Bioinformatics: Origin of Replication

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

Book Reference

Chapter 1, Bioinformatics Algorithms: An Active Learning Approach - I

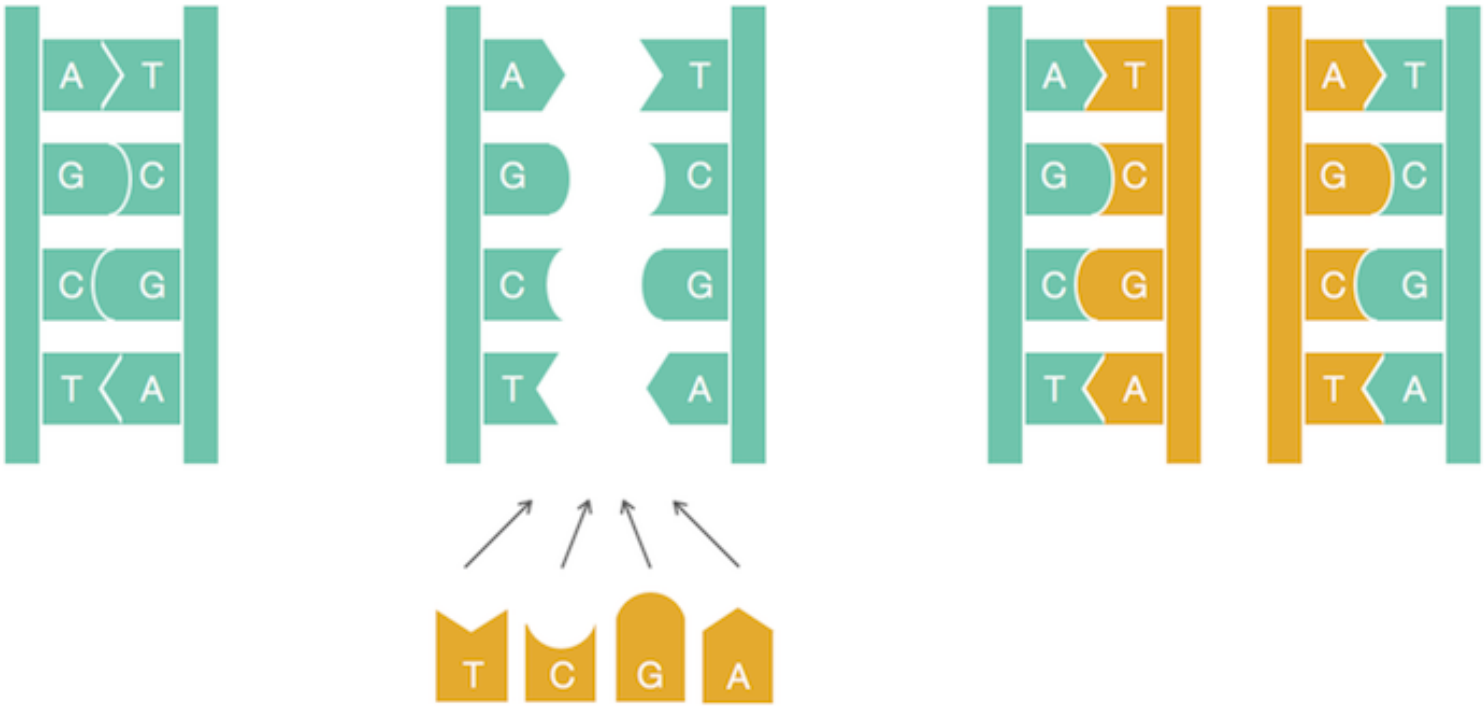
Genome Replication

- Genome replication is one of the most important tasks carried out in the cell. Before a cell can divide, it must first replicate its genome so that each of the two daughter cells inherits its own copy.
- In 1953, James Watson and Francis Crick completed their landmark paper on the DNA double helix with a now-famous phrase:
  - It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.
- They conjectured that the two strands of the parent DNA molecule unwind during replication, and then each parent strand acts as a template for the synthesis of a new strand.
- As a result, the replication process begins with a pair of complementary strands of DNA and ends with two pairs of complementary strands.

BRAC UNIVERSITY
Inspiring Excellence

Genome Replication

- They conjectured that the two strands of the parent DNA molecule unwind during replication, and then each parent strand acts as a template for the synthesis of a new strand.
- As a result, the replication process begins with a pair of complementary strands of DNA and ends with two pairs of complementary strands.



Origin of Replication

- Replication begins in a genomic region called the replication origin (denoted oriC) and is performed by molecular copy machines called DNA polymerases.
- Locating oriC presents an important task not only for understanding how cells replicate but also for various biomedical problems.

Gene Therapy

1. **viral vectors:** Genetically engineered mini-genomes, which are able to penetrate cell walls (just like real viruses).
2. Viral vectors carrying artificial genes have been used in agriculture to engineer frost-resistant tomatoes and pesticide-resistant corn.
3. 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 girl had been so vulnerable to infections that she was forced to live in a sterile environment.

Viral Vector

- 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.
- To ensure that the vector actually replicates inside the cell, biologists must know where oriC is in the vector's genome and ensure that the genetic manipulations that they perform do not affect it.
- We have the first problem - where in the genome is the OriC located?

BRAC UNIVERSITY
Inspiring Excellence

Genome Replication

The problem

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

Is this a computational problem?

What would a biologist do?

Immediately, start deleting short segments from the genome to find a segment whose deletion stops replication.

What would a computer scientist do?

- Let's start with a simple case:
  - Bacterial Genome - mostly a single chromosome
  - typically a few hundred nucleotides long

BRAC UNIVERSITY
Inspiring Excellence

OriC of Vibrio *cholerae*

1. Lets start with a known example: Vibrio cholerae.

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

- How does the bacterial cell know to begin replication exactly in this short region within the much larger Vibrio cholerae chromosome, which consists of 1,108,250 nucleotides?
- There must be some hidden message in the oriC region ordering the cell to begin replication here.

BRAC UNIVERSITY
Inspiring Excellence

Hidden Message

1. The initiation of replication is mediated by **DnaA**.
2. **DnaA:** a protein that binds to a short segment within the *oriC* known as a **DnaA** box.
3. DnaA box as a message within the DNA sequence telling DnaA: "bind here!"

Hidden Message Problem

1. **Input:** A string *Text*.
2. **Output:** A hidden message in *Text*.

BRAC UNIVERSITY
Inspiring Excellence

Hidden Message - The Gold Bug

```
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;+?;
```

Hidden Message - The Gold Bug

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;+?;

Hidden Message - The Gold Bug

```
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 Problem

1. Lets find frequent words within *oriC*.

ACAACTATGCATACTATCGGGAACTATCCT.

COUNT(ACAACTATGCATACTATCGGGAACTATCCT, ACTAT) = 3.

2. **k-mer** is a string of length k and Count(*Text*, *Pattern*) is the number of times that a k-mer Pattern appears as a substring of *Text*.

PATTERNCOUNT(*Text*, *Pattern*)

```
1  Count = 0
2  for i = 0 to |Text| − |Pattern|
3        if Text.subString(i, |Pattern|) == Pattern
4              Count = Count + 1
5  return Count
```

BRAC UNIVERSITY
Inspiring Excellence

Frequent Words Problem

Pattern is a **most frequent k-mer** in *Text* if it maximizes *Count*(*Text*, *Pattern*) among all *k*-mers.

ACTAT is a most frequent 5-mer of ACA**ACTAT**GCAT**ACTAT**CGGGA**ACTAT**CCT
ATA is a most frequent 3-mer of CG**ATATA**TCC**ATA**G.

Is it possible for a string to have multiple most frequent k-mers?

Frequent Words Problem

1. **Input:** A string *Text* and a integer *k*.
2. **Output:** All most frequent *k*-mers in *Text*.

ROSALIND:1A https://rosalind.info/problems/ba1a/

BRAC UNIVERSITY
Inspiring Excellence

A Straight Forward Algorithm

FrequentWords($Text, k$)

```
1  FrequentPatterns = φ
2  for i = 0 to |Text| − k
3        Pattern = Text.subString(i, k)
4        Count[i] = PatternCount(Text, Pattern)
5        maxCount = maximum value in array Count
6  for i = 0 to |Text| − k
7        if Count[i] == maxCount
8              FrequenPatterns.add(Text.subString(i, k))
9  return Count
```

Text	**A**	**C**	**T**	G	**A**	**C**	**T**	C	C	C	A	C	C	C	C
COUNT	**2**	1	1	1	**2**	1	1	3	1	1	1	3	3		

Vibrio cholerae

atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttttacggaagaatgatcaagctgctgctcttgatcatcgtttc

k	3	4	5	6	7	8	9
count	25	11	8	8	5	4	3
k-mers	tga	atga	gatca	tgatca	atgatca	atgatcaa	atgatcaag
		tgat	tgatc				cttgatcat
							tcttgatca
							ctcttgatc

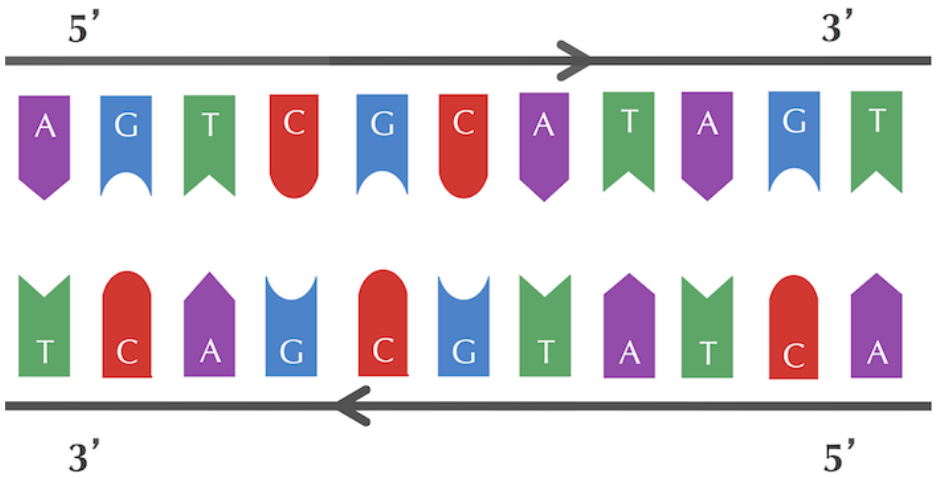
Frequent 9-mers

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaaga**ATGATCAAG**ctgctgctcttgatcatcgtttc

1. Four different 9-mers repeated three or more times in this region: ATGATCAAG, CTTGATCAT, TCTTGATCA, and CTCTTGATC
2. Does very low likelihood indicates something?
3. Which one of them? Is any of them more surprising?

BRAC UNIVERSITY
Inspiring Excellence

More Surprises: Reverse Complement



Reverse Complement Problem

1. **Input:** A DNA string *Pattern*.
2. **Output:** $\overline{Pattern}$, reverse complement of *Pattern*.

ROSALIND:1C https://rosalind.info/problems/ba1c/

BRAC UNIVERSITY
Inspiring Excellence

More Surprises: Reverse Complement

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

Pattern Matching Problem

- How frequent are they in the whole Vibrio cholerae genome?
- Are other short regions in the Vibrio cholerae genome exhibiting multiple occurrences of ATGATCAAG (or CTTGATCAT).

Pattern Matching Problem

Find all occurrences of a pattern in a string.

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

ROSALIND:1D https://rosalind.info/problems/ba1d/

ATGATCAAG appears 17 times in the following starting positions:
116556, 149355, **151913, 152013, 152394**, 186189, 194276, 200076, 224527, 307692, 479770, 610980, 653338, 679985, 768828, 878903, 985368

151913, 152013, and 152394, form clumps, i.e., appear close to each other in a small region of the genome.

BRAC UNIVERSITY
Inspiring Excellence

What about other genomes?

Thermotoga petrophila, a bacterium that strives in extremely hot environments(80 degree).

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

AACCTACCA AAACCTACC ACCTACCAC

CCTACCACC GGTAGGTTT TGGTAGGTT

Another Genome

Thermotoga petrophila, a bacterium that strives in extremely hot environments(80 degree).

aactctatacctccttttttgtcgaatttgtgtgatttatagagaaaatcttattaactga
aactaaaatggtaggtttGGTGGTAGGttttgtgtacattttgtagtatctgatttttaa
ttacataccgtatattgtattaaattgacgaacaattgcatggaattgaatatatgcaaa
acaaaCCTACCACCaaactctgtattgaccattttaggacaacttcagGGTGGTAGGttt
ctgaagctctcatcaatagactattttagtctttacaaacaatattaccgttcagattca
agattctacaacgctgttttaatgggcgttgcagaaaacttaccacctaaaatccagtat
ccaagccgatttcagagaaacctaccacttacctaccactaCCTACCACCcgggtggta
agttgcagacattattaaaaacctcatcagaagcttgttcaaaaatttcaatactcgaaa
CCTACCACCtgcgtcccctattatttactactactaataatagcagtataattgatctga

Clump Finding Problem

find every k-mer that forms a clump in the genome!

1. Given integers L and t, a k-mer Pattern forms an (L, t)-clump inside a (larger) string Genome if there is an interval of Genome of length L in which this k-mer appears at least t times.

gatcagcataagggtccC**TGCA**A**TGCA**TGACAAGCC**TGCA**GTtgttttac

Clump Finding Problem

Find patterns forming clumps in a string.

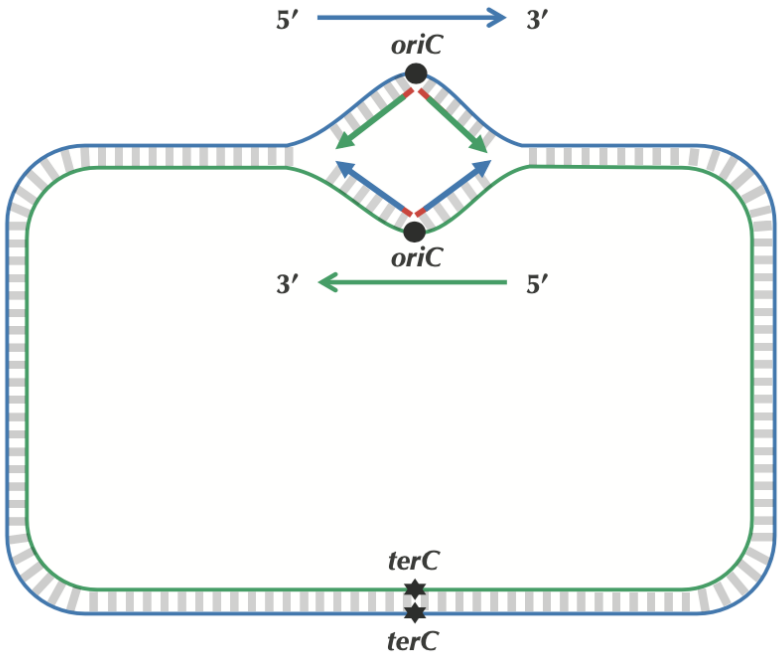
1. **Input:** String *Genome* and integers k, L and t.
2. **Output:** All distinct k-mers forming (L, t)-clumps in Genome.

ROSALIND:1E https://rosalind.info/problems/ba1e/
In the Escherichia coli (E. coli) genome, hundreds of different 9-mers forms (500, 3)-clumps, and it is absolutely unclear which of these 9-mers might represent a DnaA box.

BRAC UNIVERSITY
Inspiring Excellence

A Closer Look into Replication

- Two complementary DNA strands running in opposite directions around a circular chromosome unravel, starting at oriC.
- 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.



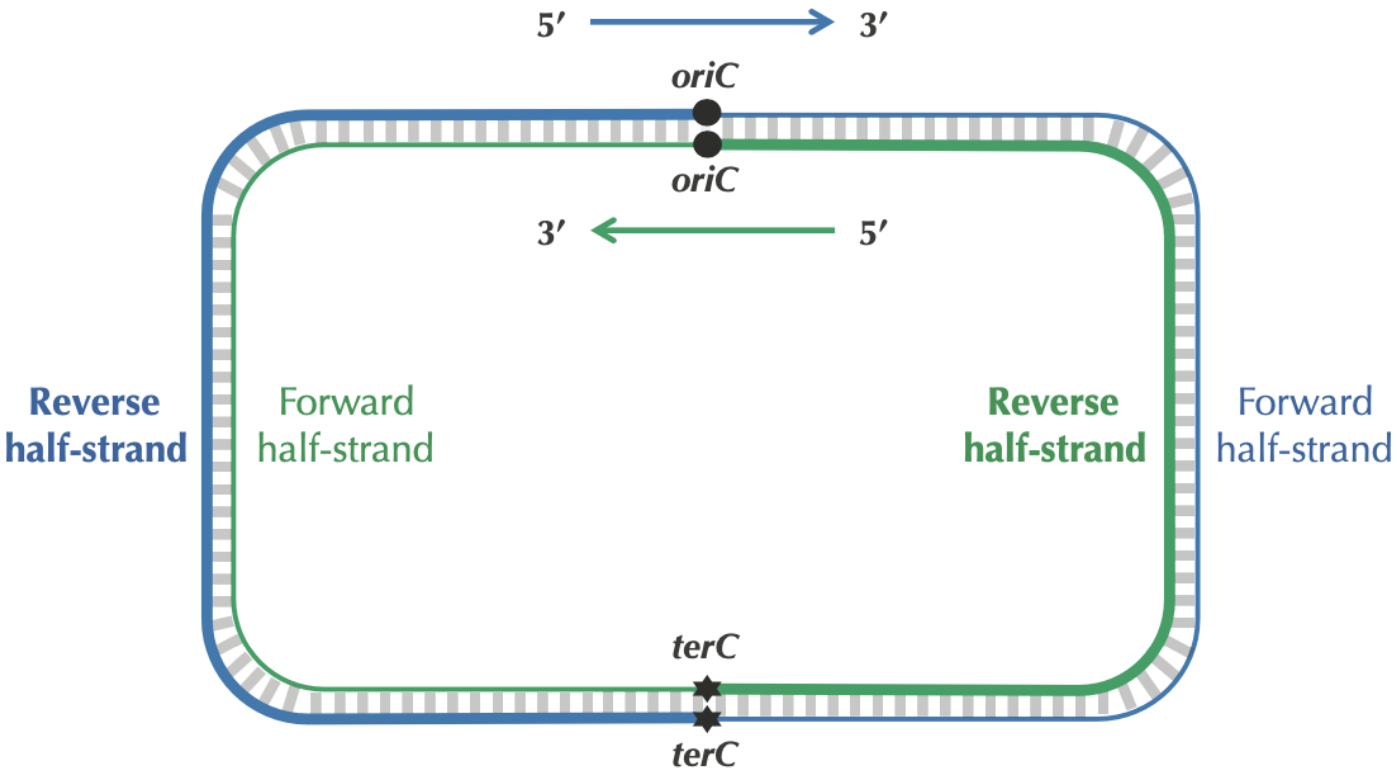
BRAC UNIVERSITY
Inspiring Excellence

DNA Replication - More Details

- 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 polymerases, 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 that binds to the parent strand and jump starts the DNA polymerase.
- 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.

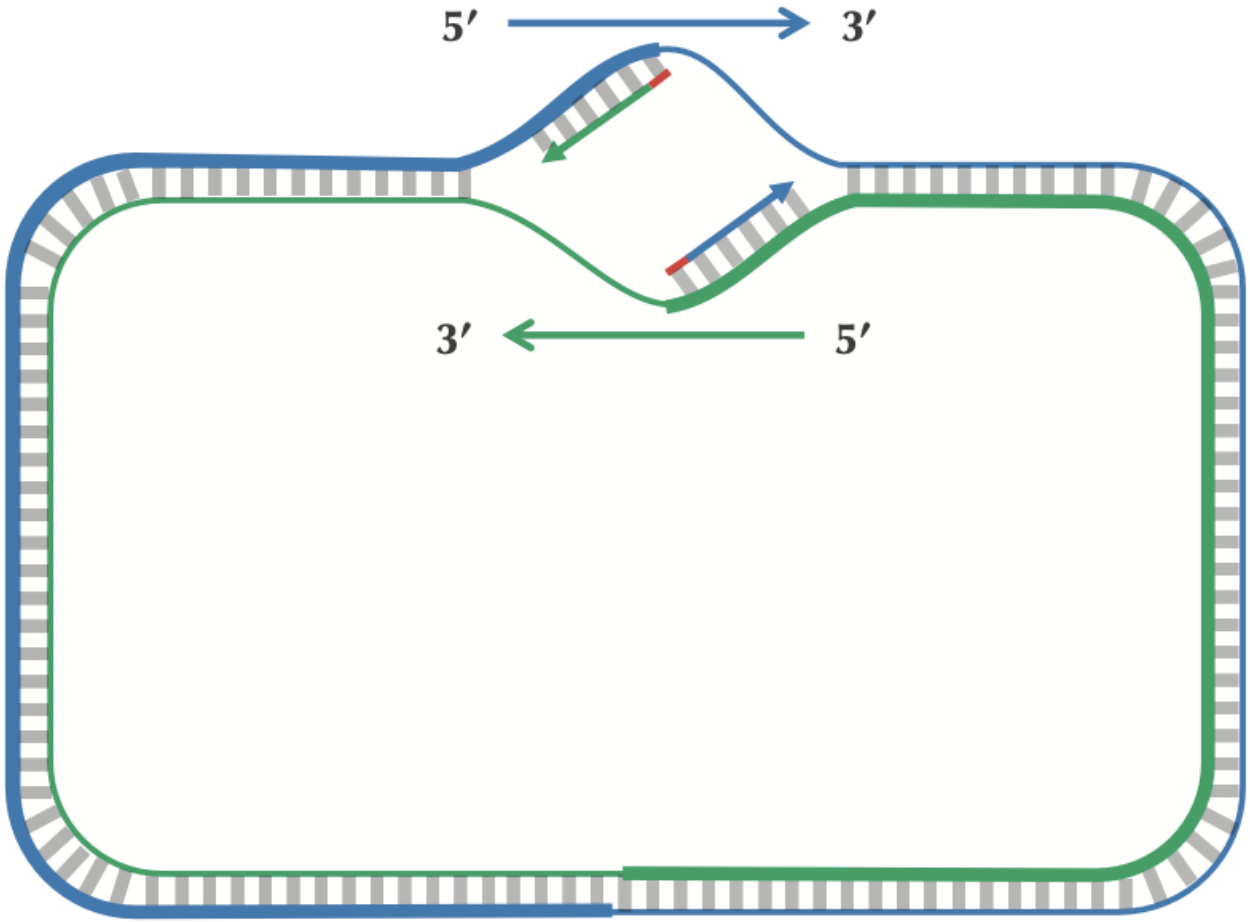
Directionality!

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



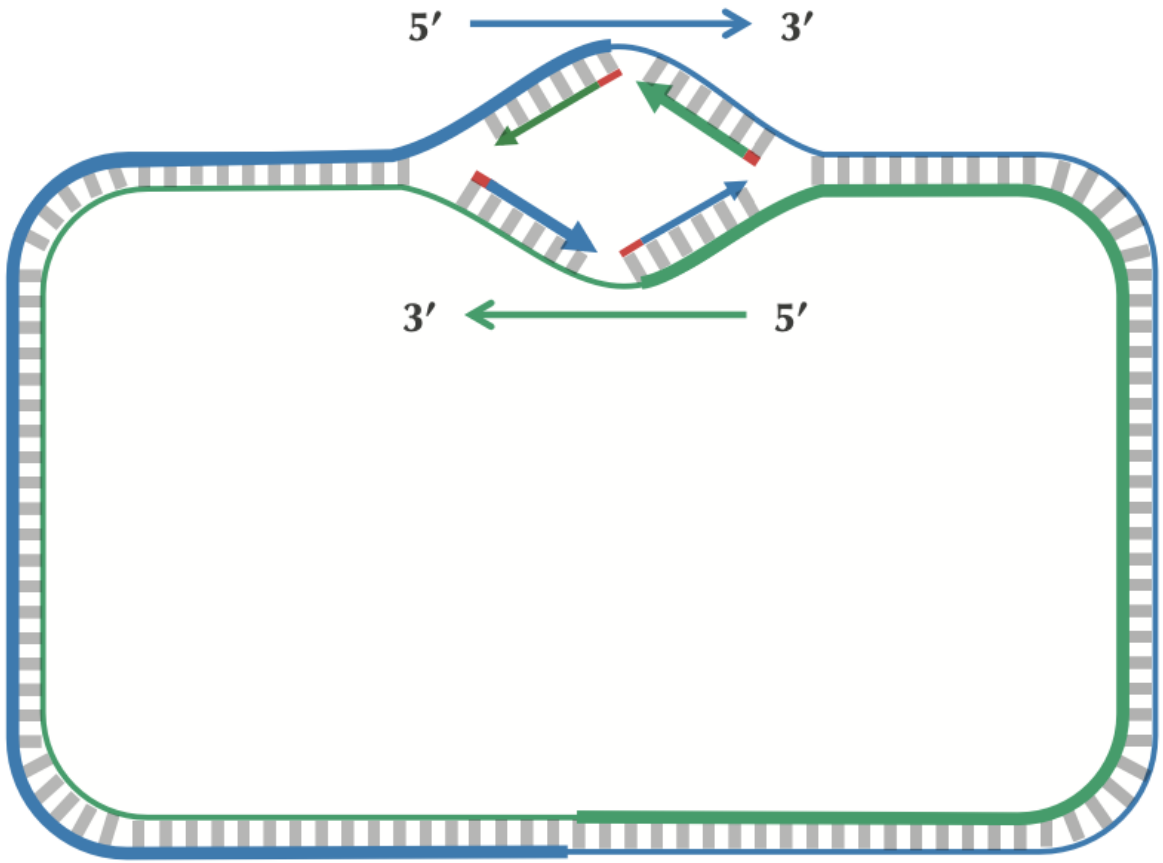
Asymmetry of replication

- Since a DNA polymerase can only move in the reverse (3' $\rightarrow$ 5') direction, it can copy nucleotides non-stop from oriC to terC along reverse half-strands.



Asymmetry of replication

- Replication on forward half-strands is very different because a DNA polymerase cannot move in the forward (5' $\rightarrow$ 3') direction; on these half-strands, a DNA polymerase must replicate backwards toward oriC.



BRAC UNIVERSITY
Inspiring Excellence

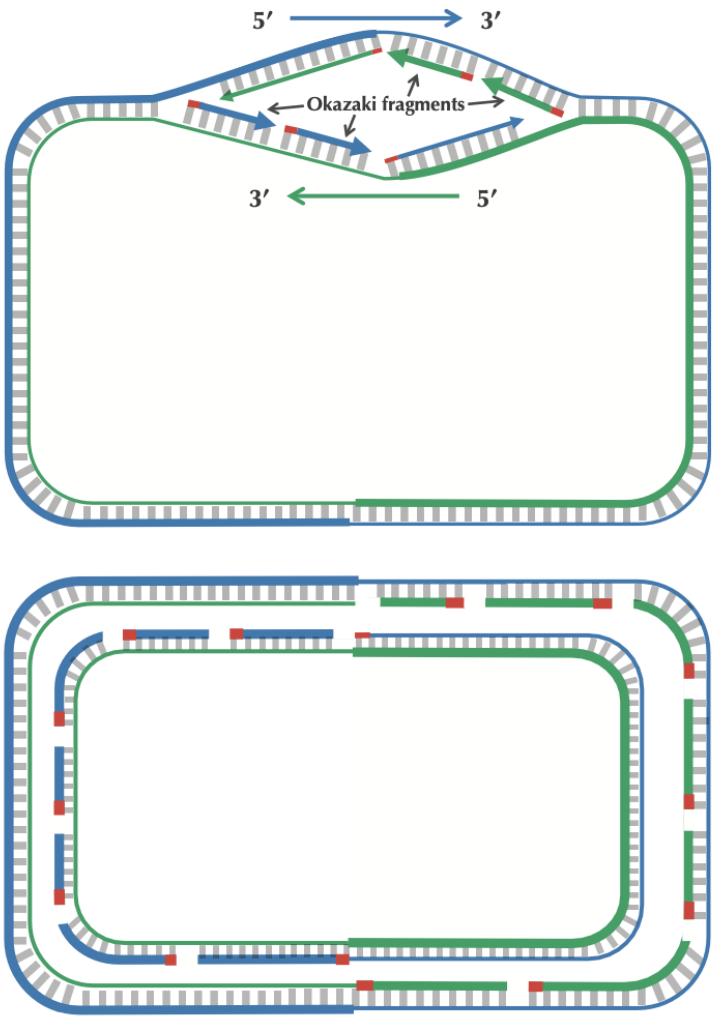
Okazaki Fragments

- 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.
- 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, which results in the synthesis of short **Okazaki fragments** that are complementary to intervals on the forward half-strand.

BRAC UNIVERSITY
Inspiring Excellence

Okazaki Fragments

- Consecutive Okazaki fragments are sewn together by an enzyme called DNA ligase.
- In reality, DNA ligase does not wait until after all the Okazaki fragments have been replicated to start sewing them together.
- Biologists call a reverse half-strand a leading half-strand since a single DNA poly- merase traverses this half-strand non-stop, and they call a forward half-strand a lagging half-strand since it is used as a template by many DNA polymerases stopping and starting replication.



Nucleotide Counts: Deamination

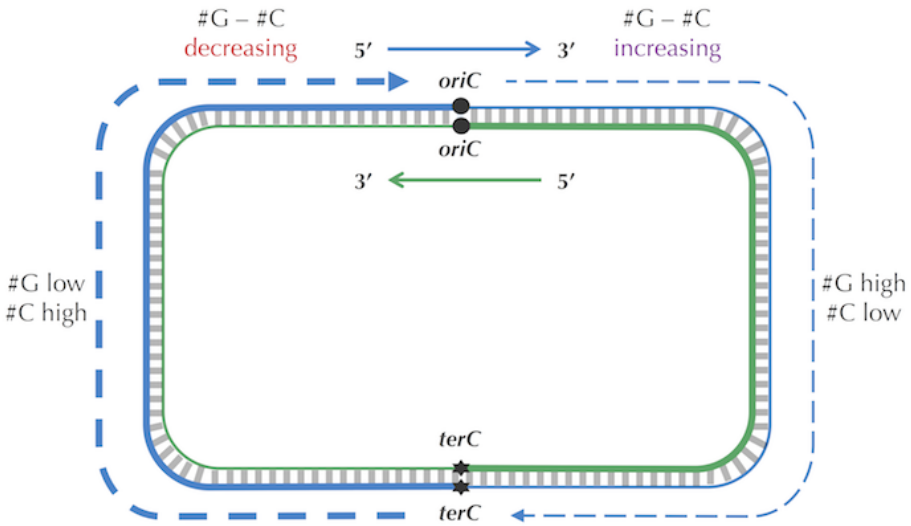
- Single-stranded DNA has a much higher mutation rate than double-stranded DNA. The nucleotide counts for Thermotoga petrophila:

	#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

1. Cytosine (C) has a tendency to mutate into thymine (T)
2. Decrease in cytosine on the forward half-strand forms mismatched base pairs T-G.
3. Mismatched pairs can further mutate into T-A pairs when the bond is repaired in the next round of replication, which accounts for the observed decrease in guanine (G) on the reverse half-strand.

BRAC UNIVERSITY
Inspiring Excellence

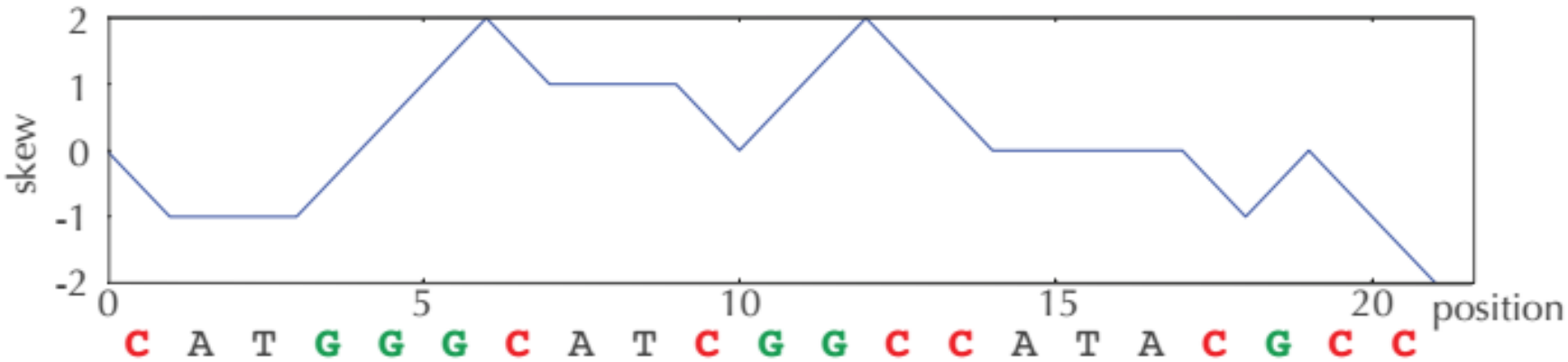
Deamination



- Difference between the total amount of guanine and the total amount of cytosine is negative on the reverse half-strand and positive on the forward half-strand.
- If this difference starts increasing, then we guess that we are on the forward half-strand; on the other hand, if this difference starts decreasing, then the we guess that we are on the reverse half-strand.

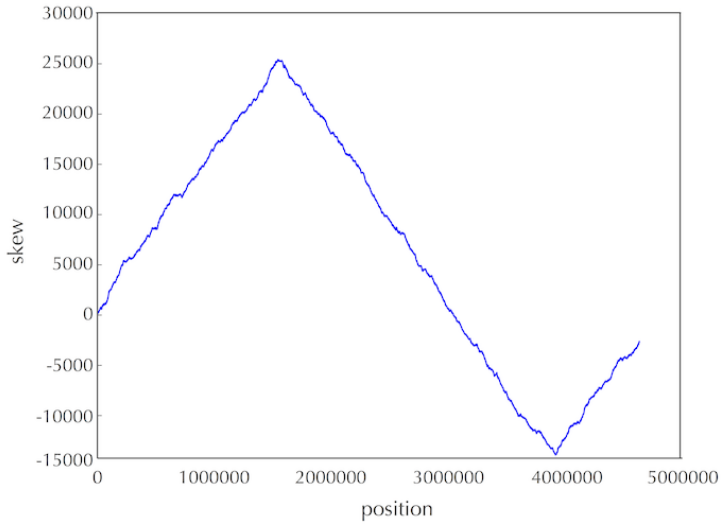
BRAC UNIVERSITY
Inspiring Excellence

The Skew Diagram



- We define $Skew_i(Genome)$ as the difference between the total number of occurrences of G and the total number of occurrences of C in the first i nucleotides of Genome.
- Where is *oriC*?

Skew Diagram for E. Coli



Minimum Skew Problem

Find a position in a genome minimizing the skew.

1. **Input:** A dna String *Genome*.
2. **Output:** All integer(s) i minimizing $Skew_i(Genome)$ among all values of i (from 0 to $|Genome|$).

ROSALIND:1F https://rosalind.info/problems/ba1f/

BRAC UNIVERSITY
Inspiring Excellence

More Elusive - OriC of E. coli

No 9-mers (along with their reverse complements) that appear three or more times! But, occurrences with approximate matches !

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

Approximate Matching

- We say that position i in k-mers $p_1, \cdots, p_k$ and $q_1, \cdots, q_k$ is a mismatch if $p_i \neq q_i$.
- CGAAT and CGGAC have two mismatches.
- The number of mismatches between strings p and q is called the Hamming distance between these strings and is denoted $HammingDistance(p, q)$.

Hamming Distance Problem

Compute the Hamming distance between two strings.

1. **Input:** Two strings of equal length.
2. **Output:** The Hamming distance between these strings.

BRAC UNIVERSITY
Inspiring Excellence

Approximate Matching

We say that a k-mer *Pattern* appears as a substring of *Text* with at most d mismatches if there is some k-mer substring *Pattern*$'$ of *Text* having d or fewer mismatches with *Pattern*, i.e.,

$$HammingDistance(Pattern, Pattern') \leq d$$

Approximate Pattern Matching Problem

Find all approximate occurrences of a pattern in a string.

1. **Input:** Strings *Pattern* and *Text* along with an integer d.
2. **Output:** All starting positions where *Pattern* appears as a substring of *Text* with at most d mismatches.

ROSALIND:1H http://rosalind.info/problems/ba1h/

BRAC UNIVERSITY
Inspiring Excellence

Approximate Frequency

- Given strings *Text* and *Pattern* as well as an integer d, we define $Count_d(Text, Pattern)$ as the total number of occurrences of *Pattern* in *Text* with at most d mismatches.
- $Count_1(AACAAGCTGATAAACATTTAAAGAG, AAAAA) = 4$ because AAAAA appears four times in this string with at most one mismatch: AACAA, ATAAA, AAACA, and AAAGA. Note that two of these occurrences overlap.
- Exercise: Compute $Count_2(AACAAGCTGATAAACATTTAAAGAG, AAAAA)$.

BRAC UNIVERSITY
Inspiring Excellence

Approximate Matching

```
ApproximatePatternCount(Text, Pattern, d)
1  Count = 0
2  for i = 0 to |Text| − |Pattern|
3        Pattern′ = Text.subString(i, |Pattern|)
4        if HammingDistance(Pattern, Pattern′) ≤ d
5              Count = Count + 1
6  return Count
```

Frequent Words with Mismatches

- A most frequent k-mer with up to d mismatches in *Text* is simply a string *Pattern* maximizing $Count_d(Text, Pattern)$ among all k-mers.
- Note that *Pattern* does not need to actually appear as a substring of *Text*.
- For example, as we already saw, AAAAA is the most frequent 5-mer with 1 mismatch in AACAAGCTGATAAACATTTAAAGAG, even though it does not appear exactly in this string.

Frequent Words with Mismatches Problem

Find the most frequent k-mers with mismatches in a string.

1. **Input:** A string *Text* as well as integers k and d.
2. **Output:** All most frequent k-mers with up to d mismatches in Text.

ROSALIND:1I http://rosalind.info/problems/ba1i/

BRAC UNIVERSITY
Inspiring Excellence

Approximate Matching

Frequent Words with Mismatches and Reverse Complements Problem

Find the most frequent k-mers (with mismatches and reverse complements) in a DNA string.

1. **Input:** A DNA string *Text* as well as integers k and d.
2. **Output:** All k-mers Pattern maximizing the sum $Count_d(Text, Pattern) + Count_d(Text, \overline{Pattern})$ over all possible k-mers.

ROSALIND:1J http://rosalind.info/problems/ba1j/

BRAC UNIVERSITY
Inspiring Excellence