to predict them reliably – can you develop it?



# Finding *oriC* in Yeast

If you feel that finding bacterial replication origins is difficult, wait until you analyze replication origins in yeast or humans.

**Open Problem:** Yeast genomes have hundreds of origins of replication, but there is no software tool to predict them reliably – can you develop such a tool?

Project Director
**Uri Keich**