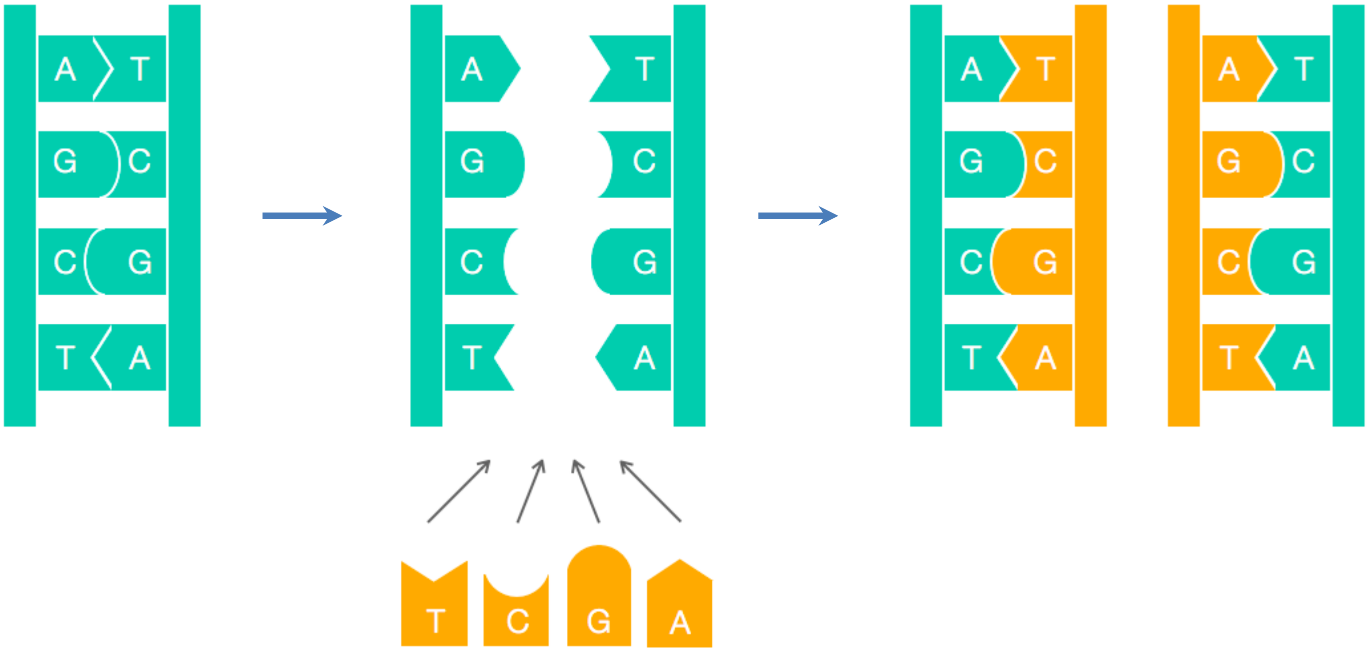
Where in a Genome Does DNA Replication Begin?

Algorithmic Warm-Up

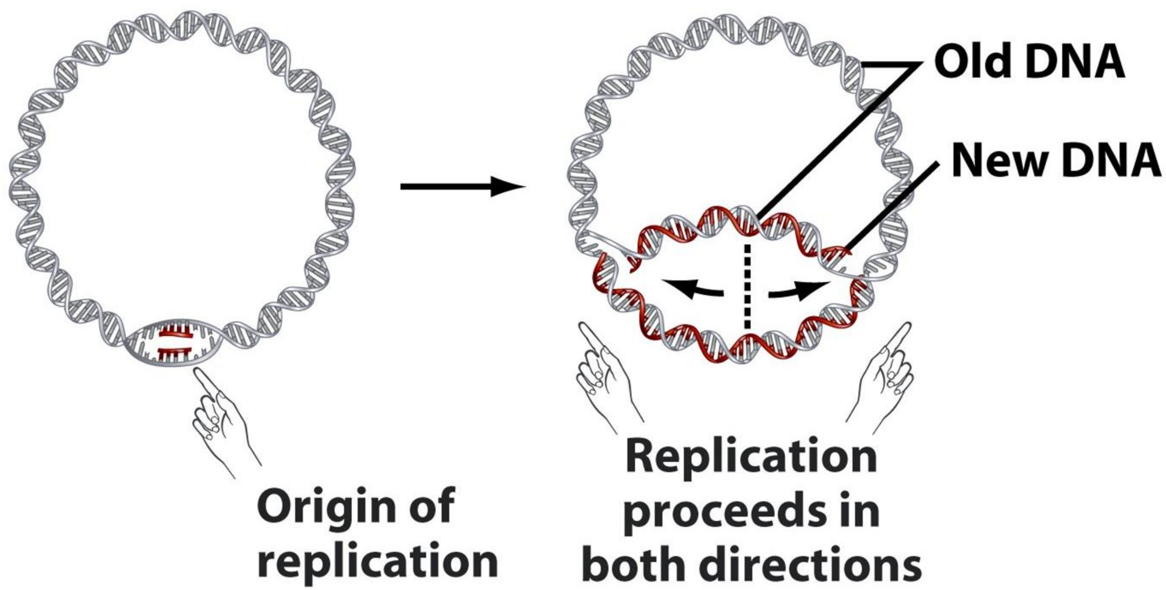
Phillip Compeau and Pavel Pevzner

Bioinformatics Algorithms: an Active Learning Approach

Before a Cell Divides, it Must Replicate its Genome



Replication begins in a region called the **replication origin (*oriC*)**



Where in a genome does it all begin?

Outline

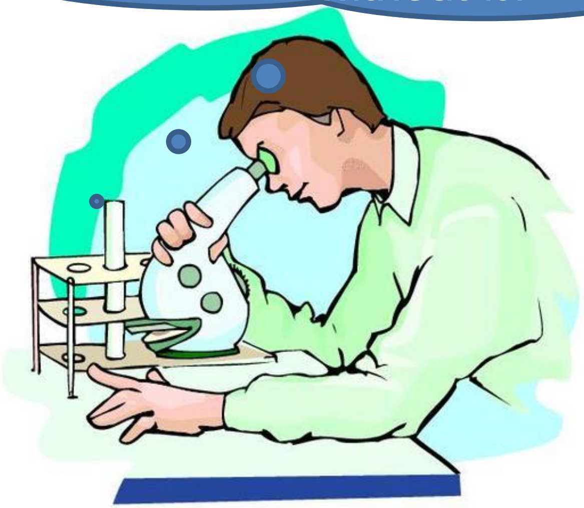
- **Search for Hidden Messages in Replication Origin**
  - **What is a Hidden Message in Replication Origin?**
  - Some Hidden Messages are More Surprising than Others
  - Clumps of Hidden Messages
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - Asymmetry of Replication
  - Why would a computer scientist care about assymetry of replication?
  - Skew Diagrams
  - Finding Frequent Words with Mismatches
  - Open Problems

Finding Origin of Replication

Finding *oriC* Problem: Finding *oriC* in a genome.

- **Input.** A genome.
- **Output.** The location of *oriC* in the genome.

OK – let’s cut out this DNA fragment. Can the genome replicate without it?



This is not a computational problem!

How Does the Cell Know to Begin Replication in Short *oriC*?

Replication origin of *Vibrio cholerae* (≈500 nucleotides):

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

There must be a **hidden message** telling the cell to start replication here.

The Hidden Message Problem

Hidden Message Problem. Finding a hidden message in a string.

- **Input.** A string *Text* (representing replication origin).
- **Output.** A hidden message in *Text*.

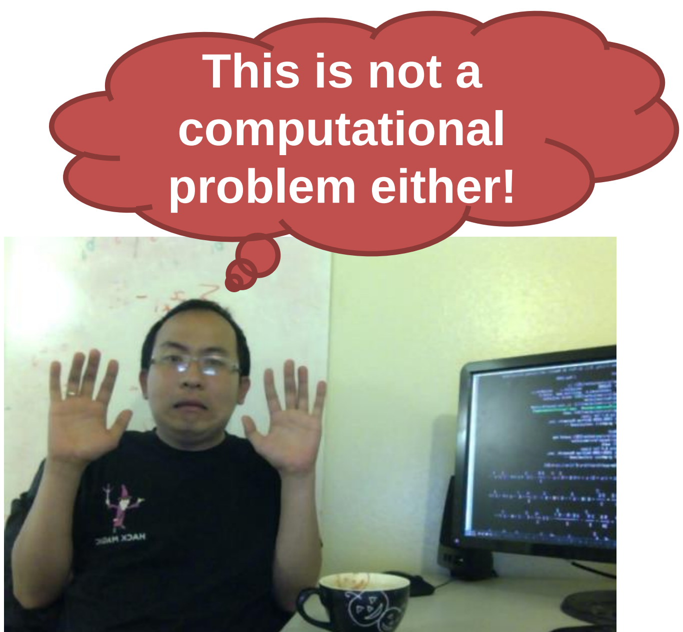
This is not a computational problem either!

The notion of "**hidden message**" is not precisely defined.

The Hidden Message Problem Revisited

Hidden Message Problem. Finding a hidden message in a string.

- **Input.** A string *Text* (representing *oriC*).
- **Output.** A hidden message in *Text*.



The notion of "**hidden message**" is not precisely defined.

Hint: For various biological signals, certain words appear surprisingly frequently in small regions of the genome.

AATTT is a surprisingly frequent 5-mer in:

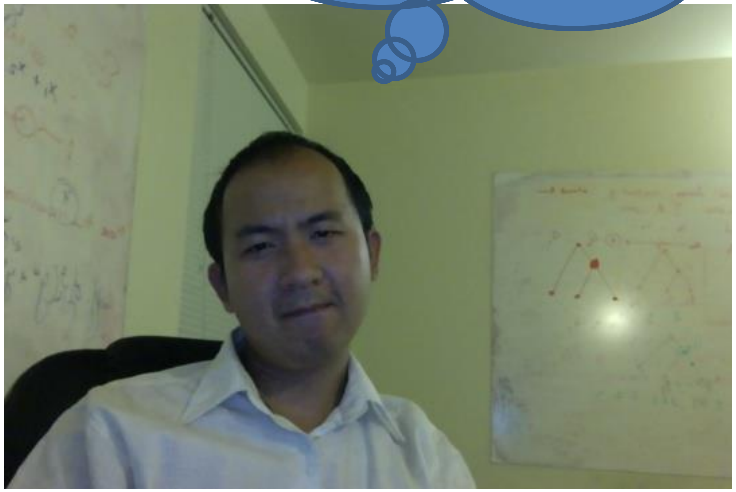
ACA**AATTT**CCAT**AATTT**CCCCA**AATTT**CC

The Frequent Words Problem

Frequent Words Problem. Finding most frequent *k*-mers in a string.

- **Input.** A string *Text* and an integer *k*.
- **Output.** All **most frequent *k*-mers** in *Text*.

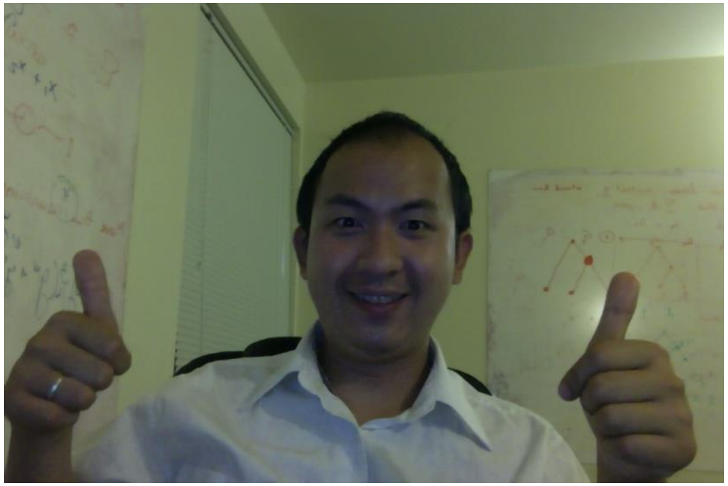
This is better, but where is the definition of “a most frequent *k*-mer?”



The Frequent Words Problem

Frequent Words Problem. Finding most frequent *k*-mers in a string.

- **Input.** A string *Text* and an integer *k*.
- **Output.** All **most frequent *k*-mers** in *Text*.



Son Pham, Ph.D., kindly gave us permission to use his photographs and greatly helped with preparing this presentation. **Thank you Son!**

A *k*-mer ***Pattern*** is a **most frequent *k*-mer** in a text if no other *k*-mer is more frequent than *Pattern.*

AATTT is a most frequent 5-mer in:

ACA**AATTT**GCAT**AATTT**CGGGA**AATTT**CCT

Does the Frequent Words Problem Make Sense to Biologists?

Frequent Words Problem. Finding most frequent *k*-mers in a string.

- **Input.** A string *Text* and an integer *k*.
- **Output.** All **most frequent *k*-mers** in *Text*.

Replication is performed by **DNA polymerase** and the initiation of replication is mediated by a protein called ***DnaA***.

DnaA binds to short (typically 9 nucleotides long) segments within the replication origin known as a ***DnaA* box**.

A *DnaA* box is a hidden message telling *DnaA*: "**bind here!**"
And *DnaA* wants to see multiple *DnaA* boxes.

Outline

- **Search for Hidden Messages in Replication Origin**
  - What is a Hidden Message in Replication Origin?
  - **Some Hidden Messages are More Surprising than Others**
  - Clumps of Hidden Messages
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - Asymmetry of Replication
  - Why would a computer scientist care about assymetry of replication?
  - Skew Diagrams
  - Finding Frequent Words with Mismatches
  - Open Problems

oriC of *Vibrio cholerae*

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaacctgagtgg
atgacatcaagataggtcgttgtatctccttcctctcgtactctcatgaccacggaaagatgatcaag
agaggatgatttcttggccatatcgcaatgaatacttgtgacttgtgcttccaattgacatcttcagc
gccatattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgttt
atcttgttttgactgagacttgttaggatagacggtttttcatcactgactagccaaagccttactct
gcctgacatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacctcttgatcatcg
atccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatca
tgtttccttaaccctctattttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

Too Many Frequent Words – Which One is a Hidden Message?

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaacctgagtgg
atgacatcaagataggtcgttgtatctccttcctctcgtactctcatgaccacggaaag**ATGATCAAG**
agaggatgatttcttggccatatcgcaatgaatacttgtgacttgtgcttccaattgacatcttcagc
gccatattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgttt
atcttgttttgactgagacttgttaggatagacggtttttcatcactgactagccaaagccttactct
gcctgacatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cg
atccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCA**
Tgtttccttaaccctctattttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

Most frequent 9-mers in this *oriC* (all appear 3 times):
ATGATCAAG, **CTTGATCAT**, **TCTTGGATCA**, **CTCTTGATC**

Is it **STATISTICALLY** surprising to find a 9-mer appearing **3 or more** times within ≈ 500 nucleotides?

Hidden Message Found!

atcaatgatcaacgtaagcttctaagcATGATCAAGgtgctcacacagtttatccacaacctgagtgg
atgacatcaagataggtcgttgtatctccttcctctcgtactctcatgaccacggaaagATGATCAAG
agaggatgatttcttggccatatcgcaatgaatacttgtgacttgtgcttccaattgacatcttcagc
gccatattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgttt
atcttgttttgactgagacttgttaggatagacggtttttcatcactgactagccaaagccttactct
gcctgacatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacctCTTGATCATcg
atccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagctCTTGATCA
TgtttccttaaccctctatttttacggaagaATGATCAAGctgctgctCTTGATCATcgtttc

ATGATCAAG

| | | | | | | | | are **reverse complements** and likely ***DnaA*** boxes

TACTAGTTC (*DnaA* does not care what strand to bind to)

It is **VERY SURPRISING** to find a 9-mer appearing **6 or more** times (counting reverse complements) within a short ≈ 500 nucleotides

Can we Now Find Hidden Messages in *Thermotoga petrophila?*

```
aactctatacctcctttttgtcgaatttgtgtgatttatagagaaaatcttattaactgaaactaa
aatggtaggtttggtggtaggttttgtgtacattttgtagtatctgattttaattacataccgta
tattgtattaaattgacgaacaattgcatggaattgaatatatgcaaaacaaacctaccaccaaac
tctgtattgaccattttaggacaacttcagggtggtaggtttctgaagctctcatcaatagactat
tttagtctttacaaacaatattaccgttcagattcaagattctacaacgctgttttaatgggcgtt
gcagaaaacttaccacctaaaatccagtatccaagccgatttcagagaaacctaccacttacctac
cacttacctaccacccgggtggtaagttgcagacattattaaaaacctcatcagaagcttgttcaa
aaatttcaatactcgaaacctaccacctgcgtcccctattatttactactactaataatagcagta
taattgatctgaaaagaggtggtaaaaaa
```

No single occurrence of **ATGATCAAG** or **CTTGATCAT** from *Vibrio Cholerae*!!!

Applying the Frequent Words Problem to this replication origin:

AACCTACCA, ACCTACCAC, GGTAGGTTT, TGGTAGGTT, AAACCTACC, CCTACCACC

Different genomes → different hidden messages (*DnaA boxes*)

Hidden Messages in *Thermotoga petrophila*

aactctatacctcctttttgtcgaatttgtgtgatttatagagaaaatcttattaactgaaactaa
aatggtaggttt**GGTGGTAGG**ttttgtgtacattttgtagtatctgattttttaattacataccgta
tattgtattaaattgacgaacaattgcatggaattgaatatatgcaaaacaaa**CCTACCACC**aaac
tctgtattgaccattttaggacaacttcag**GGTGGTAGG**tttctgaagctctcatcaatagactat
tttagtctttacaaacaatattaccgttcagattcaagattctacaacgctgttttaatgggcgtt
gcagaaaacttaccacctaaaatccagtatccaagccgatttcagagaaacctaccacttacctac
cactta**CCTACCACC**cgggtggtaagttgcagacattattaaaaacctcatcagaagcttgttcaa
aaatttcaatactcgaaa**CCTACCACC**tgcgtcccctattatttactactactaataatagcagta
taattgatctgaaaagaggtggtaaaaa

Ori-Finder software confirms that

CCTACCACC

| | | | | | | | | are candidate hidden messages.

GGATGGTGG

We learned how to find hidden messages **IF *oriC* is given.** But we have no clue **WHERE** *oriC* is located in a (long) genome.

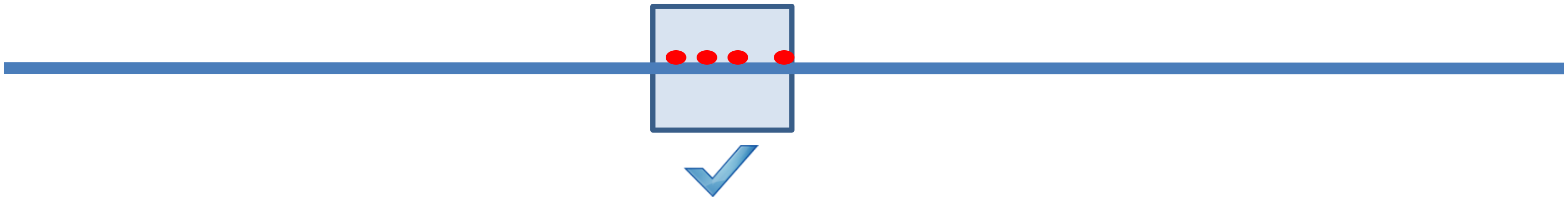
Outline

- **Search for Hidden Messages in Replication Origin**
  - What is a Hidden Message in Replication Origin?
  - Some Hidden Messages are More Surprising than Others
  - **Clumps of Hidden Messages**
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - Asymmetry of Replication
  - Why would a computer scientist care about assymetry of replication?
  - Skew Diagrams
  - Finding Frequent Words with Mismatches
  - Open Problems

Finding Replication Origin

Our strategy **BEFORE**: given a previously **known** *oriC* (a 500-nucleotide window), find **frequent words** (clumps) in *oriC* as candidate *DnaA* boxes.

replication origin → **frequent words**



Finding Replication Origin

Our strategy **BEFORE**: given previously **known** *oriC* (a 500-nucleotide window), find **frequent words** (clumps) in *oriC* as candidate *DnaA* boxes.

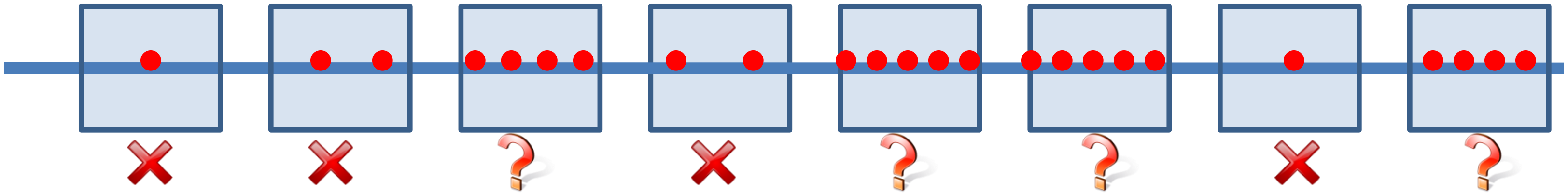
replication origin → **frequent words**

But what if the position of the replication origin within a genome is **unknown**!

Finding Replication Origin

Our strategy **BEFORE**: given previously **known** *oriC* (a 500-nucleotide window), find **frequent words** (clumps) in *oriC* as candidate *DnaA* boxes.

replication origin → **frequent words**



NEW strategy*:* find frequent words in **ALL** windows within a genome. Windows with **clumps** of frequent words are candidate replication origins.

frequent words → **replication origin**

What is a Clump?

Formal: A *k*-mer forms an (***L***, ***t***)-**clump** inside *Genome* if there is a **short** (length ***L***) interval of *Genome* in which it appears **many** (at least ***t***) times.

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

- **Input.** A string *Genome* and integers *k* (length of a pattern), ***L*** (window length), and ***t*** (number of patterns in a clump).
- **Output.** All *k*-mers forming (***L***, ***t***)-**clumps** in *Genome*.

There exist **1904** *different* 9-mers forming (500,3)-clumps in *E. coli* genome. **It is absolutely unclear which of them point to the replication origin…**

Where in a Genome Does DNA Replication Begin?

Algorithmic Warm-Up

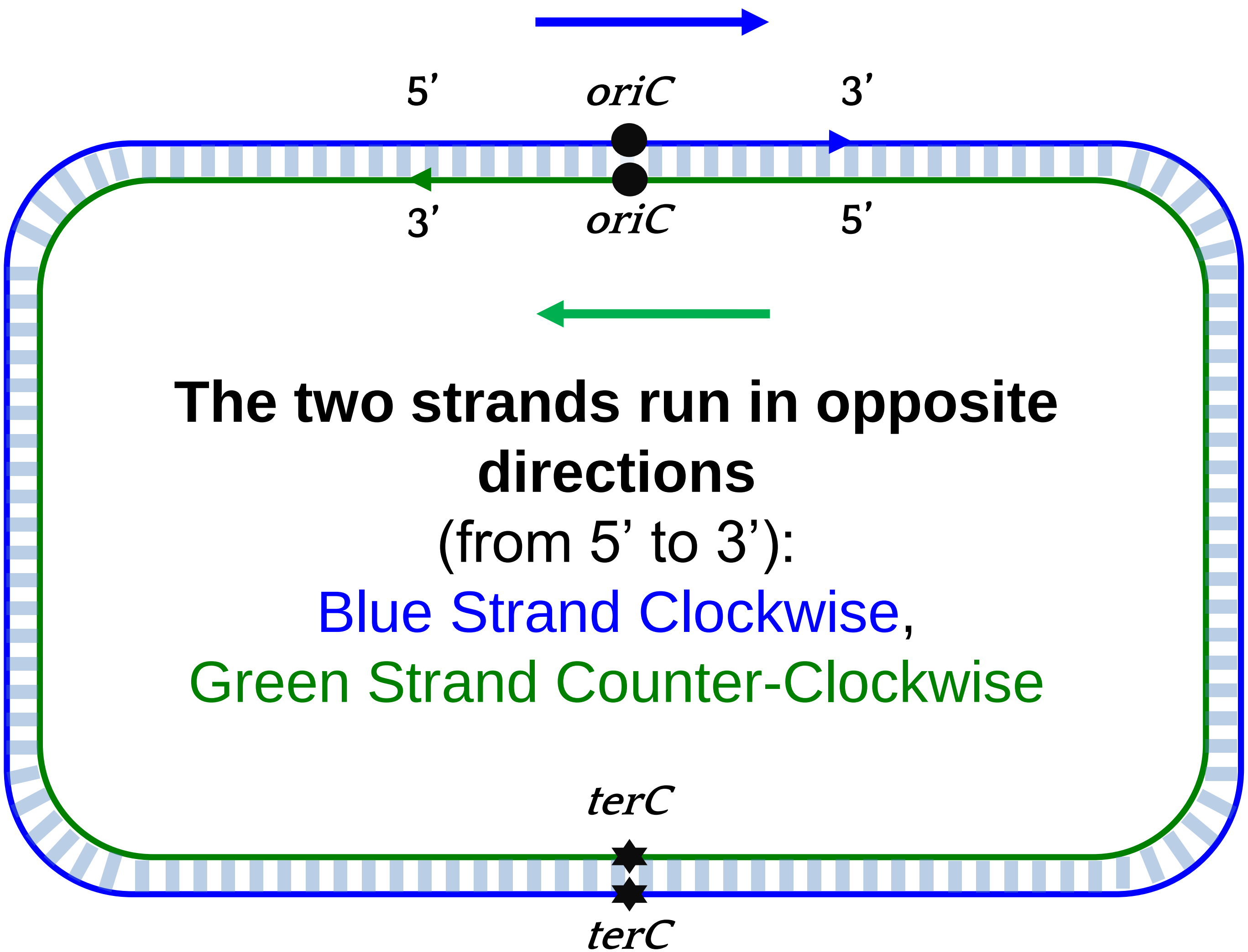
Phillip Compeau and Pavel Pevzner

Bioinformatics Algorithms: an Active Learning Approach

Outline

- **Search for Hidden Messages in Replication Origin**
  - What is a Hidden Message in Replication Origin?
  - Some Hidden Messages are More Surprising than Others
  - Clumps of Hidden Messages
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - **Asymmetry of Replication**
  - Why would a computer scientist care about assymetry of replication?
  - Skew Diagrams
  - Finding Frequent Words with Mismatches
  - Open Problems

DNA Strands Have Directions!



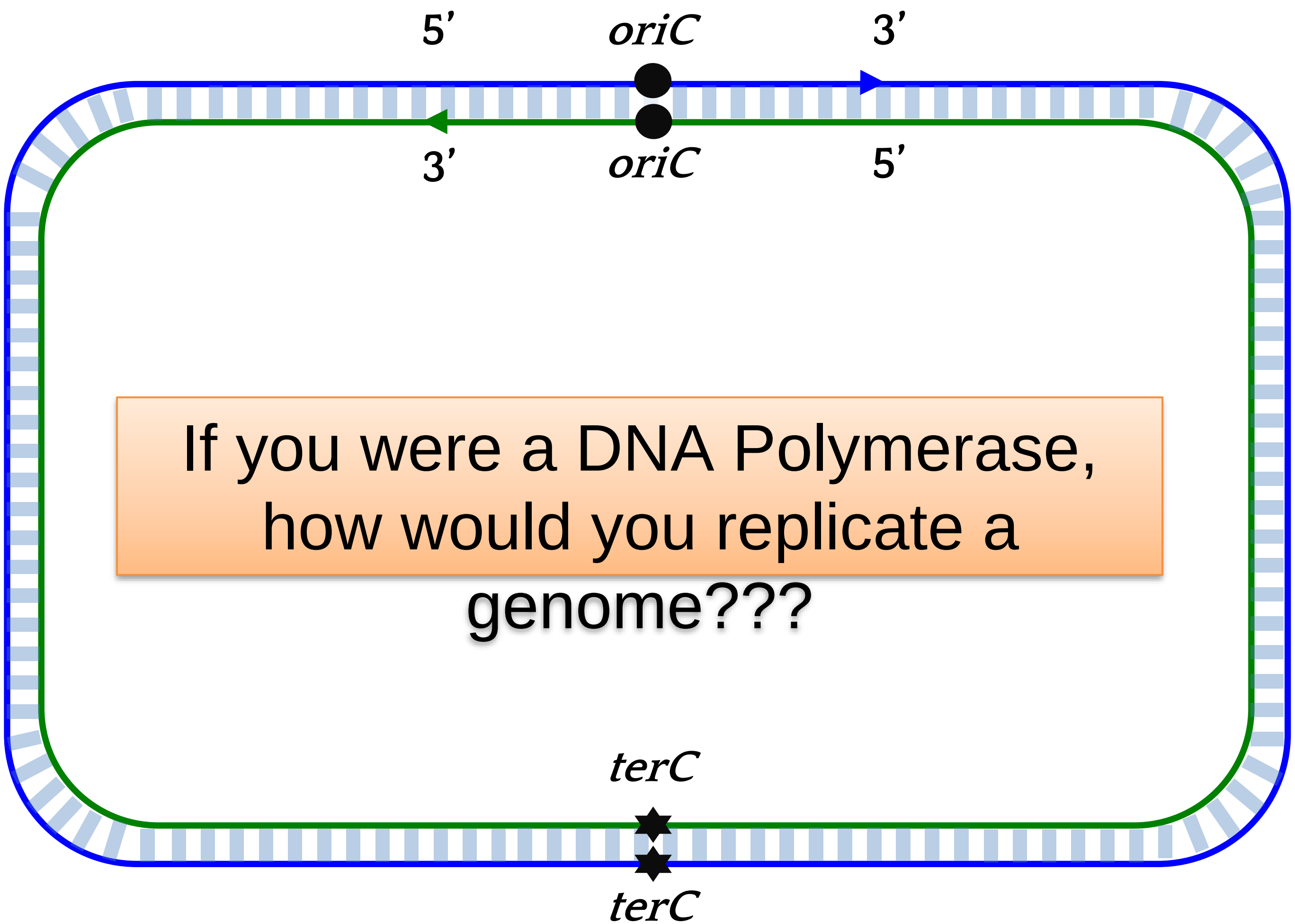
The two strands run in opposite directions

(from 5’ to 3’):

Blue Strand Clockwise,

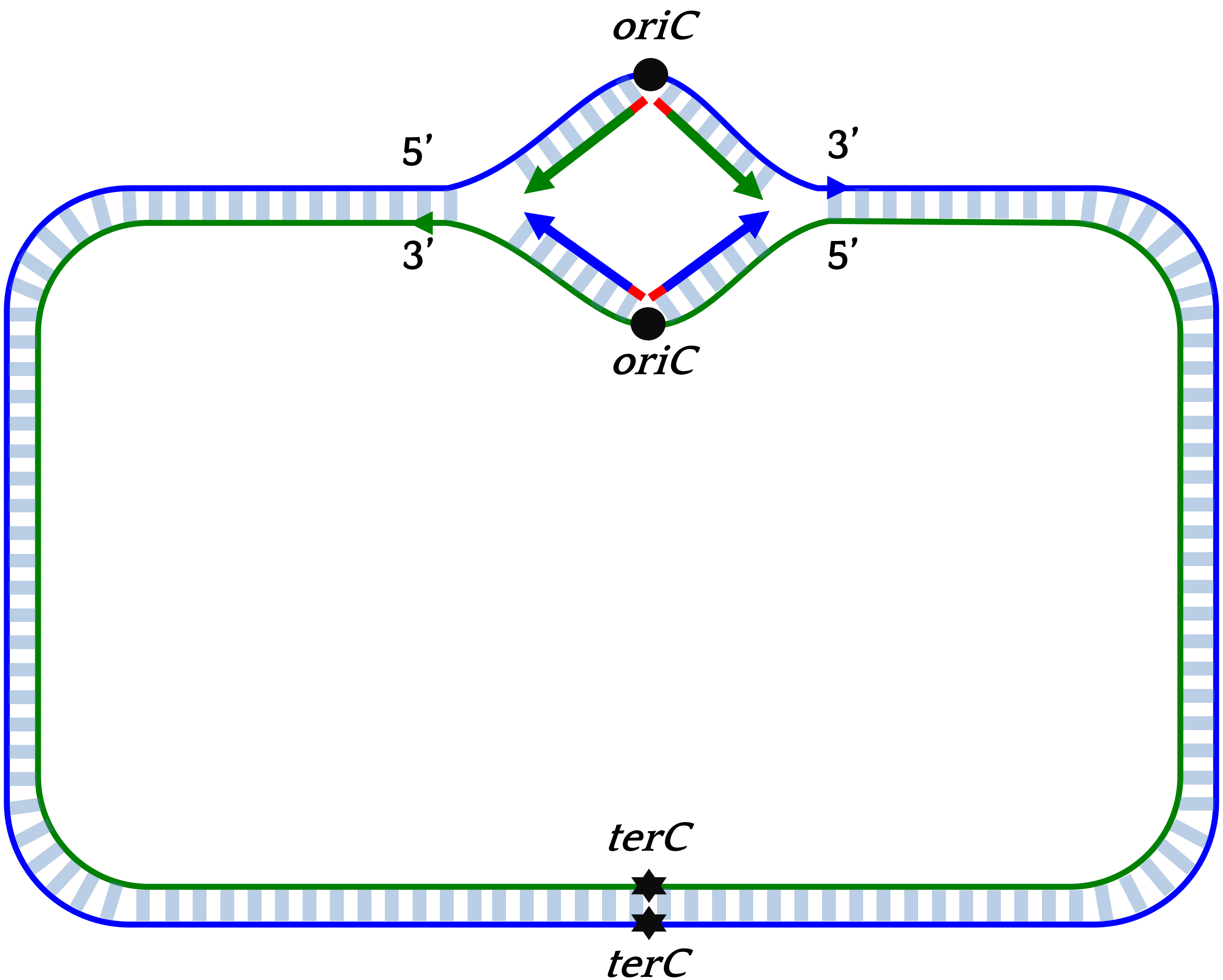
Green Strand Counter-Clockwise

DNA Strands Have Directions

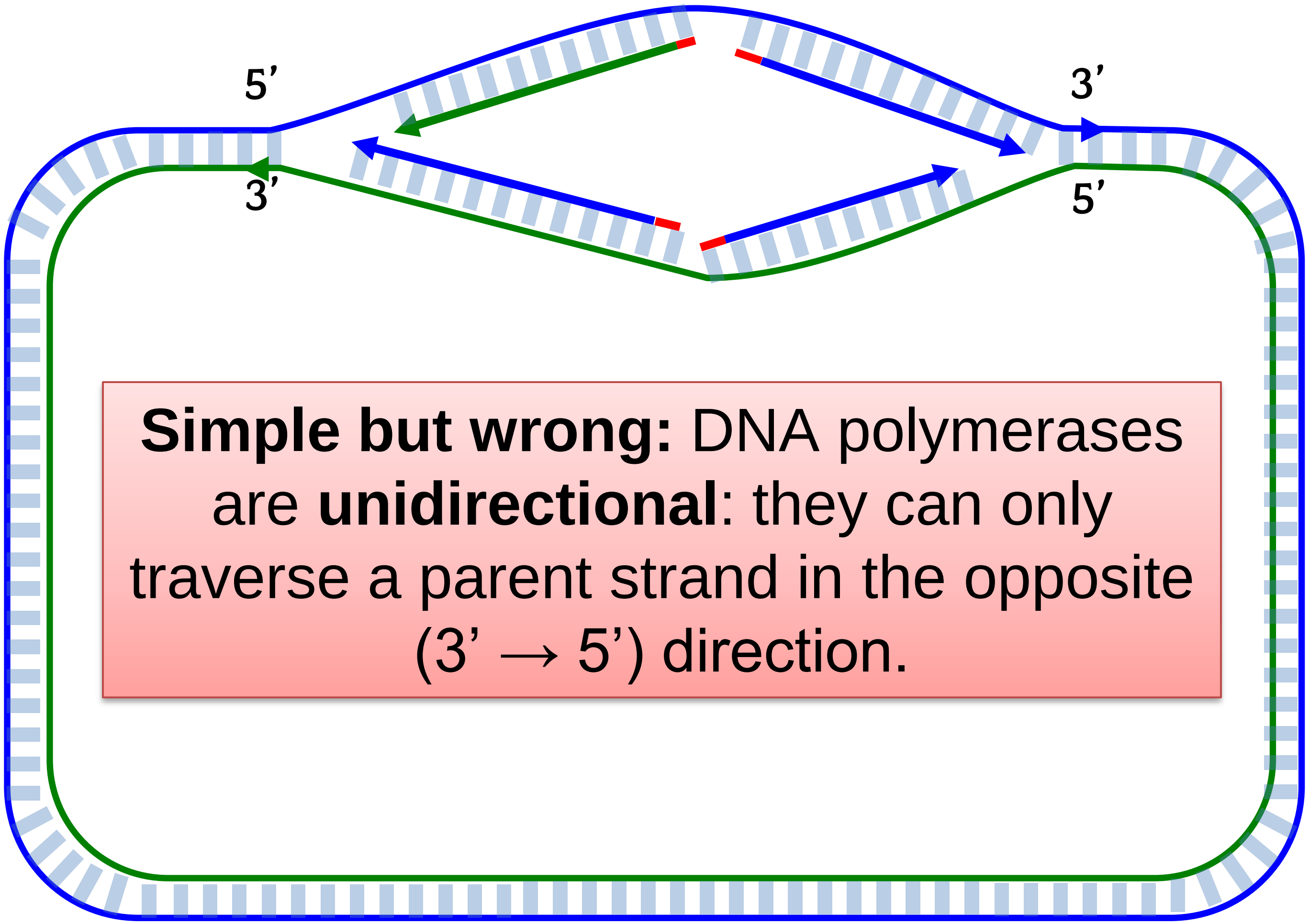


If you were a DNA Polymerase, how would you replicate a genome???

Four DNA Polymerases Do the Job

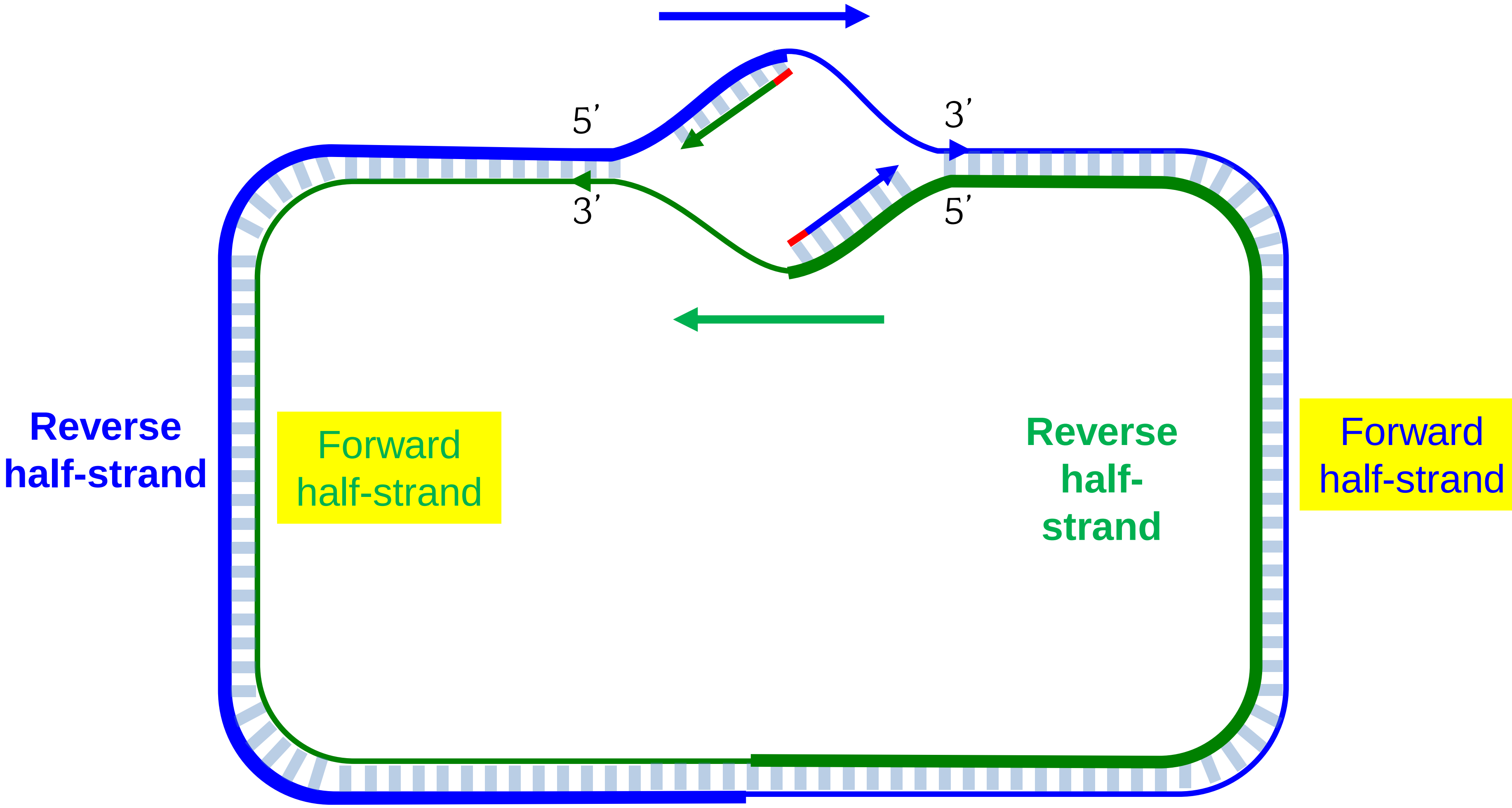


Continue as Replication Fork Enlarges



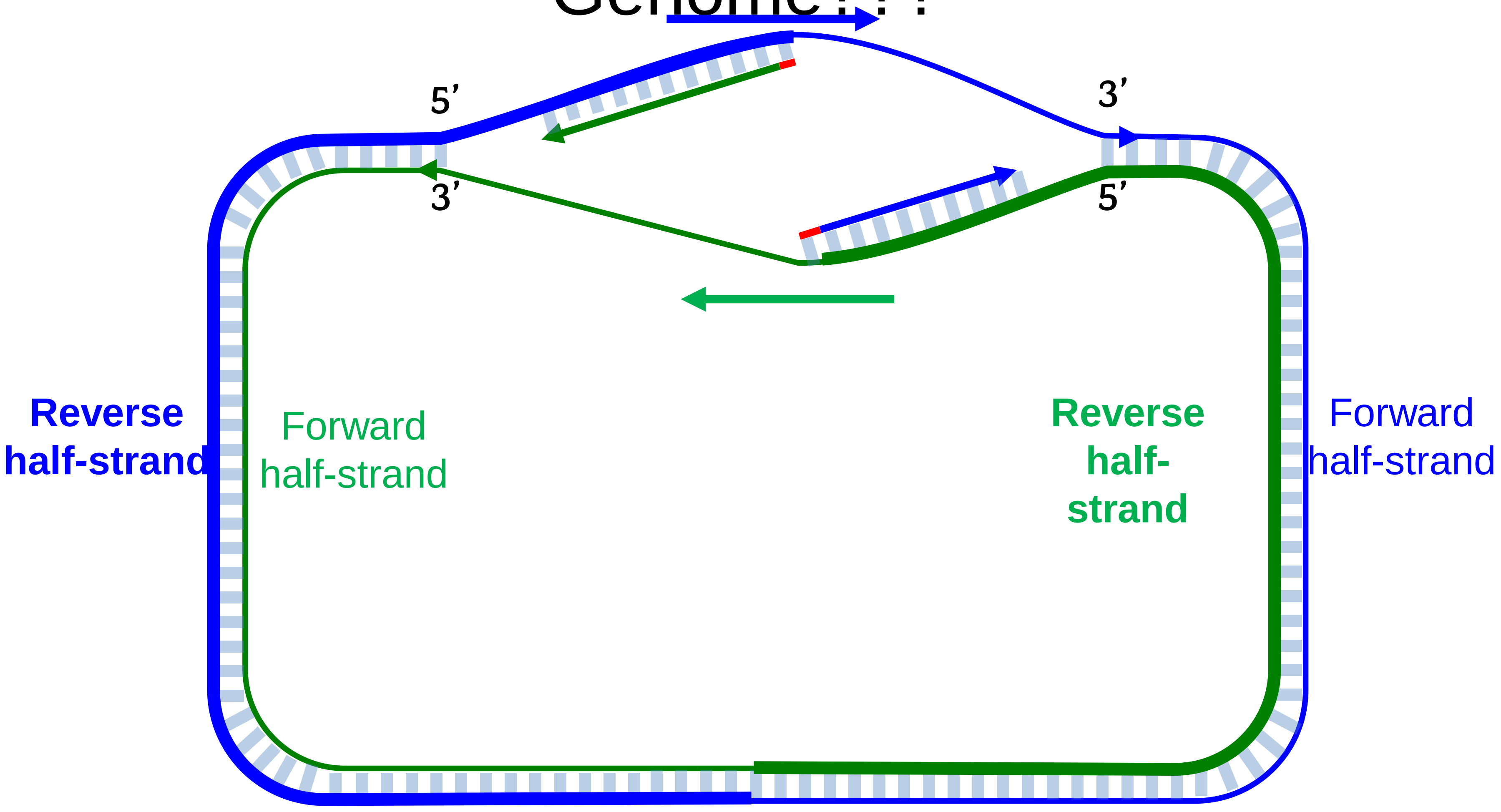
Simple but wrong: DNA polymerases are **unidirectional**: they can only traverse a parent strand in the opposite (3’ → 5’) direction.

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

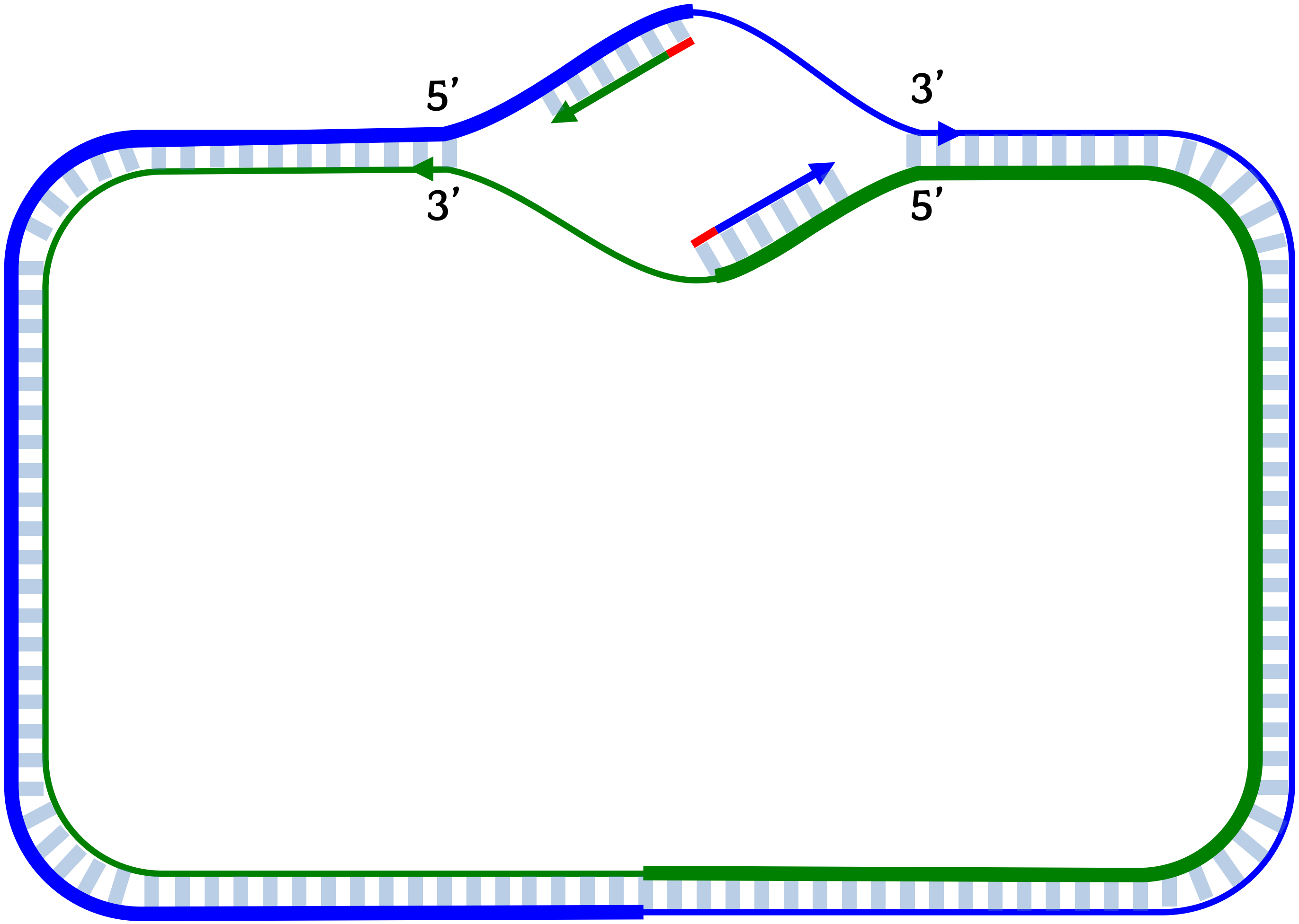


Big problem replicating forward half-strands (thin (thick) lines)

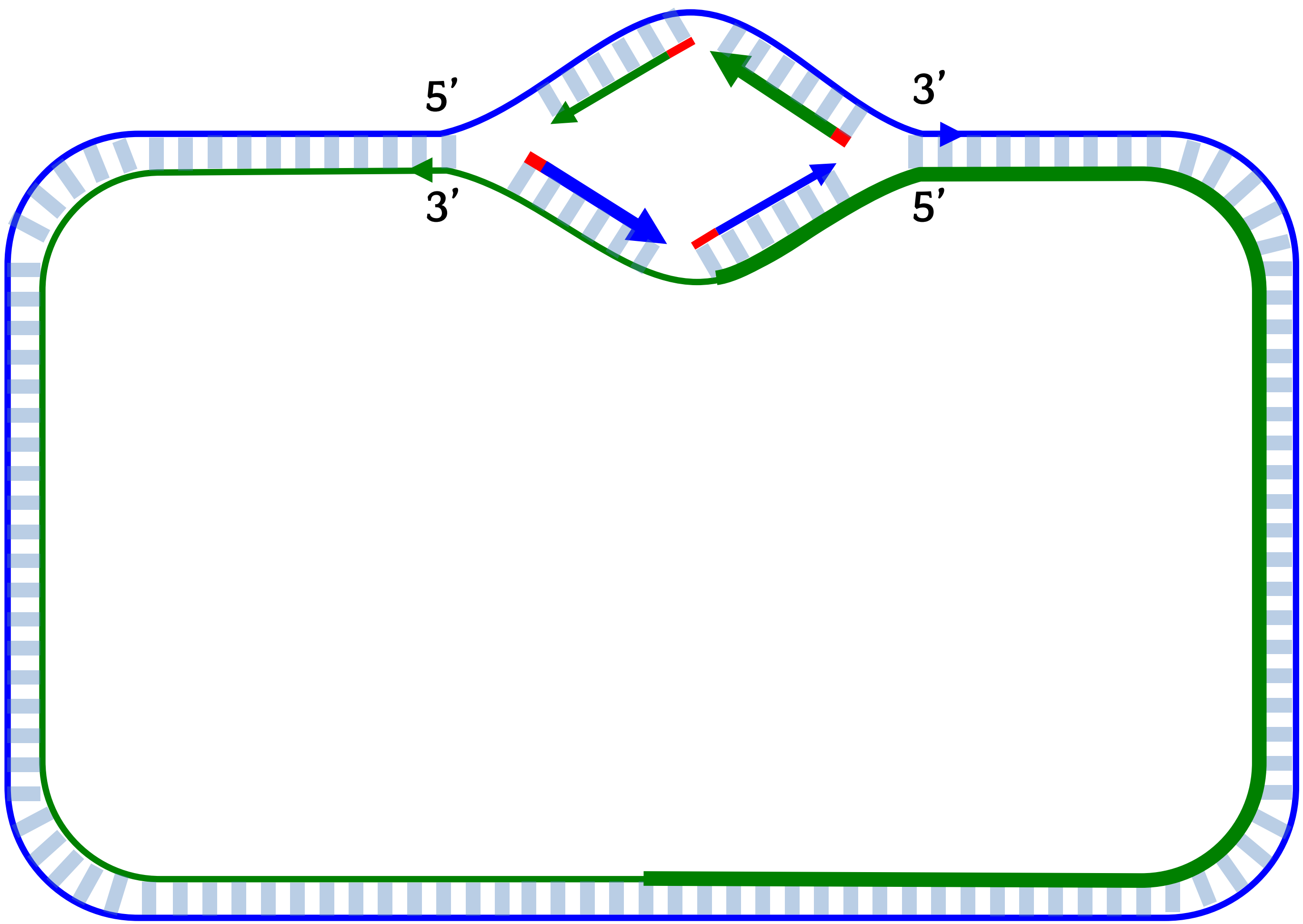
If you Were a **UNIDIRECTIONAL** DNA Polymerase, How Would you Replicate a Genome???



Wait until the Fork Opens and…

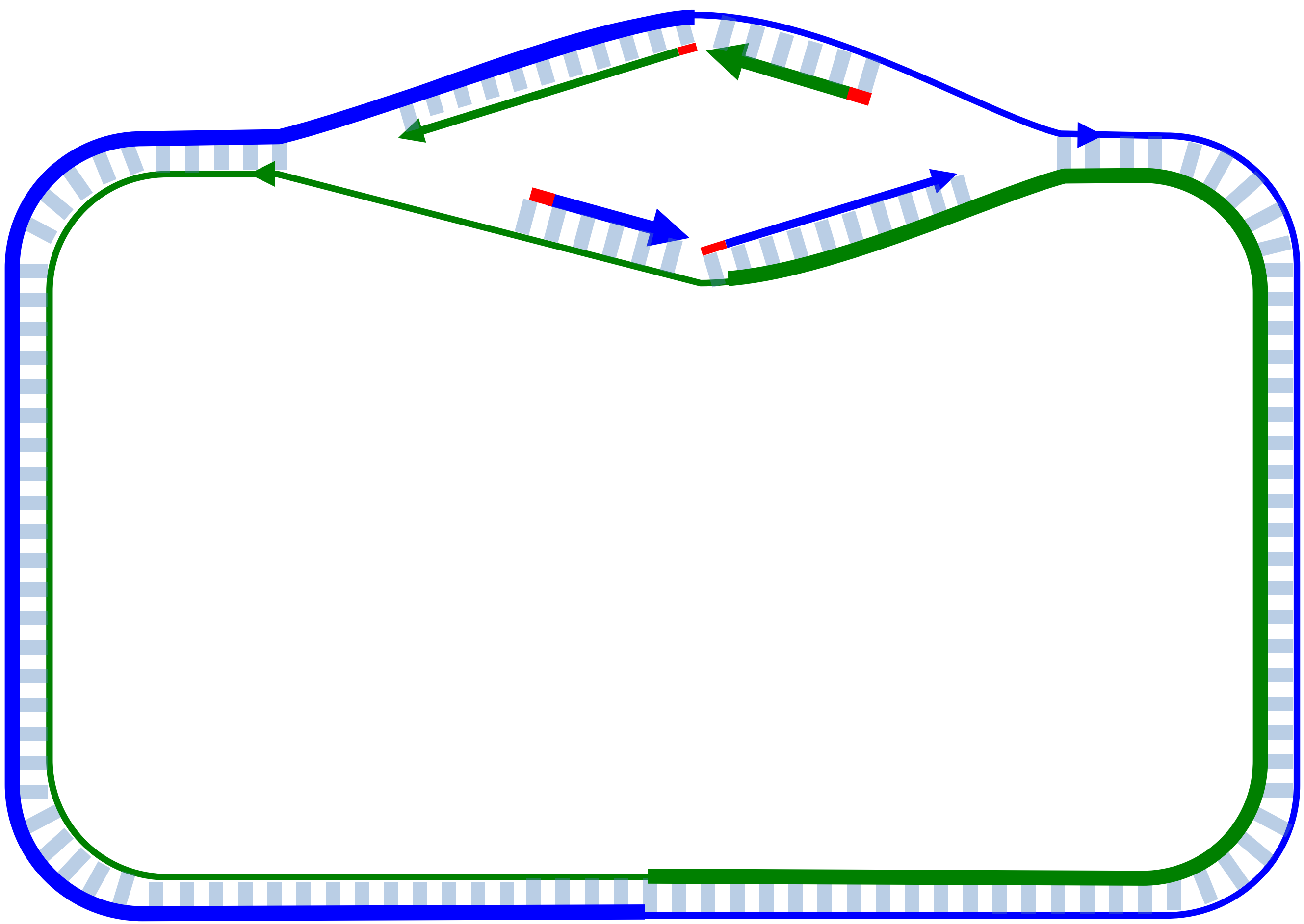


Wait until the Fork Opens and Replicate

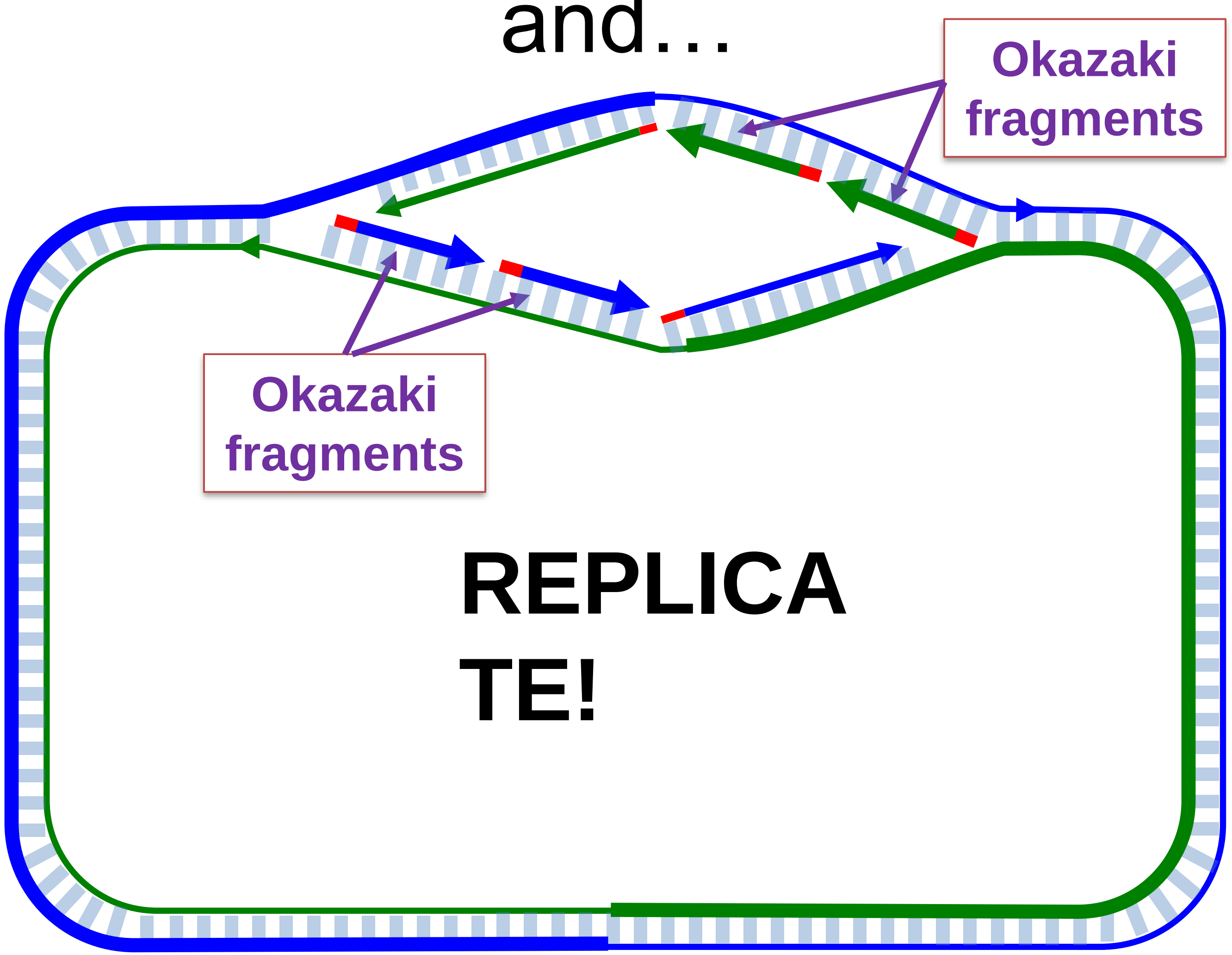


Replicate

Wait until the Fork Opens Even More and…

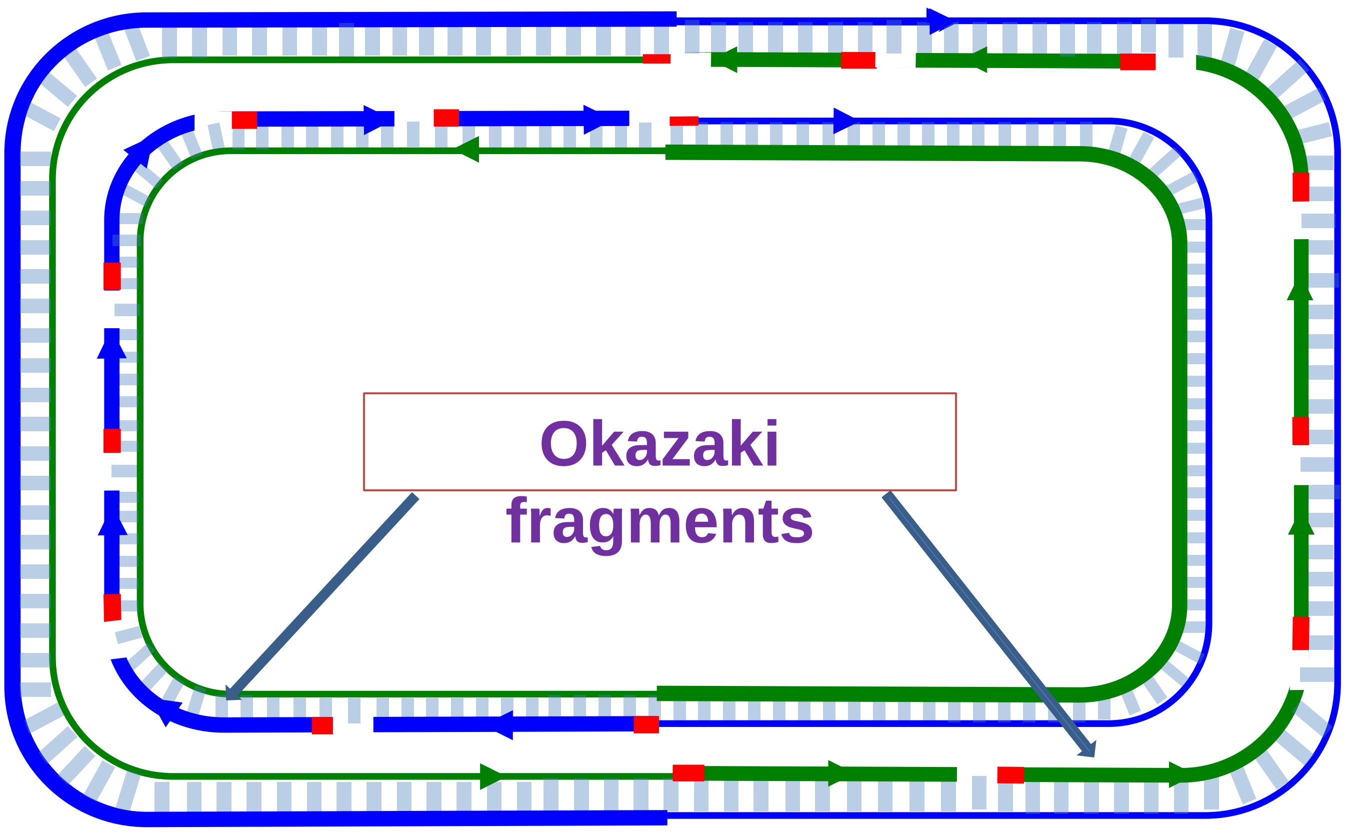


Replicate
Wait until the Fork Opens Even More and…



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

Okazaki Fragments Need to be Ligated to Fill in the Gaps

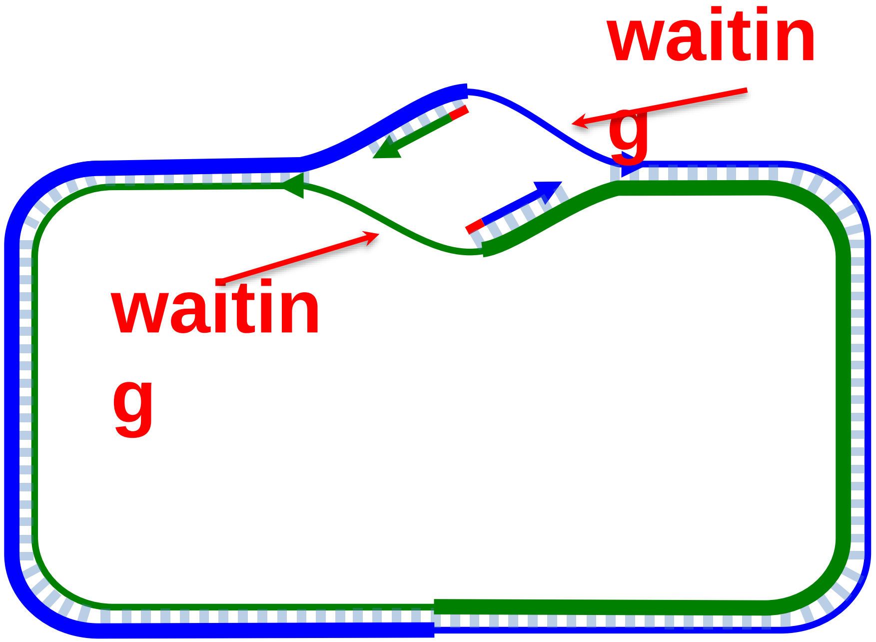


The genome has been replicated!

Different Lifestyles of Reverse and Forward Half-Strands

The **reverse half-strand** lives a **double-stranded** life most of the time.

The **forward half-strand** spends a large portion of its life **single-stranded**, waiting to be replicated.



But why would a computer scientist care?

Outline

- **Search for Hidden Messages in Replication Origin**
  - What is a Hidden Message in Replication Origin?
  - Some Hidden Messages are More Surprising than Others
  - Clumps of Hidden Messages
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - Asymmetry of Replication
  - **Why would a computer scientist care about assymetry of replication?**
  - Skew Diagrams
  - Finding Frequent Words with Mismatches
  - Open Problems

Asymmetry of Replication Affects Nucleotide Frequencies

Single-stranded DNA has a much higher mutation rate than double-stranded DNA.

Thus, if one nucleotide has a greater mutation rate, then we should observe its **shortage** on the forward half-strand that lives single-stranded life!

Which nucleotide (A/C/G/T) has the highest mutation rate? Why?

The Peculiar Statistics of #G - #C

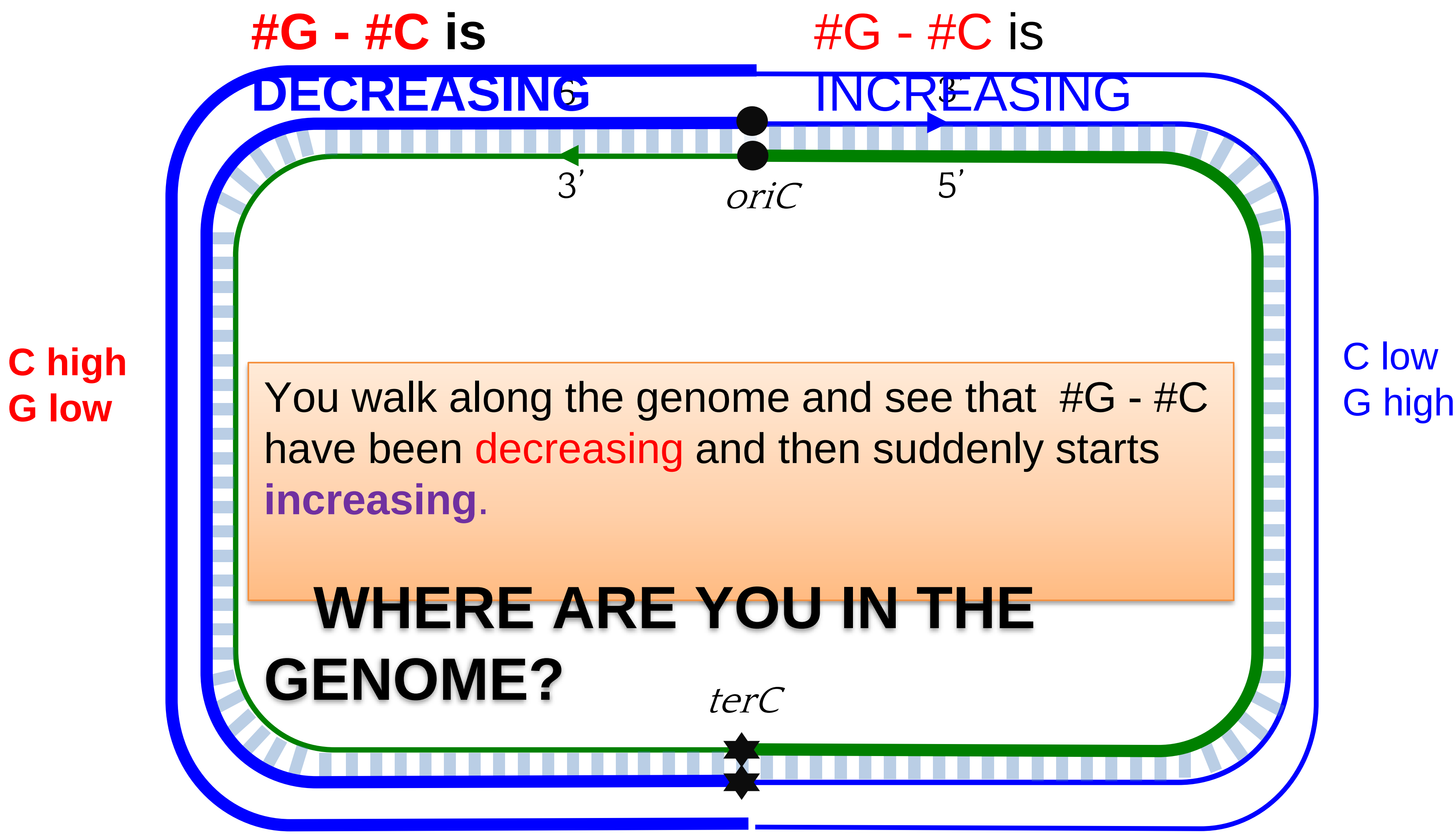
Cytosine (C) rapidly mutates into thymine (T) through **deamination**; deamination rates rise 100-fold when DNA is single stranded!

Forward half-strand (single-stranded life): **shortage of C, normal G**

Reverse half-strand (double-stranded life): **shortage of G, normal C**

	#C	#G	#G - #C
Reverse half-strand	219518	201634	**-17884**
Forward half-strand	207901	211607	**+3706**
Difference	**+11617**	**-9973**	

Take a Walk Along the Genome



You walk along the genome 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

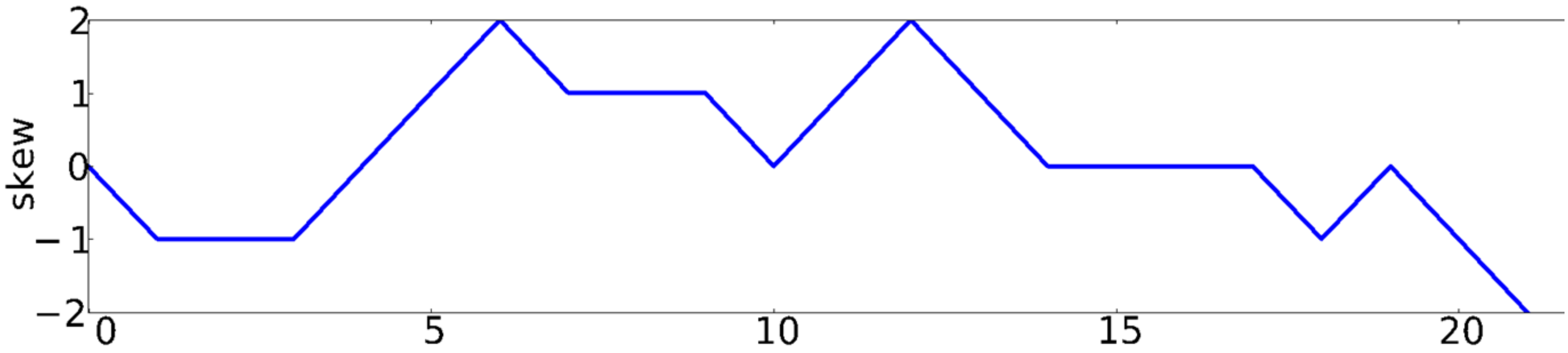
Outline

- **Search for Hidden Messages in Replication Origin**
  - What is a Hidden Message in Replication Origin?
  - Some Hidden Messages are More Surprising than Others
  - Clumps of Hidden Messages
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - Asymmetry of Replication
  - Why would a computer scientist care about assymetry of replication?
  - **Skew Diagrams**
  - Finding Frequent Words with Mismatches
  - Open Problems

Skew Diagram

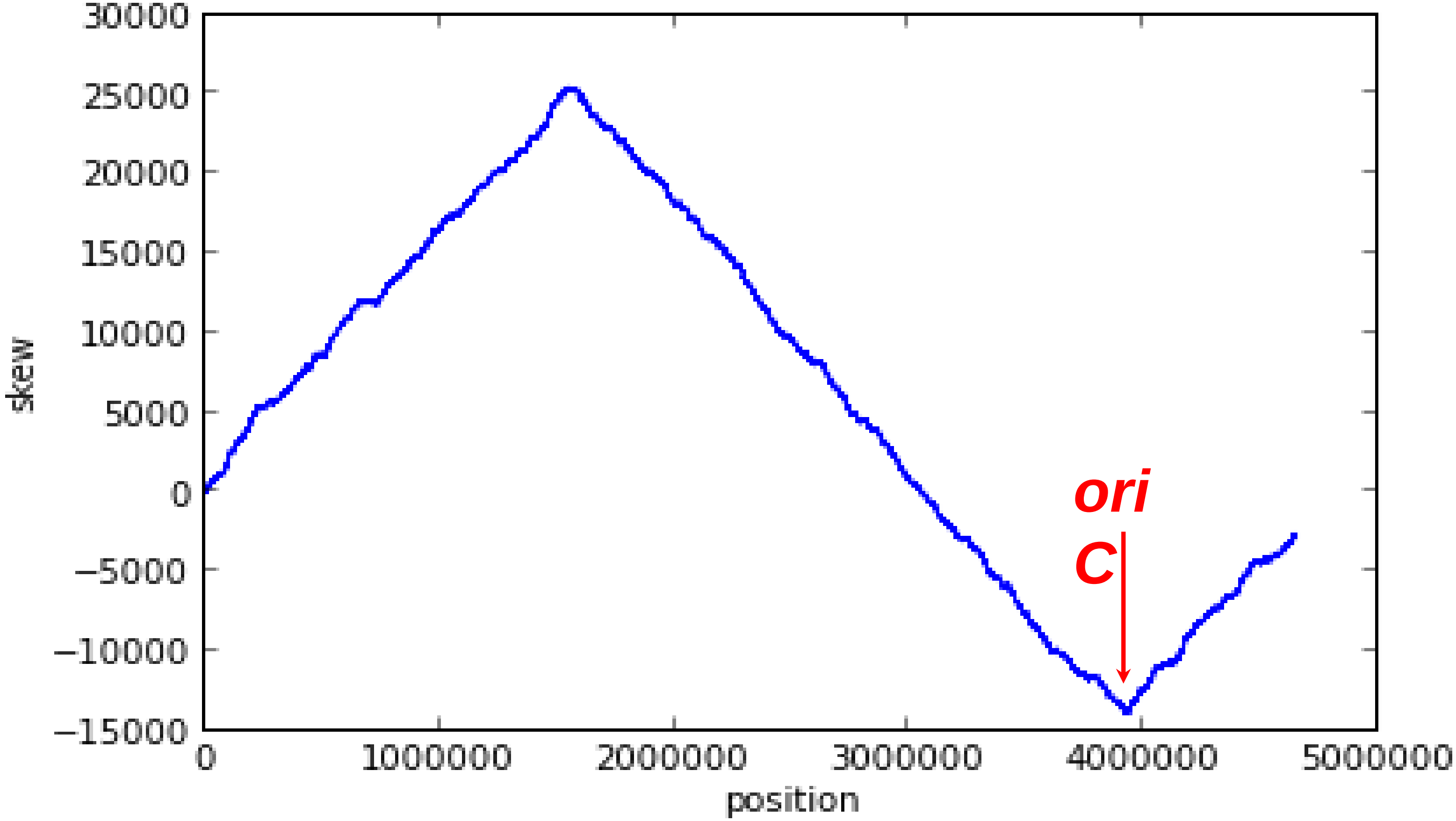
Skew(k): #G - #C for the first *k* nucleotides of *Genome*.

Skew diagram: Plot *Skew*(*k*) against *k*



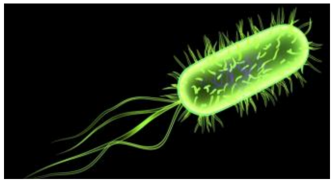
CATGGGCATCGGCCATACGCC

Skew Diagram of *E. Coli*: Where is the Origin of Replication?



You walk along the genome and see that #G - #C have been decreasing and then suddenly starts **increasing**: **WHERE ARE YOU IN THE GENOME?**

We Found the Replication Origin in *E. Coli* **BUT**...



The minimum of the Skew Diagram points to this region in *E. coli*:

```
aatgatgatgacgtcaaaaggatccggataaaacatggtgattgcctcgcataacgcggta
tgaaaatggattgaagcccgggccgtggattctactcaactttgtcggcttgagaaagacc
tgggatcctgggtattaaaaagaagatctatttatttagagatctgttctattgtgatctc
ttattaggatcgcactgccctgtggataacaaggatccggcttttaagatcaacaacctgg
aaaggatcattaactgtgaatgatcggtgatcctggaccgtataagctgggatcagaatga
ggggttatacacaactcaaaaactgaacaacagttgttctttggataactaccggttgatc
caagcttcctgacagagttatccacagtagatcgcacgatctgtatacttatttgagtaaa
ttaacccacgatcccagccattcttctgccggatcttccggaatgtcgtgatcaagaatgt
tgatcttcagtg
```

But there are **no** frequent 9-mers (that appear three or more times) in this region!

SHOULD WE GIVE UP?

Outline

- **Search for Hidden Messages in Replication Origin**
  - What is a Hidden Message in Replication Origin?
  - Some Hidden Messages are More Surprising than Others
  - Clumps of Hidden Messages
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - Asymmetry of Replication
  - Why would a computer scientist care about assymetry of replication?
  - Skew Diagrams
  - **Finding Frequent Words with Mismatches**
  - Open Problems

Searching for Even More Elusive Hidden Messages

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtt
tccttaaccctctattttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

oriC in *Vibrio cholerae* has 6 *DnaA* boxes – **can you find more?**

Previously Invisible *DnaA* Boxes

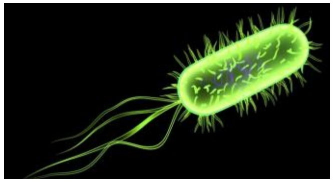
oriC in *Vibrio cholerae* contains ATGATCAAC and CATGATCAT, which differ from canonical *DnaA* boxes ATGATCAAG/CTTGATCAT in a single mutation:

atca**ATGATCAAC**gtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaag**CATGATCAT**ggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtt
tccttaaccctctattttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

Frequent Words with Mismatches Problem. Find the most frequent *k*-mers with mismatches in a string.

- **Input.** A string *Text*, and integers *k* and *d*.
- **Output.** All most frequent *k*-mers with up to *d* mismatches in *Text*.

Finally, *DnaA* Boxes in *E. Coli!*

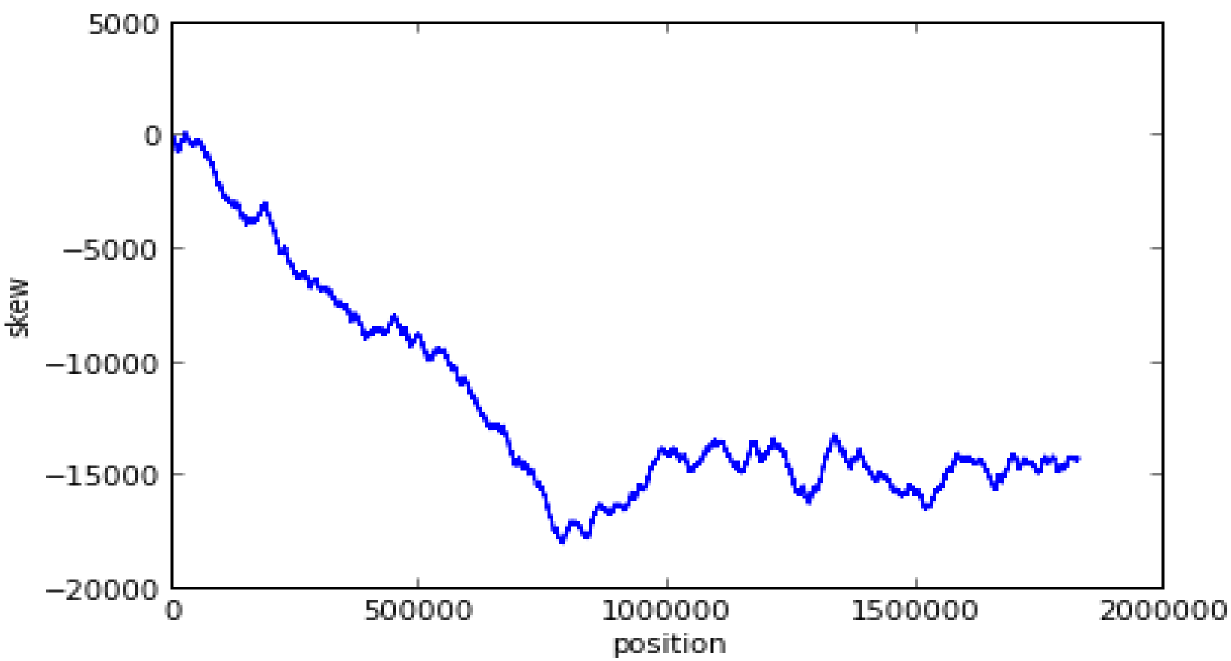


Frequent 9-mers (with 1 Mismatch and Reverse Complements) in putative *oriC* of *E. coli*

aatgatgatgacgtcaaaaggatccggataaaacatggtgattgcctcgcataacgcggt
atgaaaatggattgaagcccgggccgtggattctactcaactttgtcggcttgagaaaga
cctgggatcctgggtattaaaaagaagatctatttatttagagatctgttctattgtgat
ctcttattaggatcgcactgccc**TGTGGATAA**caaggatccggcttttaagatcaacaac
ctggaaaggatcattaactgtgaatgatcggtgatcctggaccgtataagctgggatcag
aatgagggg**TTATACACA**actcaaaaactgaacaacagttgttc**TTTGGATAAC**taccgg
ttgatccaagcttcctgacagag**TTATCCACA**gtagatcgcacgatctgtatacttattt
gagtaaattaacccacgatcccagccattcttctgccggatcttccggaatgtcgtgatc
aagaatgttgatcttcagtg

Complications

- Some bacteria have fewer *DnaA* boxes.
- Terminus of replication is often not located directly opposite to *oriC*.
- The skew diagram is often more complex than in the case of *E. coli*.



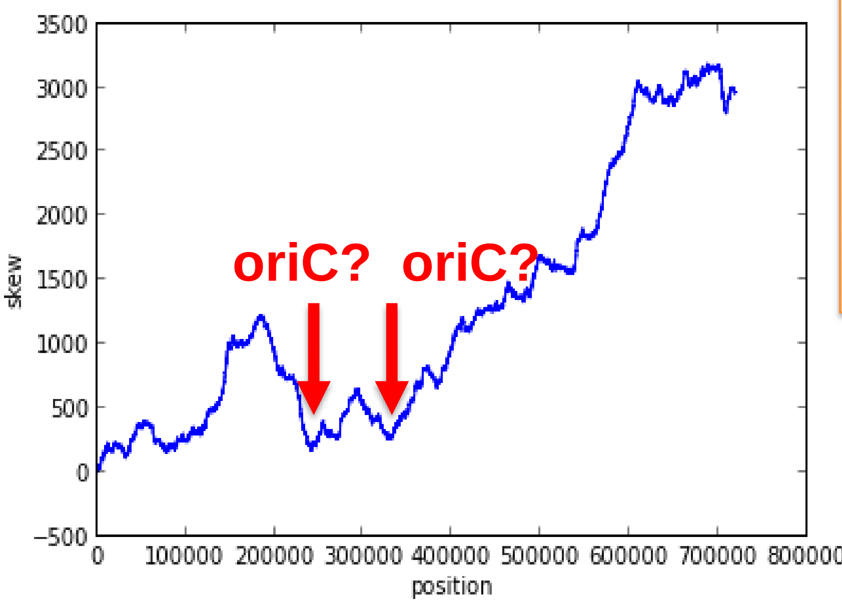
The skew diagram of *Thermotoga*

Outline

- **Search for Hidden Messages in Replication Origin**
  - What is a Hidden Message in Replication Origin?
  - Some Hidden Messages are More Surprising than Others
  - Clumps of Hidden Messages
- **From a Biological Insight toward an Algorithm for Finding Replication Origin**
  - Asymmetry of Replication
  - Why would a computer scientist care about assymetry of replication?
  - Skew Diagrams
  - Finding Frequent Words with Mismatches
  - **Open Problems: From Massive Open Online Courses (MOOC) to Massive Open Online Research (MOOR)**

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.

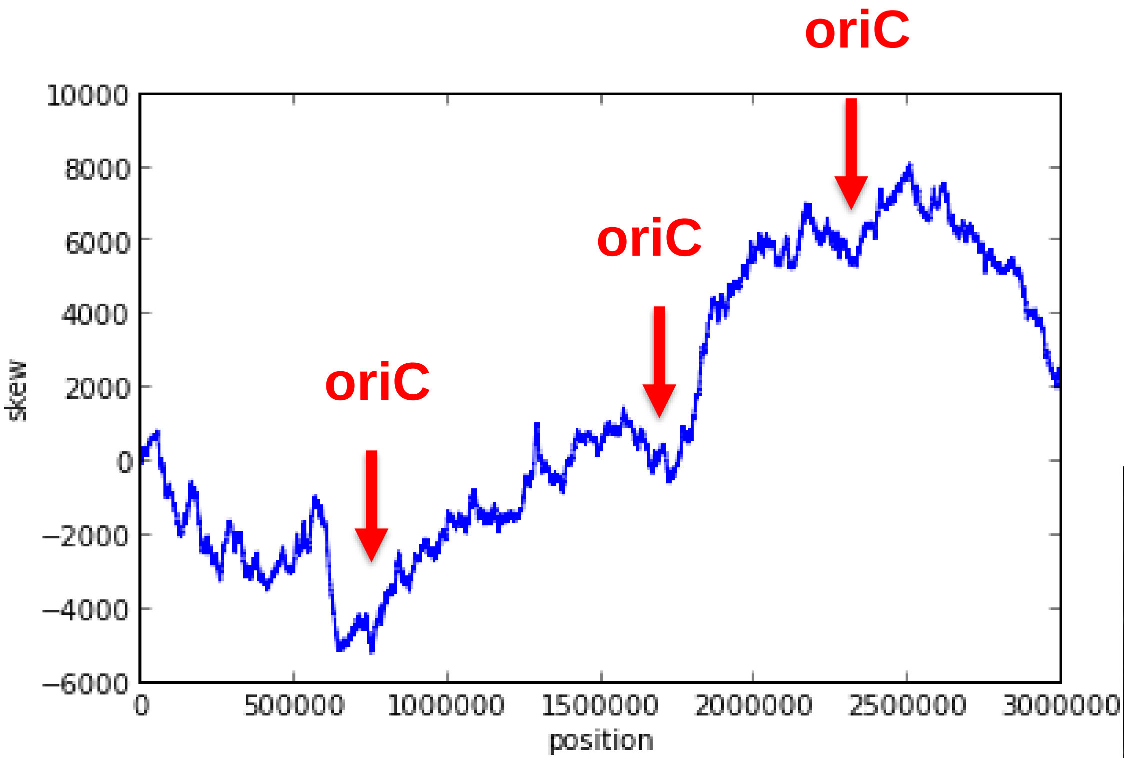


Skew diagram of *Wigglesworthia glossinidia*

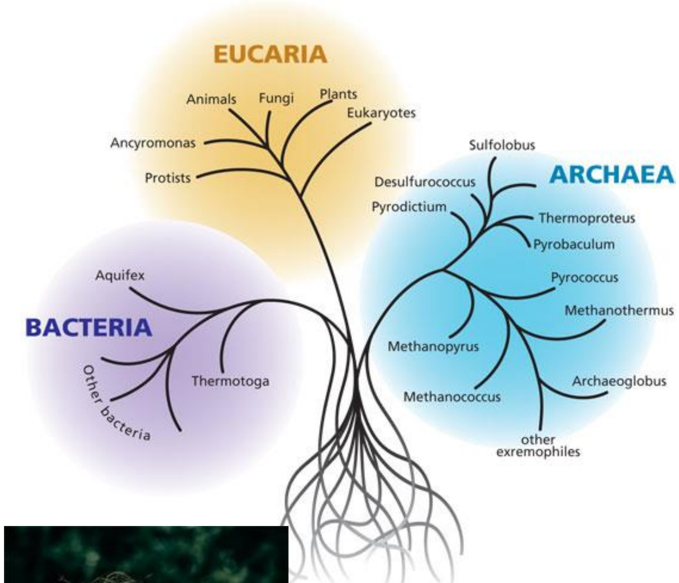
Open Problem: Can you confirm or refute the Xia conjecture that this bacterial genome indeed has multiple replicat[illegible]s?

Project Director **Mikhail Gelfand**

Finding *oriC* in Archaea

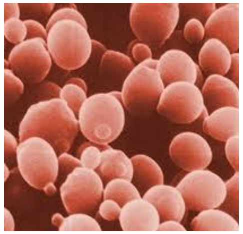


The skew diagram for *Sulfolocus salfataricus*



Project Director **Mikhail Gelfand**

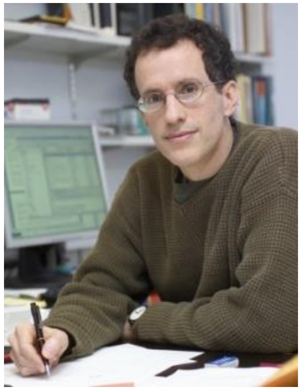
Open Problem: *Archaea* do have multiple origins of replication (3 in *Sulfolocus salfataricus)* but there is no algorithm and software tool yet to predict them reliably – can you develop it?



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

Happy Rosalind!