أكاديمية كاوست
KAUST ACADEMY

Advanced Bioinformatics

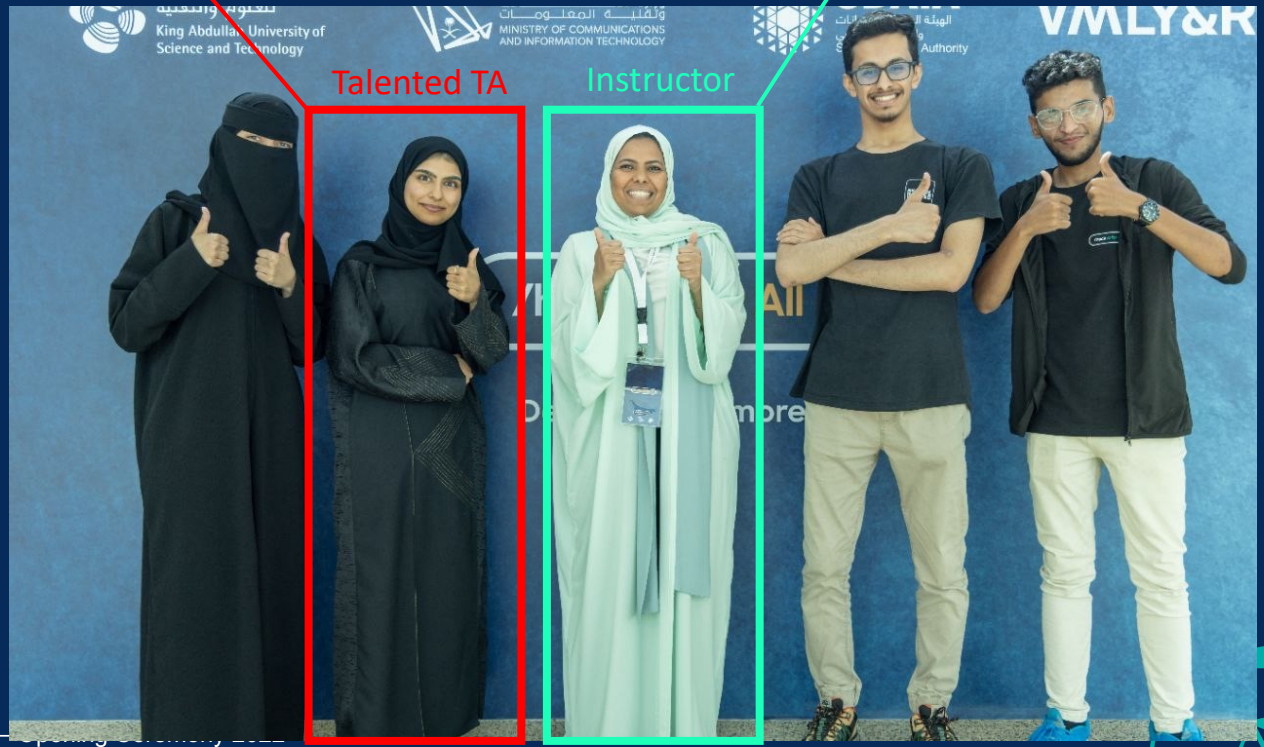
Instructor :Sakhaa Alsaedi
Sakhaa.Alsaedi@kaust.edu.sa

Day 1: Genome Replication (Part 1)
25th Feb. 2024

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

- Ebtihal Hani
- B.s in Artificial Intelligence
- Alumni-KAUST Academy (AI program)

- Sakhaa Alsaedi
- Ph.D. Candidate in Computer Science Program at KAUST
- Researcher at CBRC, KAUST Smart-Health Initiative
- Lecturer at Taibah University,



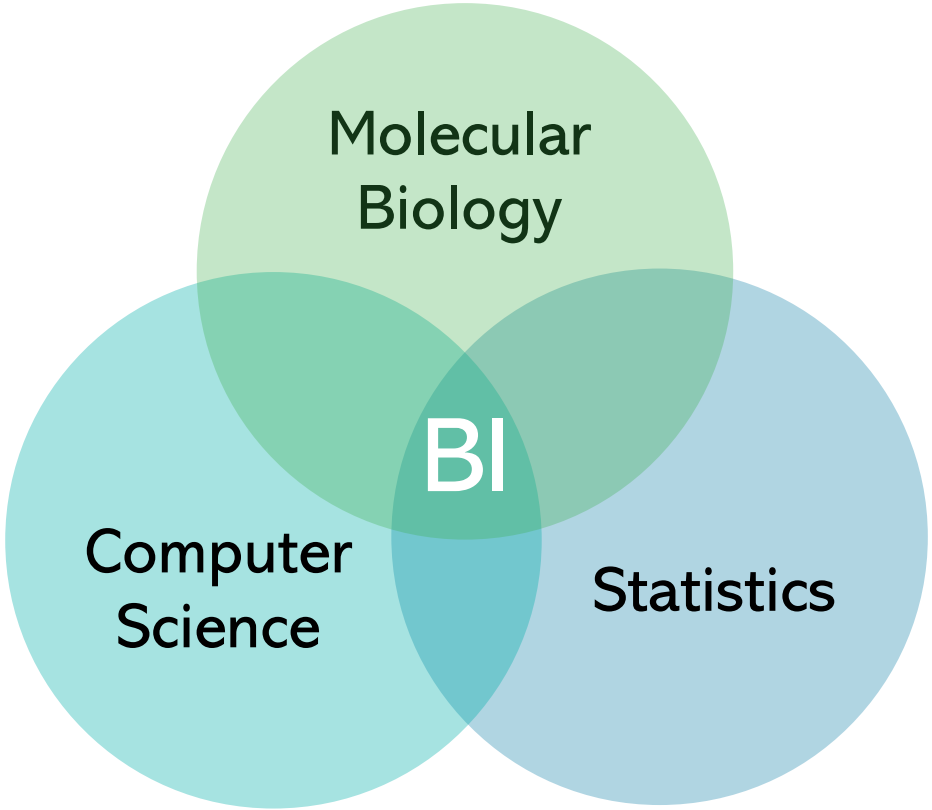
Outlines

- Overview in Bioinformatics
- Introduction to Python
- Introduction to Genome Replication
- Genome Replication Problem
- Bioinformatics Challenges with using python

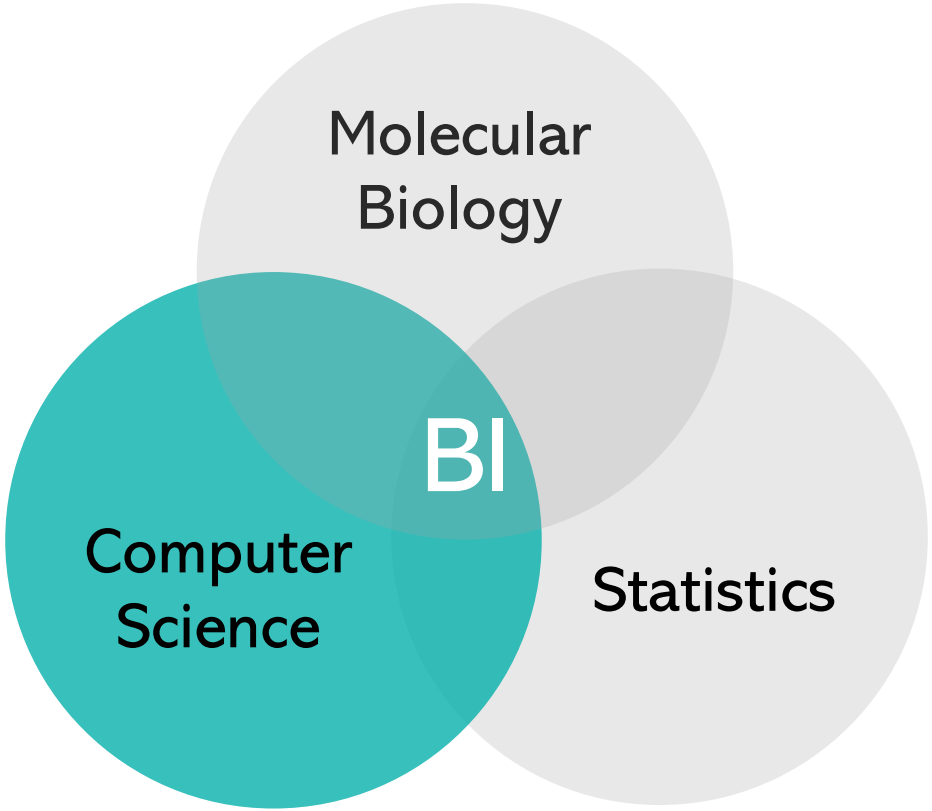
Outlines

- Overview in Bioinformatics
- Introduction to Python
- Introduction to Genome Replication
- Genome Replication Problem
- Bioinformatics Challenges with using python

Bioinformatics



Bioinformatics



Outlines

- Overview in Bioinformatics
- Introduction to Python
- Introduction to Genome Replication
- Genome Replication Problem
- Bioinformatics Challenges with using python

Introduction to Python

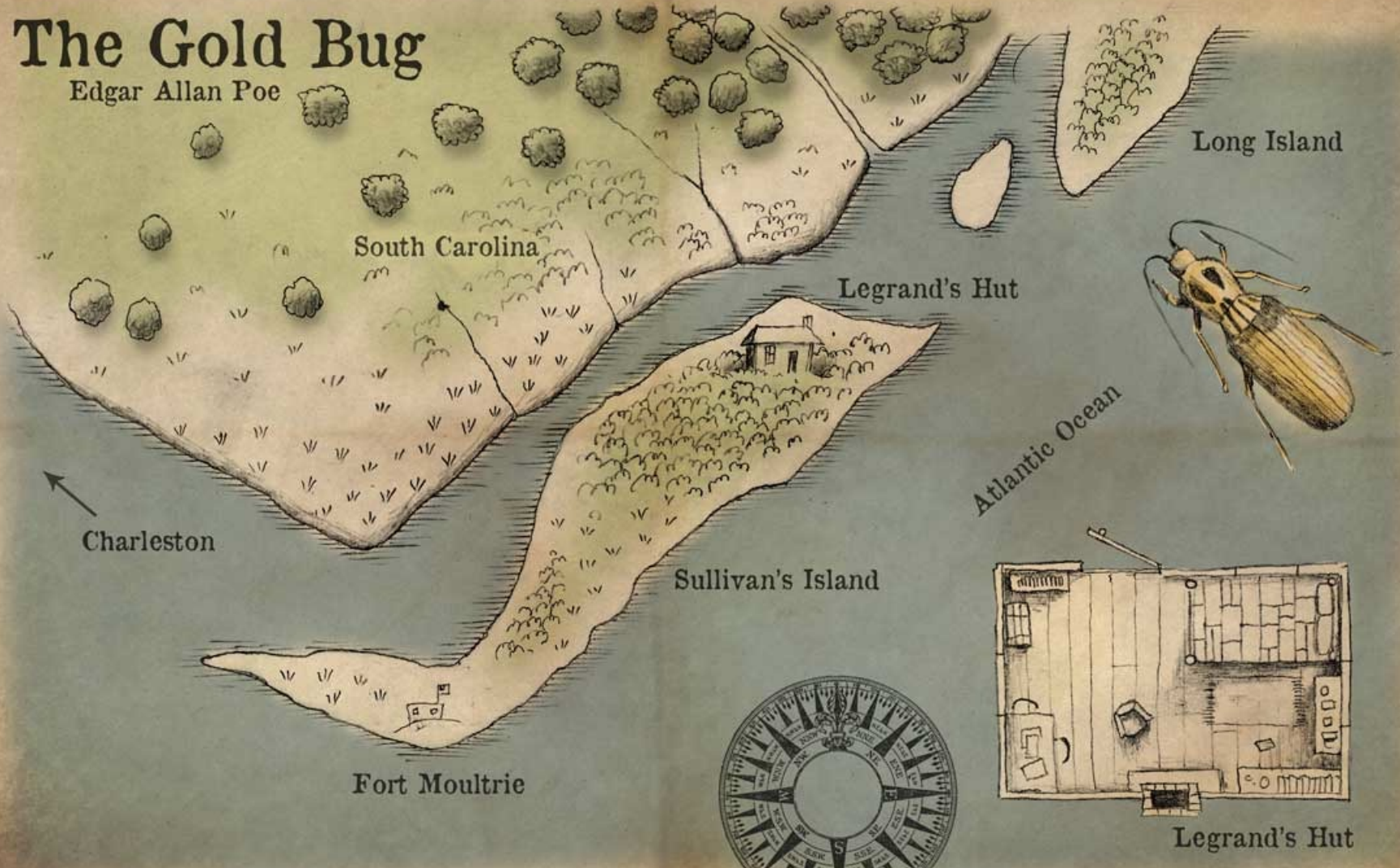
GitHub

- Basic Syntax and Variables
- Operators and Expressions
- Data Types and Structures
- Control Structures and Functions
- Basic Statistics and Analysis

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

The Gold Bug

Edgar Allan Poe



“The Gold-Bug” Problem

أكاديمية كاوست
KAUST ACADEMY

```
53++!305))6*;4826)4+.)4+);806*;48
!8`60))85;]8*:+*8!83(88)5*!;46(;8
8*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 secret message left by pirates

(“The Gold-Bug” by Edgar Allan Poe)

Why is “;48” so Frequent?

أكاديمية كاوست
KAUST ACADEMY

```
53++!305))6*;4826)4+.)4+);806*;48
!8`60))85;]8*:+*8!83(88)5*!;46(;8
8*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;
+?;
```

Hint: The message is in English

“THE” is the Most Frequent English Word

أكاديمية كاوست
KAUST ACADEMY

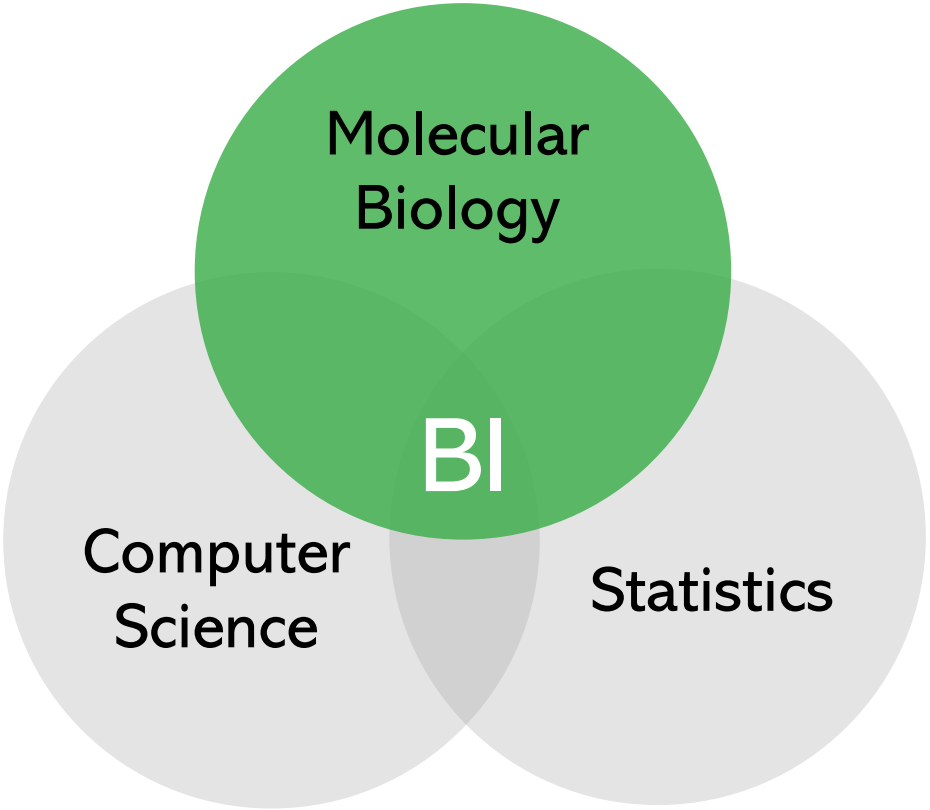
53++!305))6*THE26)4+.)4+)806*THE!
8`60))85;]8*:+*8!83(88)5*!;46(;88
96?;8)*+(THE5);5*!2:*+(;4956*2(
5*4)8`8*;4069285);)6!8)4++;1(+9TH
E081;8:8+1THE!85;4)485!528806*81(
+9THE;(88;4(+?34THE)4+;161;:188;+
?;

Could you Complete Decoding the Message?

أكاديمية كاوست
KAUST ACADEMY

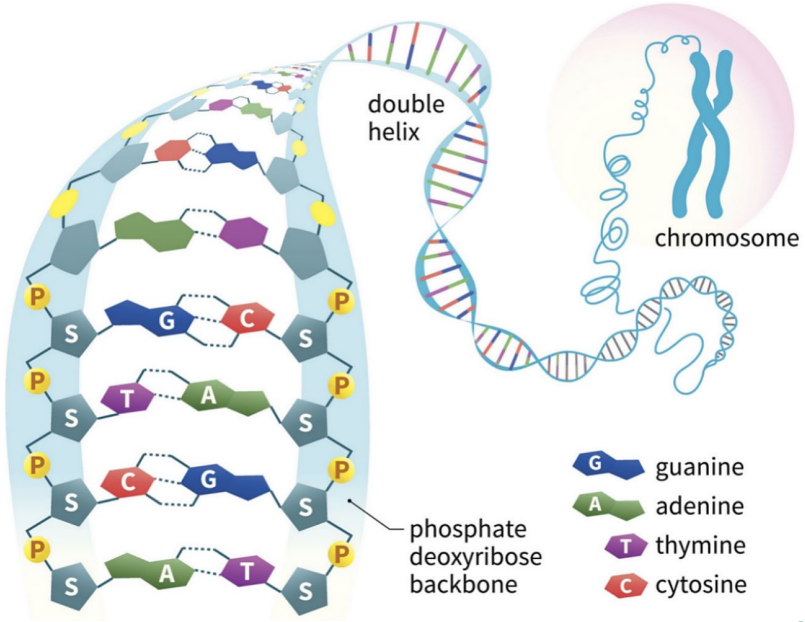
53++!305))6*THE26)H+.)H+)806*THE!
E`60))E5;]E*:+*E!E3(EE)5*!TH6(TEE
96?;E)*+(THE5)T5*!2:*+(TH956*2(
5*H)E`E*TH0692E5)T)6!E)H++T1(+9TH
E0E1TE:E+1THE!E5T4)HE5!52880 6*E1(
+9THET(EETH(+?34THE)H+T161T:1EET+
?T

Bioinformatics

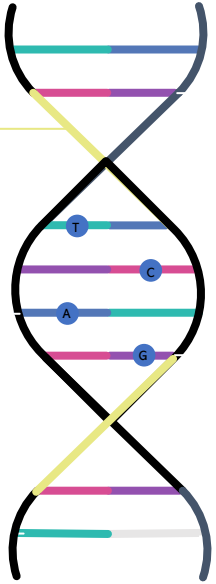


Genome

- Cell → Nucleus
- Nucleus → Chromosomes
- Chromosomes→ DNA (Genome)
- DNA→ genes (Coding –noncoding)
- Genes → Nucleotide acid



DNA (Double Helix)



Non-Coding

Thymine

Adenine

Single helix

Coding

Cytosine

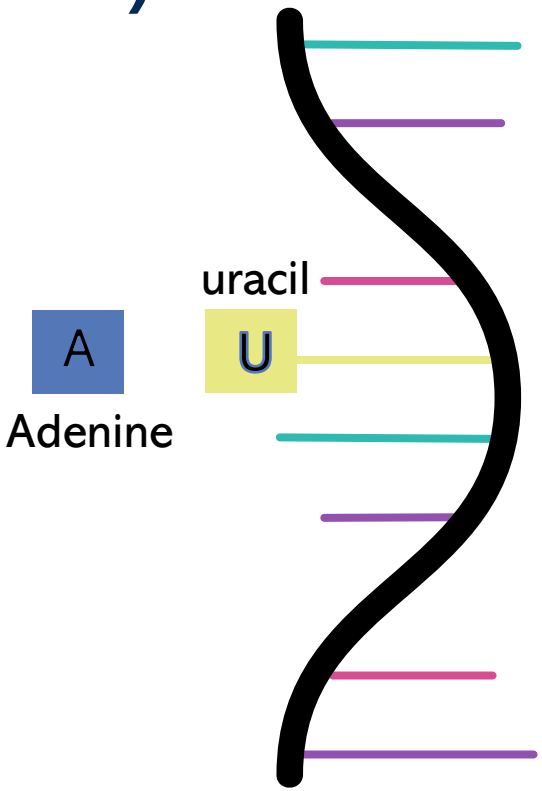
Guanine

Gene

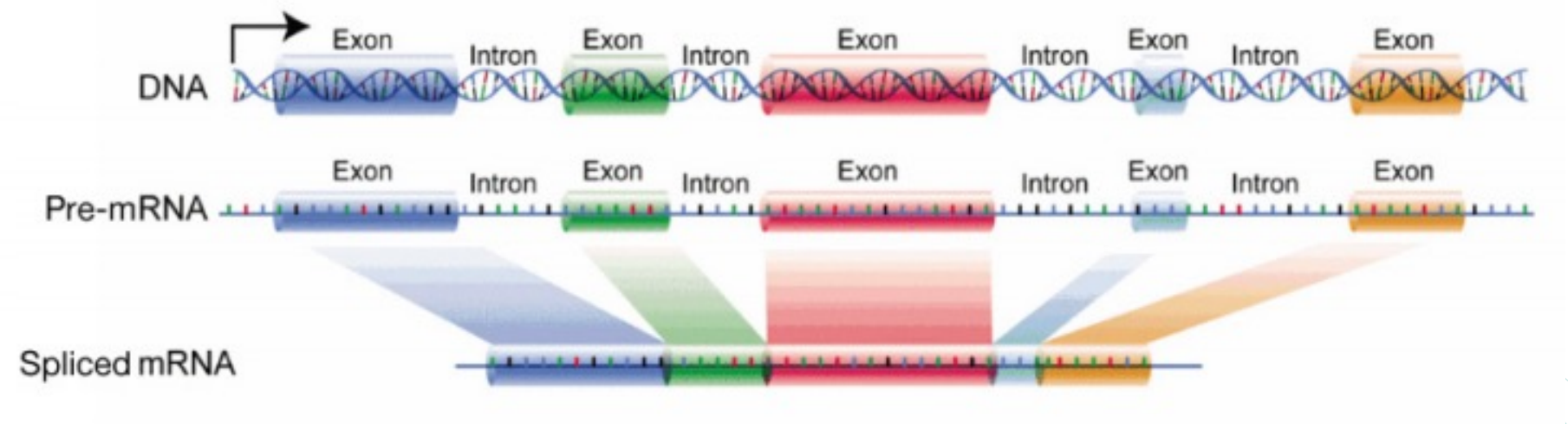
DNA segment

Single helix

RNA (Single Helix)



Quick Summary of Nucleotide Structure

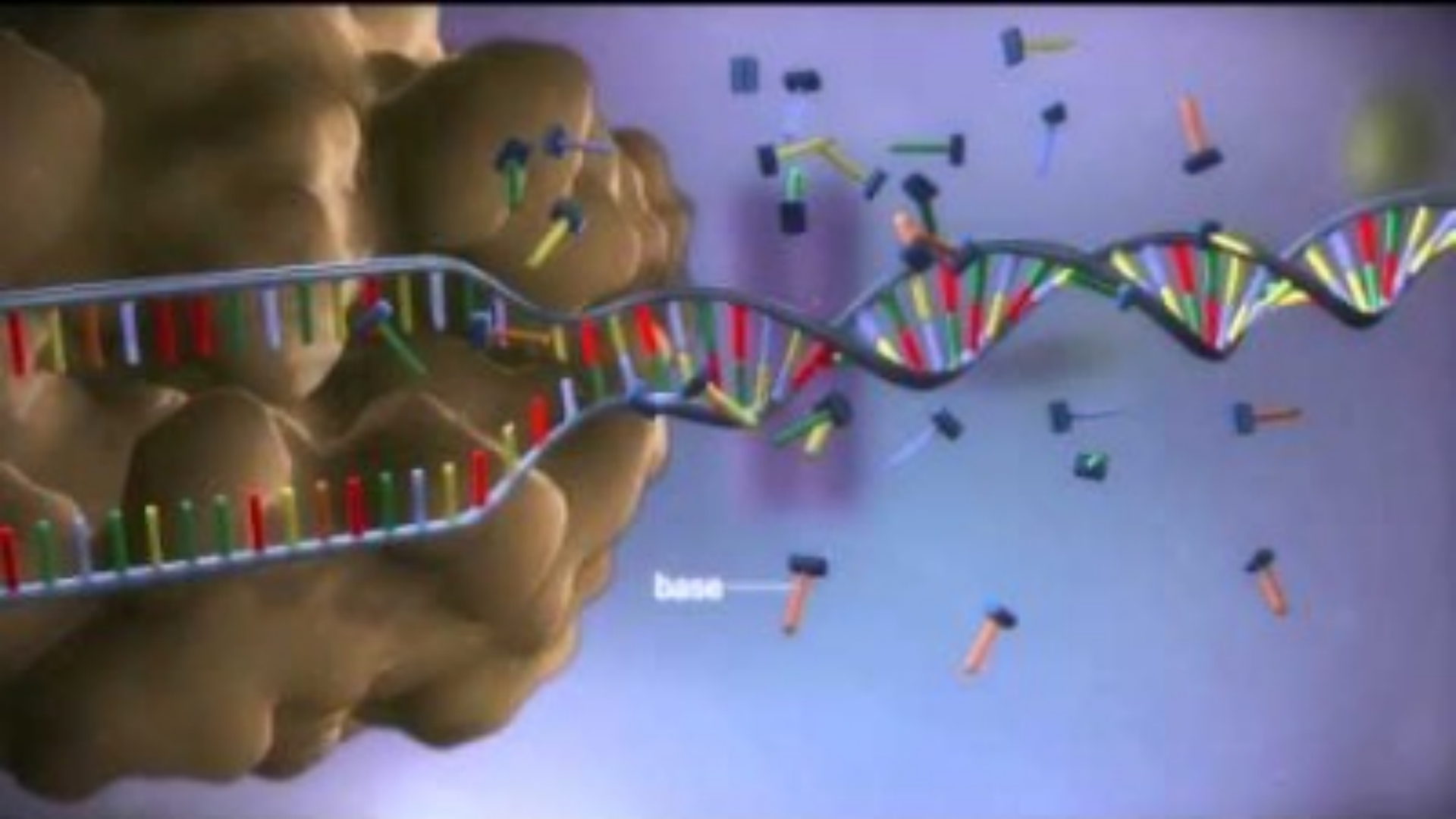


The Central Dogma Main Processes

DNA Replication

RNA Transcription

RNA Translation


base

Outlines

- Overview in Bioinformatics
- Introduction to Python
- Introduction to Genome Replication
- Genome Replication Problem
- Bioinformatics Challenges with using python

The Central Dogma Main Processes

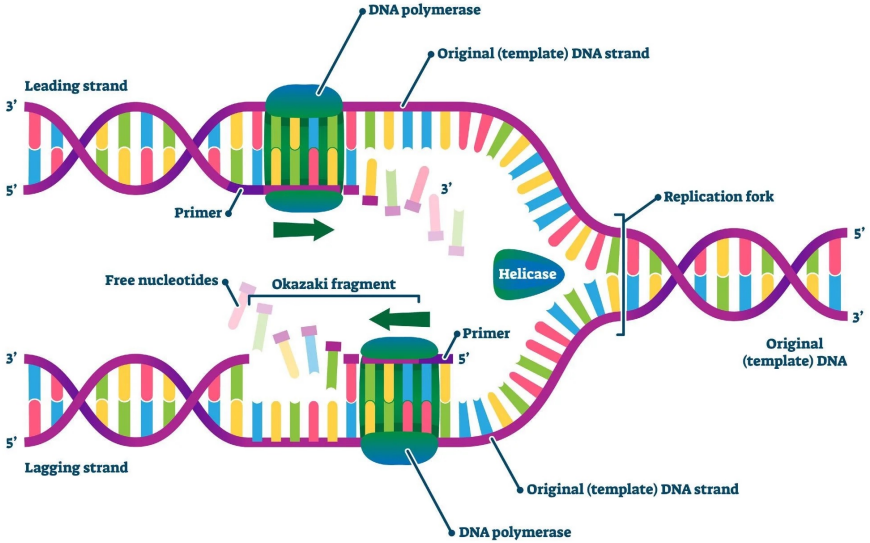
DNA Replication

RNA Transcription

RNA Translation

DNA Replication and Repair: Understanding how DNA is copied (replication) and maintained (repair) is crucial for maintaining genetic stability

- Helicase
- DNA polymerase



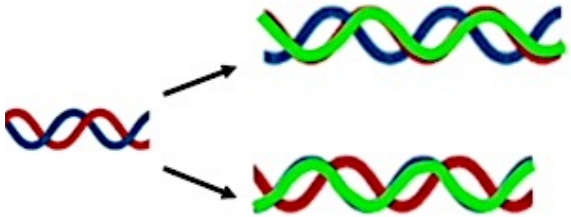
Significance of DNA Replication

أكاديمية كاوست
KAUST ACADEMY

- Is the process by which DNA makes a copy of itself
- Accurate transmission of genetic information from parent to offspring
- Essential for cell division during growth and repair of damaged tissue
- Errors - lead to mutations or changes in the genome and are associated with several clinical disorders

DNA Replication

- Replication is bidirectional
- Replication is Semiconservative
- New strands only synthesise in 5' → 3' direction

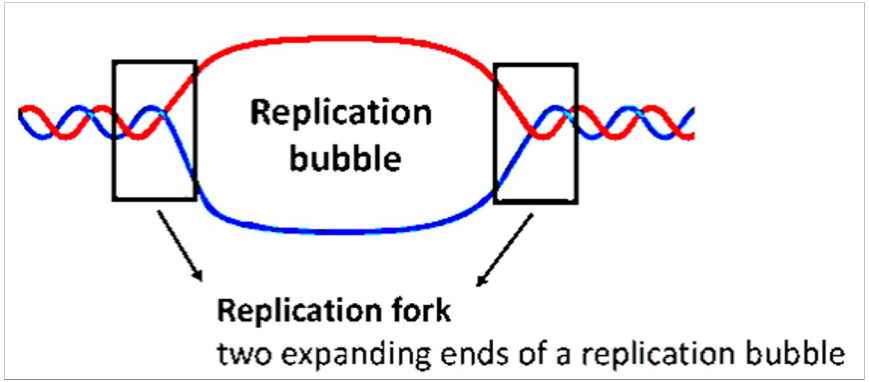


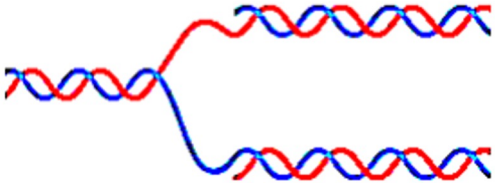
Stages of DNA Replication

1. **Initiation**

2. Elongation

3. Termination





أكاديمية كاوست
KAUST ACADEMY

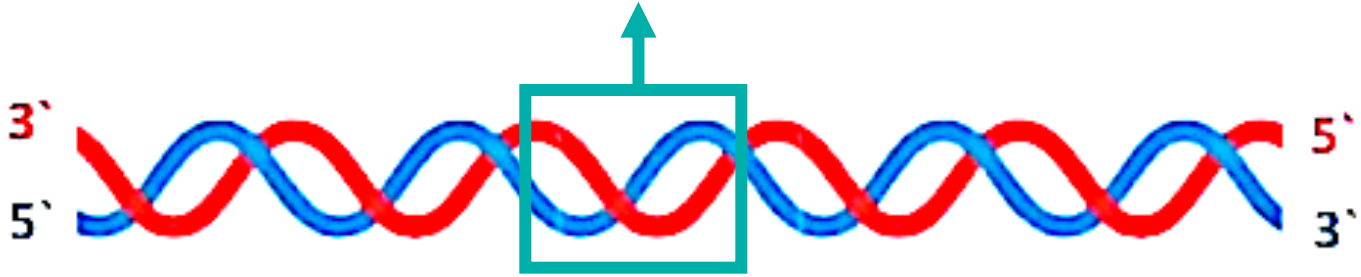
Where in a Genome Does DNA Replication Begin?

Think and share your thoughts :)

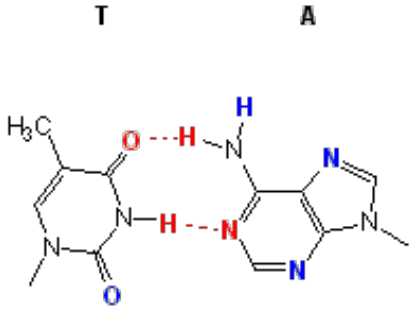
جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

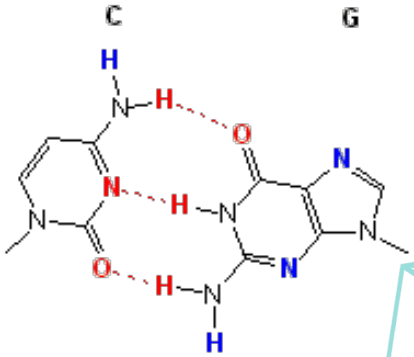
Where in a genome does it all begin?

- Replication → the replication origin (*oriC*)



A and T

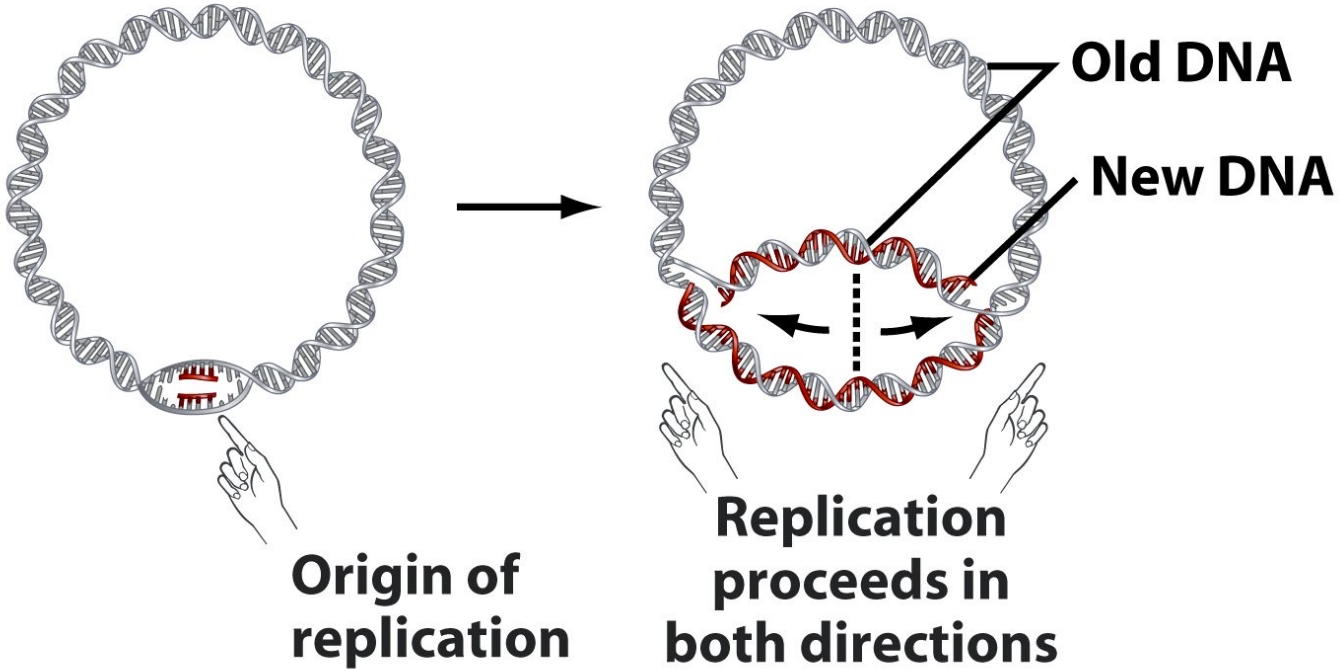




Where in a genome does it all begin?

أكاديمية كاوست
KAUST ACADEMY

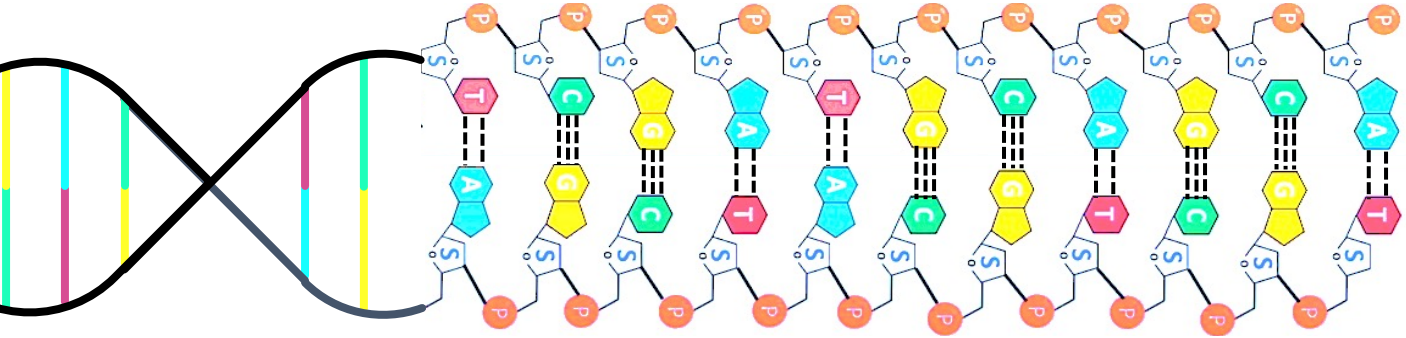
Replication begins in a region called the replication origin (*oriC*)

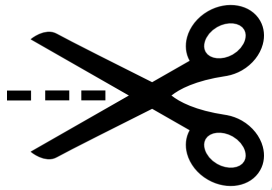


Separation of the two complementary DNA strands

أكاديمية كاوست
KAUST ACADEMY

- Helicase: Unwind DNA helix by disrupting hydrogen bonds

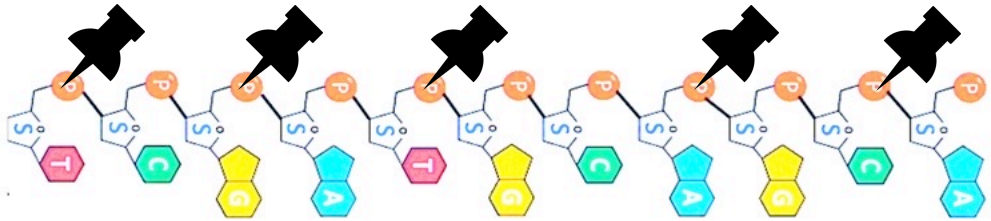


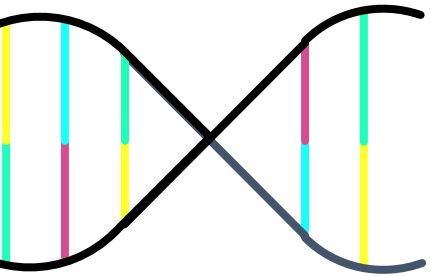


Separation of the two complementary DNA strands

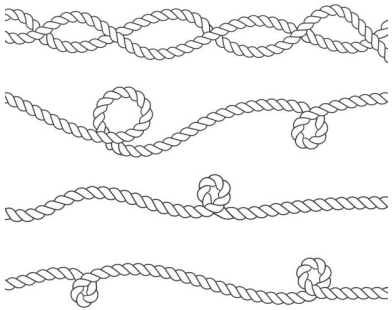
أكاديمية كاوست
KAUST ACADEMY

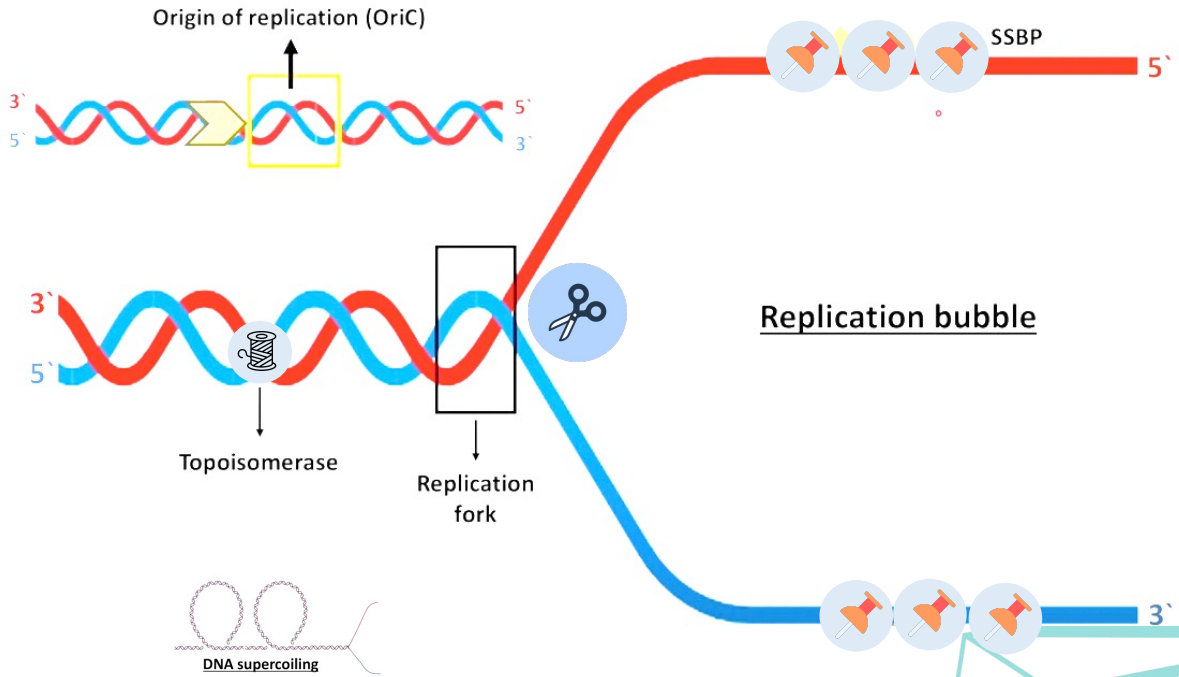
- Single-stranded DNA-binding protein (SSBP)

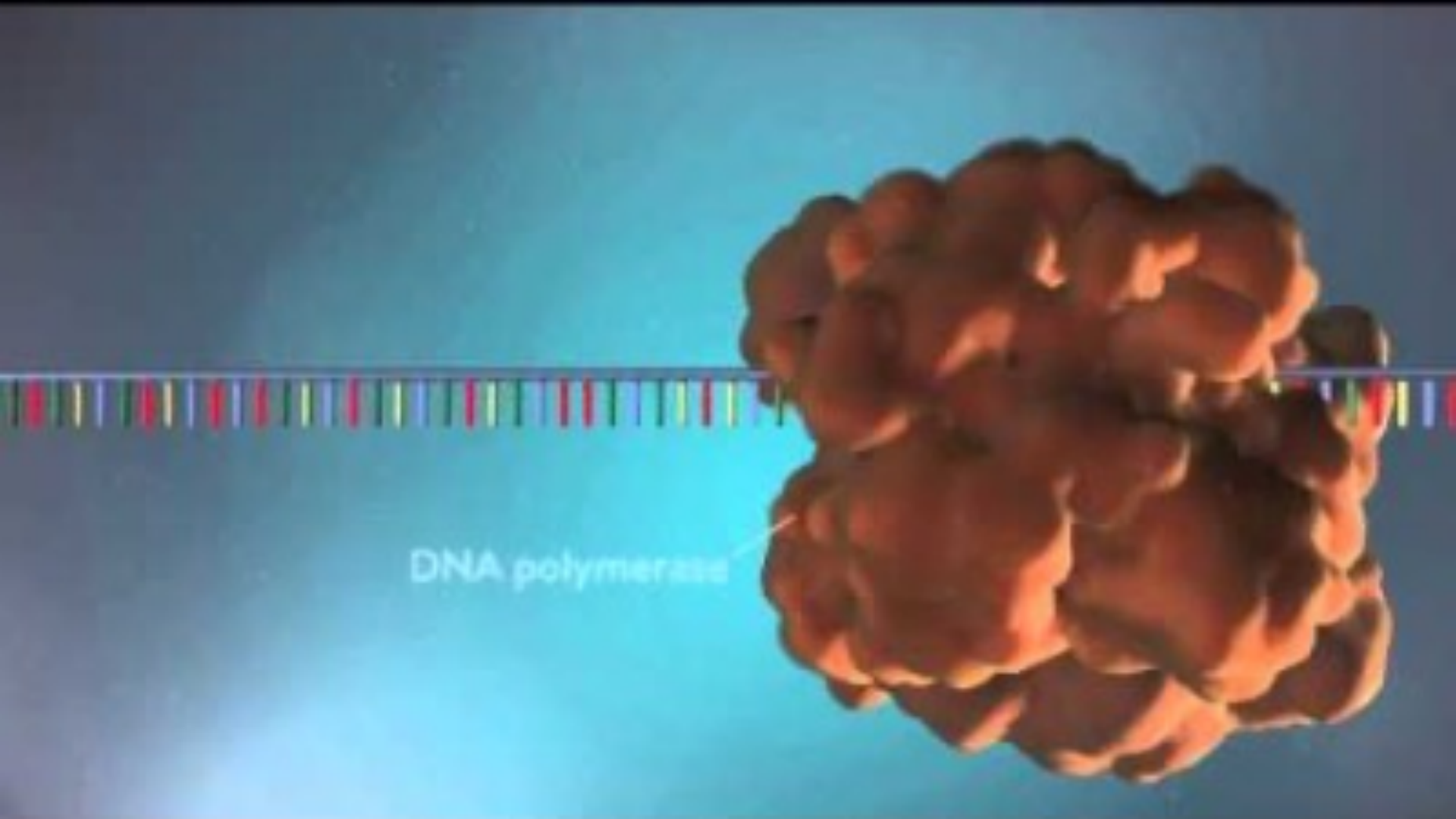




Topoisomerase: relieve torsional strain that results from helicase-induced unwinding





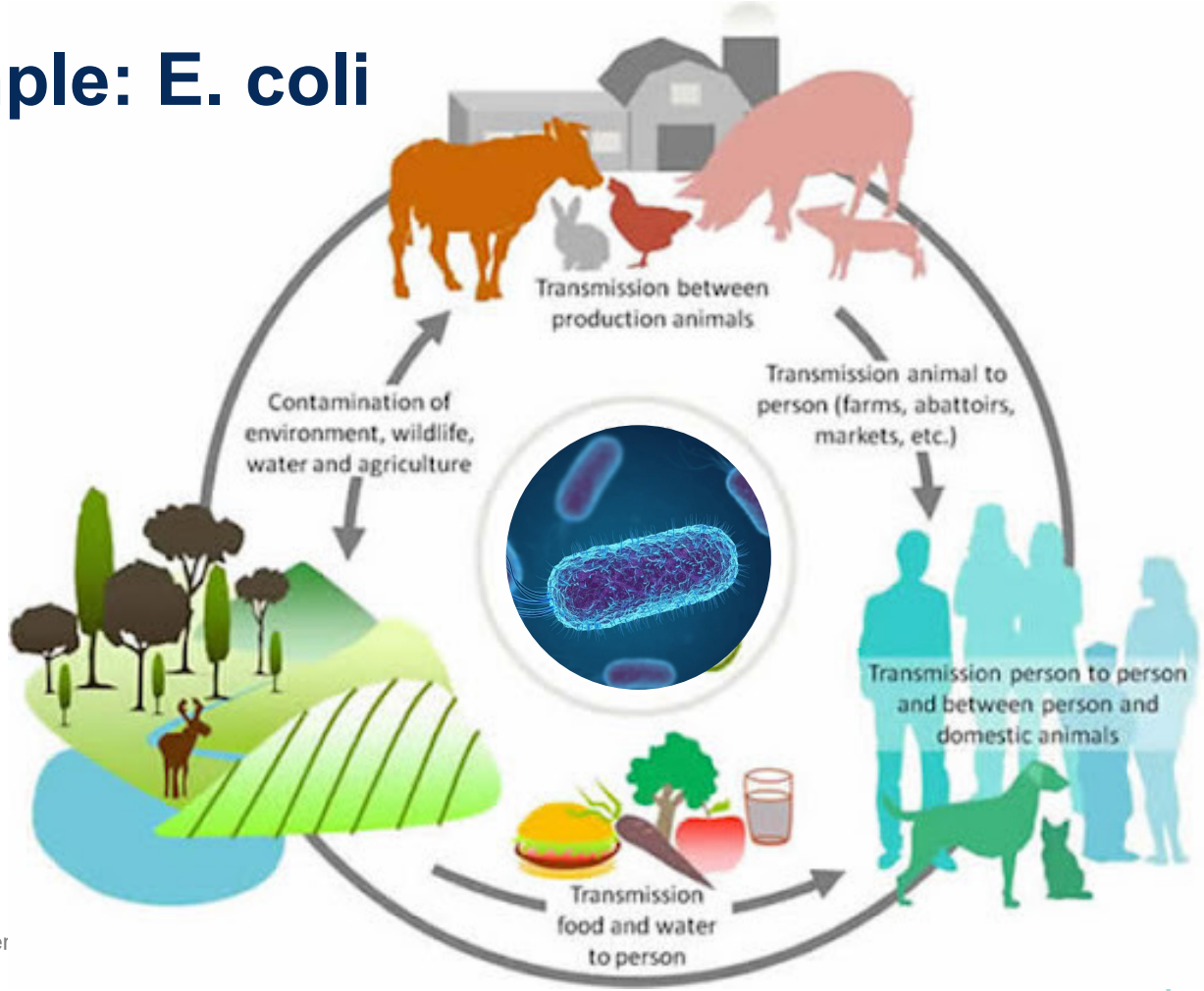

DNA polymerase

Number of Origins of Replication across Different Species

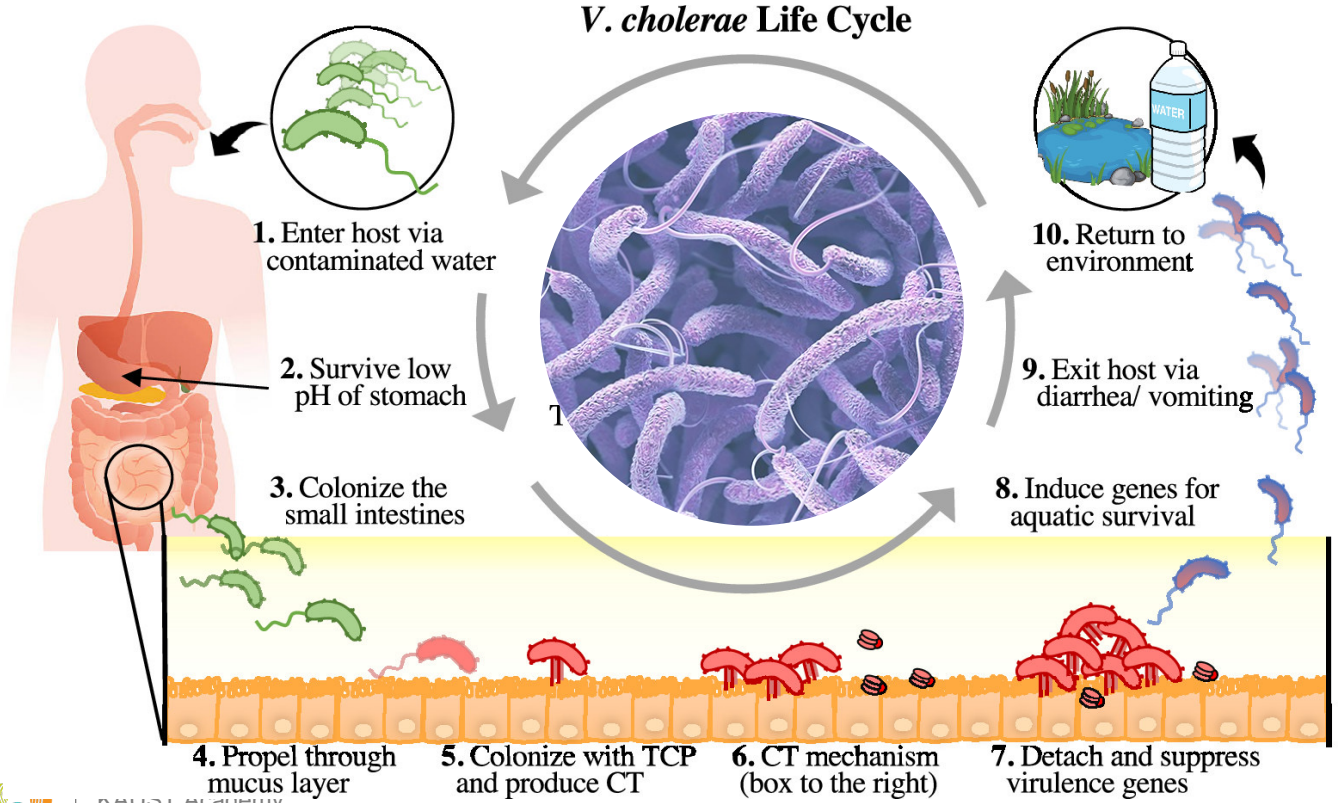
Species	Type of Organism	Genome Size	Number of Origins of Replication
Human	Eukaryote	~3 billion bp	Tens of thousands
Escherichia coli	Prokaryote	~4.6 million bp	1 (OriC)
Yeast	Eukaryote	~12 million bp	~400
Fruit fly	Eukaryote	~140 million bp	Several hundred to a few thousand

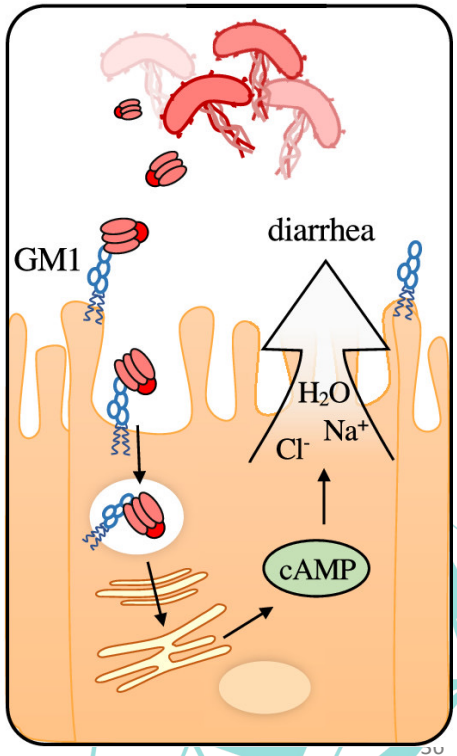
Metagenomics Example

Example: E. coli



Example: Vibrio Cholerae





Outlines

- Overview in Bioinformatics
- Introduction to Python
- Introduction to Genome Replication
- Genome Replication Problem
- Bioinformatics Challenges with using python

أكاديمية كاوست
KAUST ACADEMY

How to Find the Origin of Replication?

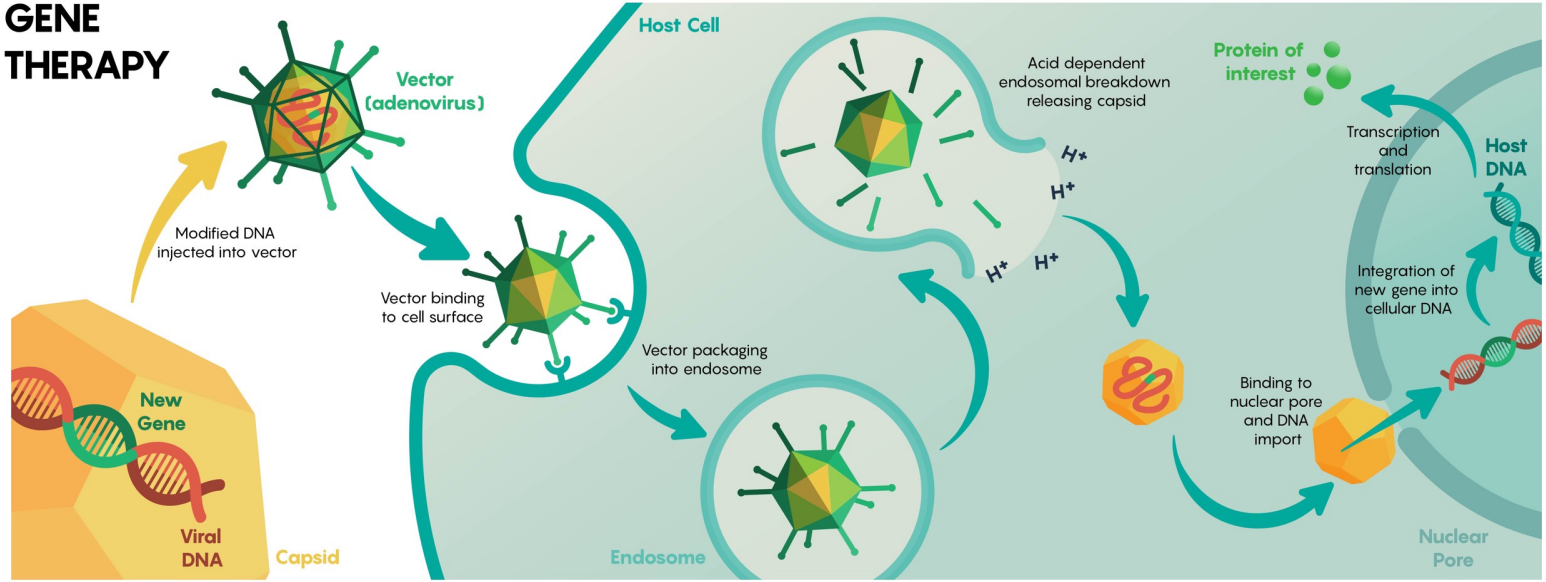
Think and share your thoughts :)

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

أكاديمية وست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Why Finding Origin of Replication is an important task ?



Viral vectors , some gene therapy methods use genetically engineered mini-genomes

Viral Vector Applications - OriC Replication

أكاديمية كاوست
KAUST ACADEMY

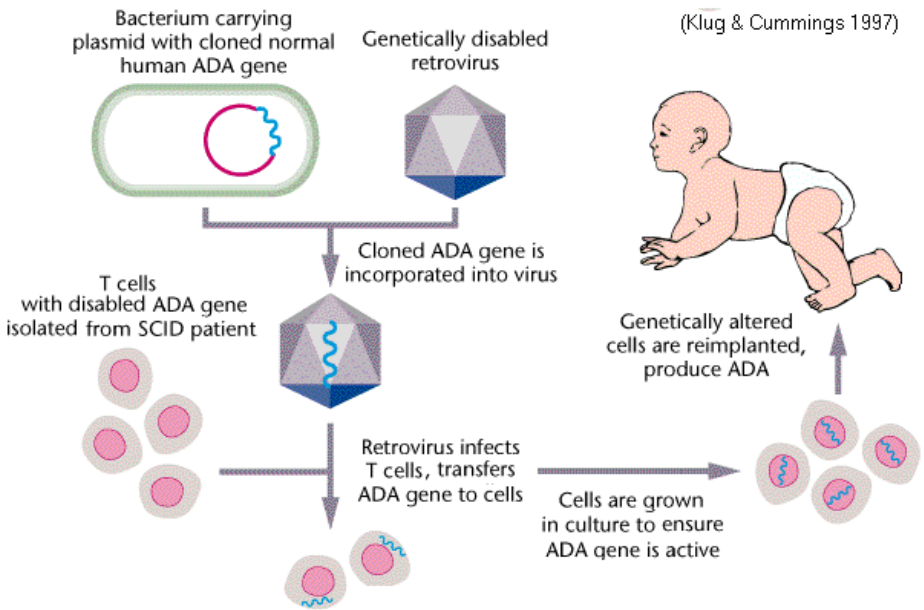
- **Gene tharpy** → Deliver therapeutic genes to patient cells
- **Vaccine development** → Cancer
- **Genetic engineering of plants** → Frost-resistant tomatoes
- **Basic genetics research** → Gene function and expression
- **Therapeutic protein production** → Upstream production
- **Gene Silencing** → Deliver RNAi constructs or CRISPR-Cas systems

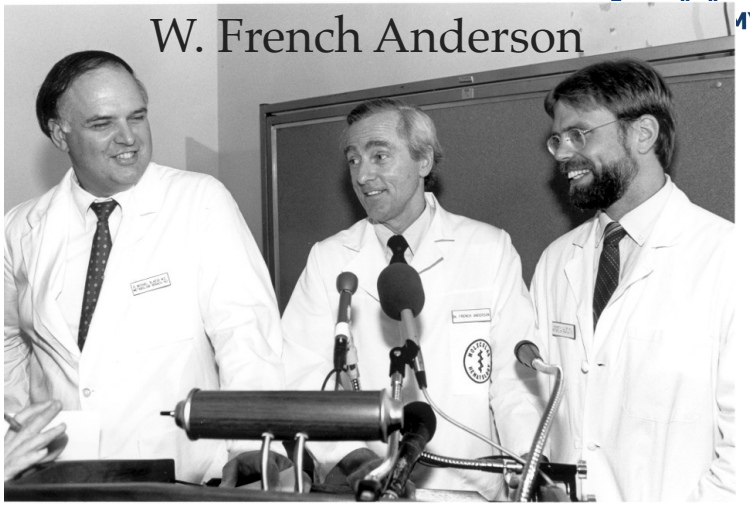
In 1992: The First Human Gene Therapy

David Vetter, The Bubble Boy

Severe Combined Immunodeficiency Disorder (SCID)

ADA-SCID Gene Therapy





- ADA-SCID

1990 — RV transduction of T cells
Safe but no evidence of efficacy

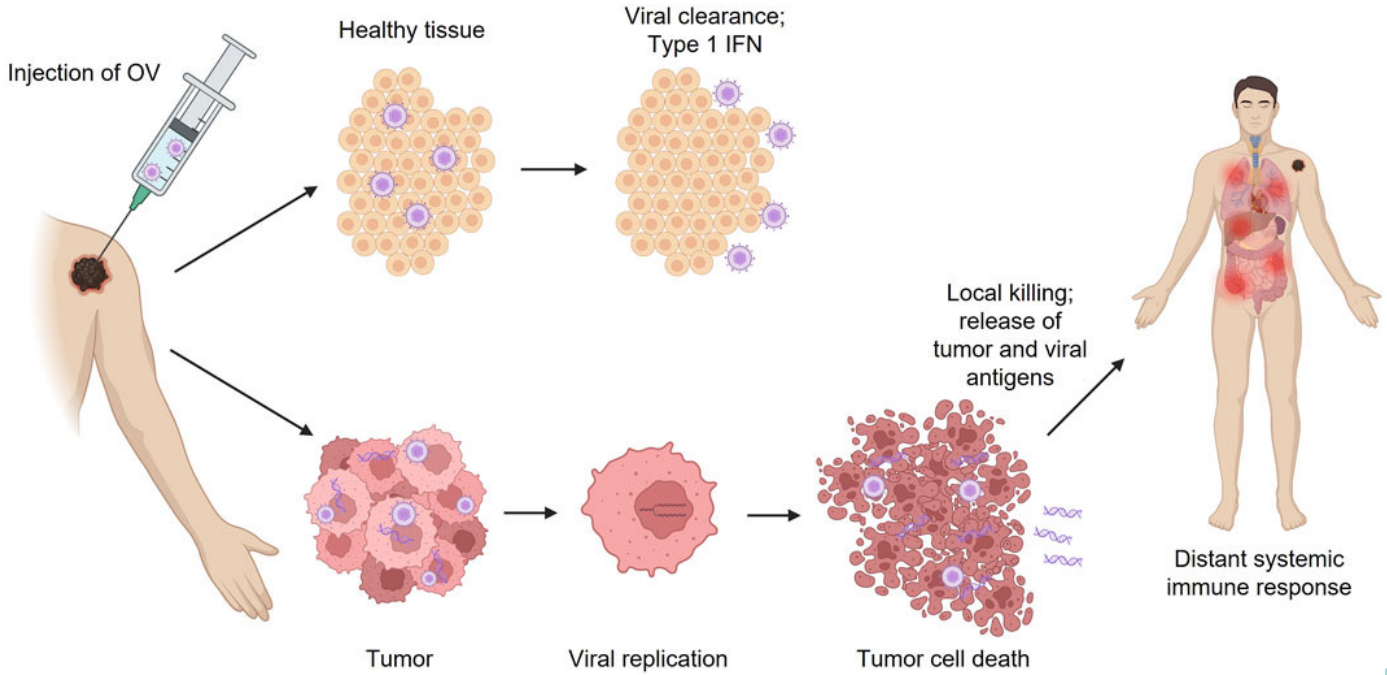
2000 — HSR, Milano, Italy
Modified trial shows therapeutic effect

2009 — Results published
NEJM Aiuti et al.

The First Human Gene Therapy

- Severe Combined Immunodeficiency Disorder

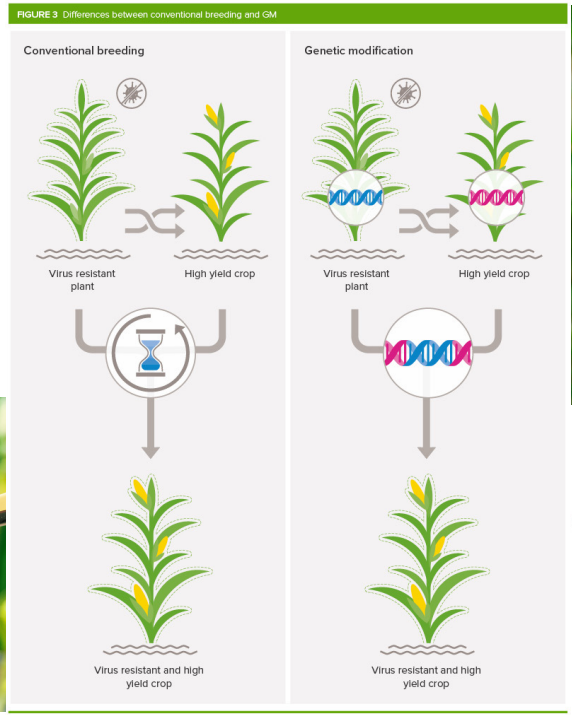
Cancer Vaccine Development

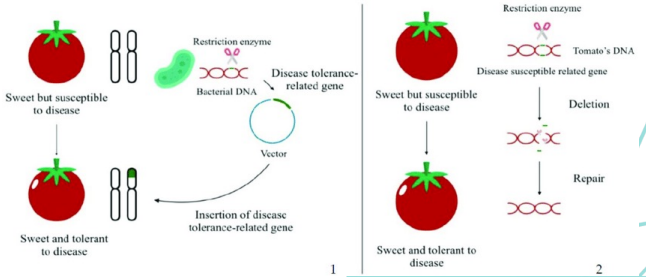


Agriculture and Food Genetic Engineering

أكاديمية كاوست
KAUST ACADEMY

- Engineer frost-resistant tomatoes
- Pesticide-resistant corn





أكاديمية وست
KAUST ACADEMY

How?

ow to Find the Origin of epli ation

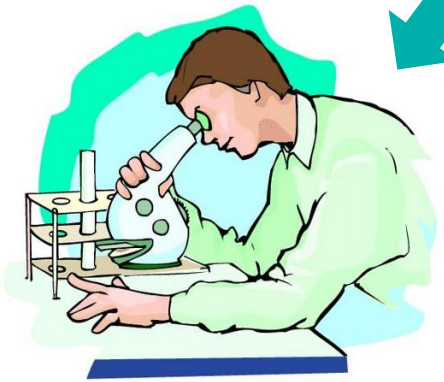
Think and share your thoughts :)

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Finding Origin of Replication

أكاديمية كاوست
KAUST ACADEMY

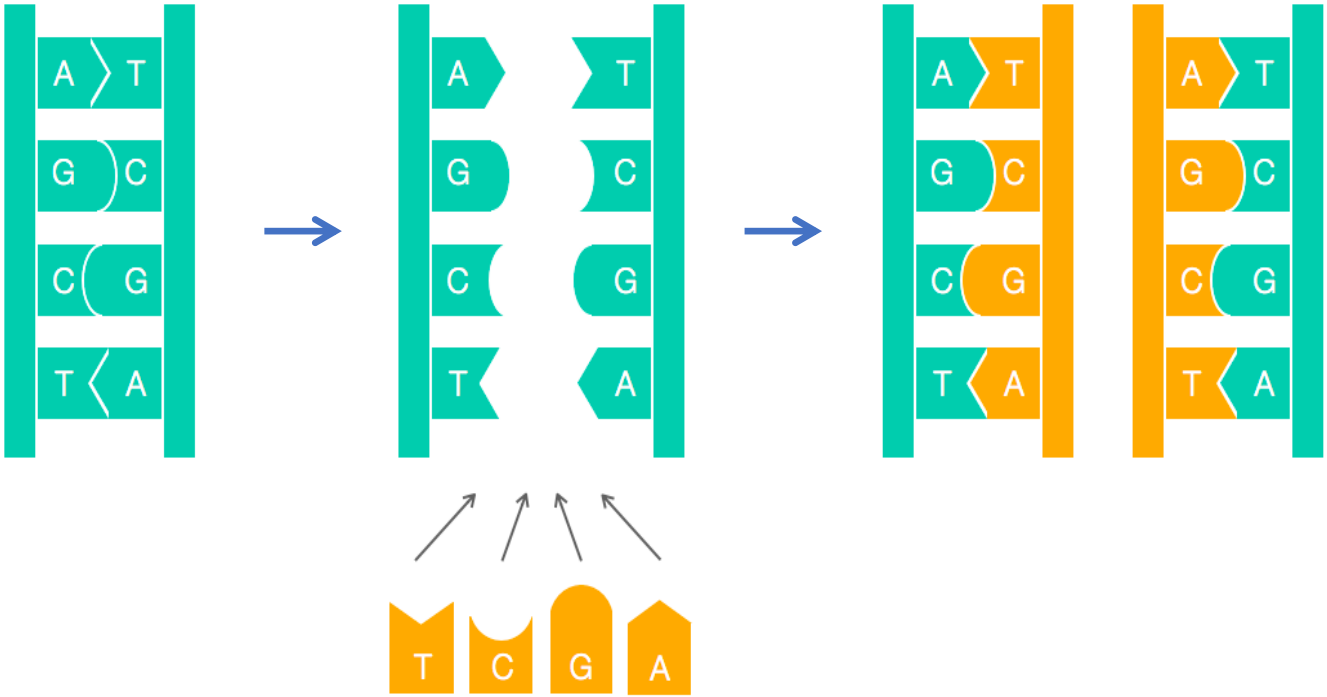
Finding OriC



Wet-Lab: OK – let's cut out this DNA fragment. Can the genome replicate without it?

Dry-Lab: ?????????

Genome Replication Problem



Finding Origin of Replication

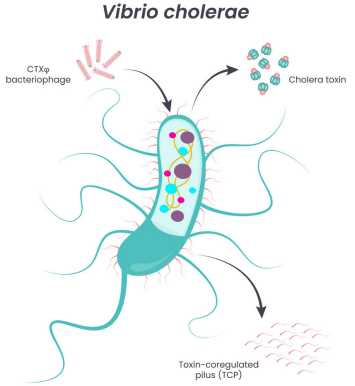
أكاديمية كاوست
KAUST ACADEMY

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

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

1. How does the cell know to begin replication in short *OriC*?

Example: Finding Origin of Replication



- genome length of Vibrio cholerae (4,033,460 base pairs)
- Replication origin of *Vibrio cholerae* (≈500 nucleotides)

atcaatgatcaacgtaagcttctaagcatgatcaaggtgctc acagttta ccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcc ctcgtactctca gacca
cggaaagatgatcaagagaggatgatttcttggccatatc caatgaatacttgt tactt
gtgcttccaattgacatcttcagcgccatattgcgctggc aaggtgacggagcg gatt
acgaaagcatgatcatggctgttgttctgtttatcttgtt tgactgagacttgt agga
tagacggtttttcatcactgactagccaaagccttactct cctgacatcgacc aaat
tgataatgaatttacatgcttccgcgacgatttacctcttg tcatcgatccg tgaag
atcttcaattgttaattctcttgcctcgactcatagccatgat agctctt catgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctc ttc

The is a H**idden message** telling the cell to start replication he

The Hidden Message Problem

أكاديمية كاوست
KAUST ACADEMY

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

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

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

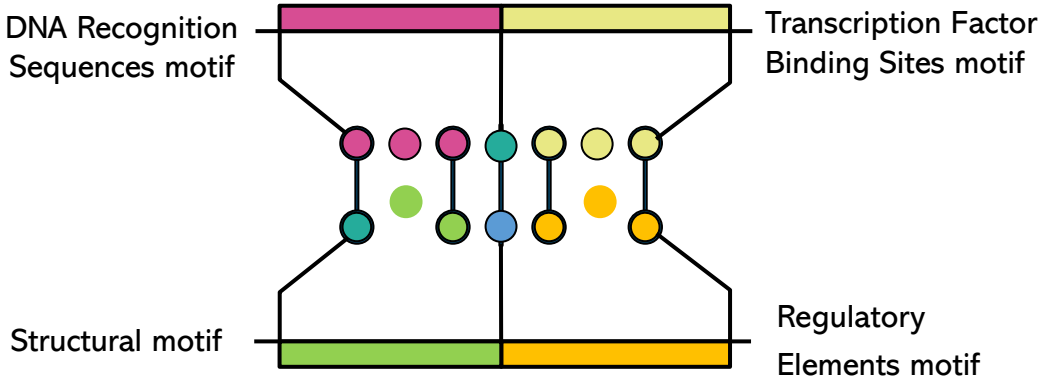
Finding Hidden Messages in DNA

DNA sequence reads

DNA Motifs: Decoding the Language of Genes

أكاديمية كاوست
KAUST ACADEMY

DNA motifs is a specific recurring sequence patterns in DNA



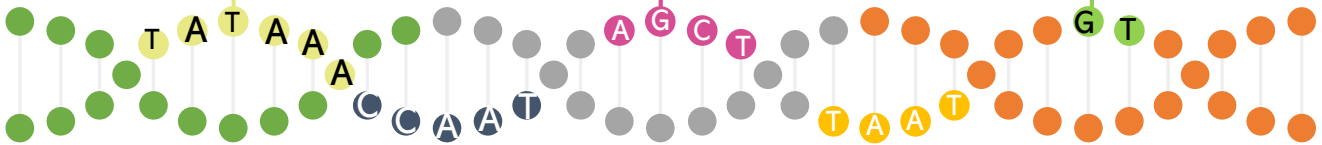
GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGACTCCTCTTC
GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGTGCATCTGAC
CCTGAGGAGAAGCACGTAGACTGG
GACTCCTCTTCGACTCCTCTTCGTGC
GACTCCTGAGGAGAAGCACGTAGA
CTGAGGACTCCTCTTCATTGCCTTT

Example of DNA Motifs

1. Transcription Factor Binding Sites (TFBS):
•**Motif:** TATA Box (TATAAA)
•**Group:** Promoter Motif
•**Function:** It's a promoter motif that helps initiate transcription by binding with transcription factors.

3. Repetitive Elements:
•**Motif:** Alu Sequence (AGCT)
•**Group:** Repetitive Motif
•**Function:** These sequences are repeated many times throughout the genome and may play a role in genetic recombination.

5. Splicing Sites:
•**Motif:** 5' Splice Site (GT)
•**Group:** Splicing Motif
•**Function:** This motif marks the beginning of an intron and is crucial for mRNA splicing.



2. Enhancer Motifs:
•**Motif:** CAAT Box (CCAAT)
•**Group:** Enhancer Motif
•**Function:** Enhancer motifs increase the rate of transcription by binding to enhancer regions and facilitating the binding of transcription factors.

4. Protein Binding Sites:
•**Motif:** Homeodomain (TAAT)
•**Group:** Protein Interaction Motif
•**Function:** This motif is recognized by specific proteins (homeodomain proteins) involved in gene regulation and development.

The Frequent Words Problem

أكاديمية كاوست
KAUST ACADEMY

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*.

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

Count(ACAACTATGCATACTATCGGGAACTATCCT, ACTAT) = 3

We will use the term k-mer to refer to a string of length k and define Count(Text, Pattern) as the number of times that a k-mer Pattern appears as a substring of Text.

Does the Frequent Words Problem Make Sense to Biologists?

أكاديمية كاوست
KAUST ACADEMY

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 ← **DNA polymerase**
- Replication initiation ← protein called ***DnaA***.

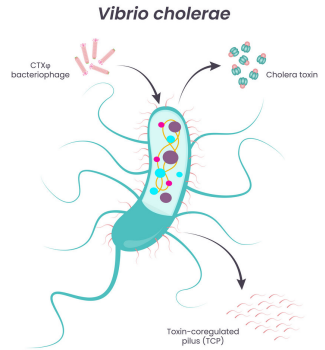
- → *DnaA* binds to short (typically 9 nucleotides long) segments within the replication origin known as a ***DnaA* box**.

- DnaA box ← 9-mer (5'-A-T-G-A-T-C-A-A-G-3')

For example : Frequent Words in *Vibrio cholerae*

أكاديمية كاوست
KAUST ACADEMY

OriC of VC



atcaatgatcaacgtaagcttctaagcATGATCAAGgtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagATGATCAAGagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaATGATCAAGctgctgctcttgatcatcgtttc

Step1: highlight a most frequent 9-mer
→ Bacterial *DnaA* boxes are usually nine nucleotides long.

- The 9-mer **ATGATCAAG** appears three times in the *ori* region of *Vibrio cholerae*

For example : Frequent Words in *Vibrio cholerae*

أكاديمية كاوست
KAUST ACADEMY

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

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

أكاديمية كاوست
KAUST ACADEMY

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaacctgagtggatgacatcaagata
ggtcgttgtatctccttcctctcgtactctcatgaccacggaaag**ATGATCAAG**agaggatgatttcttggccatatcgcaa
tgaatacttgtgacttgtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggattacgaaa
gcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttaggatagacggtttttcatcactgactagcca
aagccttactctgcctgacatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgat
ccgattgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtttccttaaccctc
tatttttacggaaga**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!

أكاديمية كاوست
KAUST ACADEMY

atcaatgatcaacgtaagcttctaagc**ATGATCAAG**gtgctcacacagtttatccacaacctgagtgga
tgacatcaagataggtcgttgtatctccttcctctcgtactctcatgaccacggaaag**ATGATCAAG**ag
aggatgatttcttggccatatcgcaatgaatacttgtgacttgtgcttccaattgacatcttcagcgcc
atattgcgctggccaaggtgacggagcgggattacgaaagcatgatcatggctgttgttctgtttatct
tgttttgactgagacttgttaggatagacggtttttcatcactgactagccaaagccttactctgcctg
acatcgaccgtaaattgataatgaatttacatgcttccgcgacgatttacct**CTTGATCAT**cgatccga
ttgaagatcttcaattgttaattctcttgcctcgactcatagccatgatgagct**CTTGATCAT**gtttcc
ttaaccctctatttttacggaaga**ATGATCAAG**ctgctgct**CTTGATCAT**cgtttc

ATGATCAAG
| | | | | | | | |
TACTAGTTC

are **reverse complements** and likely ***DnaA*** boxes
(*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.

Outlines

- Overview in Bioinformatics
- Introduction to Python
- Introduction to Genome Replication
- Genome Replication Problem
- Bioinformatics Challenges using python

Bioinformatics Challenges

GitHub

- 1A: The Hidden Message Problem
- 1B: The Frequent Words Problem
- 1C: Reverse Complement Problem
- 1D: Pattern Matching Problem

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Bioinformatics Challenges

GitHub

- **1A: The Hidden Message Problem**
- 1B: The Frequent Words Problem
- 1C: Reverse Complement Problem
- 1D: Pattern Matching Problem

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

1. The Hidden Message Problem

أكاديمية كاوست
KAUST ACADEMY

```
PatternCount(Text, Pattern)
    count ← 0
    for i ← 0 to |Text| - |Pattern|
        if Text(i, |Pattern|) = Pattern
            count ← count + 1
    return count
```

Bioinformatics Challenges

GitHub

- 1A: The Hidden Message Problem
- **1B: The Frequent Words Problem**
- 1C: Reverse Complement Problem
- 1D: Pattern Matching Problem

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

2. The Frequent Words Problem


```
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
```

Bioinformatics Challenges

GitHub

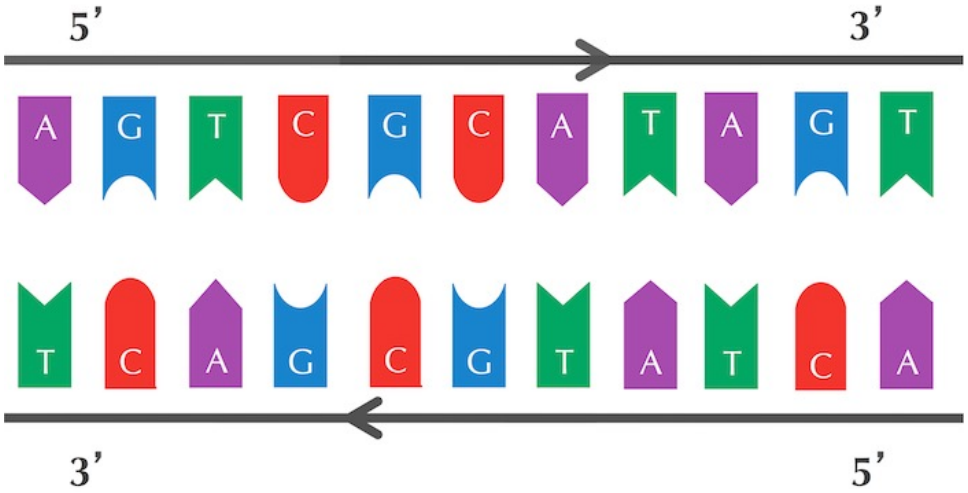
- 1A: The Hidden Message Problem
- 1B: The Frequent Words Problem
- **1C: Reverse Complement Problem**
- 1D: Pattern Matching Problem

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Complementary strand on a **template strand**

أكاديمية كاوست
KAUST ACADEMY

- Template strand: **AGTCGCATAGT**
- Complementary strand :**ACTATGCGACT**
- The beginning and end of a DNA strand are denoted 5’ and 3’, respectively.



3. Reverse Complement Problem

Reverse Complement Problem: Find the reverse complement of a DNA string.
Input: A DNA string Pattern.
Output: $Pattern_{rc}$, the reverse complement of Pattern.

Bioinformatics Challenges

GitHub

- 1A: The Hidden Message Problem
- 1B: The Frequent Words Problem
- 1C: Reverse Complement Problem
- **1D: Pattern Matching Problem**

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

3. Pattern Matching Problem

أكاديمية كاوست
KAUST ACADEMY

Pattern Matching Problem :

- Objective: Find all occurrences of a Pattern in string
- Input: A strings Pattern and Genome.
- Output: All position on genome where Pattern appears as substring.

Done with Day 1, Heyyyyy!

Thank You !

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology