



# **Virus Ancestry Detection**



A divide and conquer approach

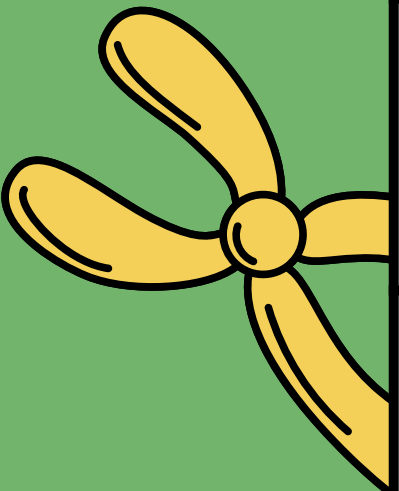
# Team Members

Hamsa Saber 1210359

Alia Tarek 4220121

Salsabil Mostafa 1210171

Youssef Affify 1200883



A decorative border featuring a stylized DNA double helix. The helix is composed of orange and green ribbons forming the sugar-phosphate backbone, with yellow and blue rectangular blocks representing the nitrogenous base pairs. The helix is shown in a 3D perspective, winding around the edges of the slide.

# **PROBLEM STATEMENT**

## **Challenges:**

- Genome sequences are long and complex, making manual analysis impossible.
- Comparing multiple sequences is computationally expensive

## **Need for Efficiency:**

- Bioinformatics requires optimized algorithms for accuracy and speed.
- Divide-and-conquer algorithms simplify the comparison process.





## **PROJECT OVERVIEW**

### **Goal:**

Identify the family of an unknown virus using DNA sequencing.

- Compare the unknown sequence with a database of known virus genomes.
- Determine the closest match and identify its family.

### **Relevance:**

- Helps track virus evolution and origins.
- Supports research and response to pandemics.

# Workflow Overview

## Input:

- An unknown DNA sequence from a .fasta file.
- A database of known DNA sequences from viruses.

## Processing Steps:

1. Use a divide and conquer algorithm to break sequences into smaller chunks.
2. Compare each chunk of the unknown DNA to the database sequences.
3. Assign scores based on matches, mismatches, and gaps.

## Output:

- Identify the closest match and its virus family.



## **DIVIDE AND CONQUER ALGORITHM**

### **Divide:**

- Split both the unknown DNA and database sequences into smaller chunks.
- Example: Split into halves recursively until the length reaches a minimum threshold.



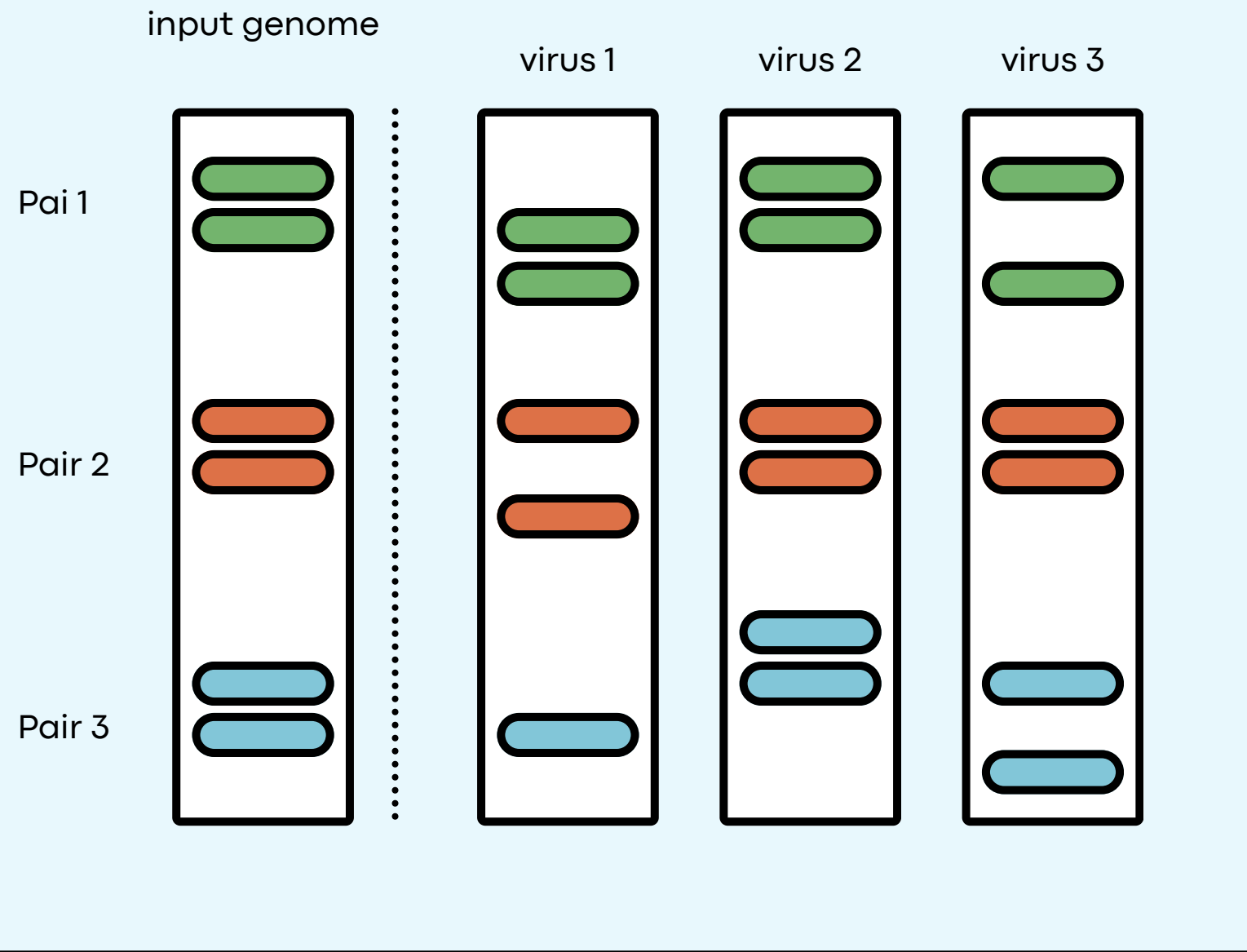
## DIVIDE AND CONQUER ALGORITHM

### Conquer

- Compare chunks using the scoring system:
- Match: +2 points.
- Mismatch: -2 points.
- Gap penalties: Open gap = -5
- Extend gap = -1.

### Combine

- Merge the results of all chunk comparisons.
- Sum scores to get the overall similarity score for each sequence.



# PSUEDOCODE

## 1. Loading DNA Sequences:

- Start by reading DNA sequences from all `.fasta` files in the database folder.
- If a file matches the name of the unknown virus, treat it as the *target sequence*.
- For all other sequences, store them in a database for comparison.
- Skip very short sequences since they might not be useful for analysis.

## 2. Aligning Two Sequences:

- Use a scoring system to align two DNA sequences:
  - Add 2 points for every matching base.
  - Subtract 2 points for mismatches.
  - Apply penalties for gaps: -5 to open a gap and -1 for extending it.
- Normalize the score by dividing it by the length of the alignment to account for different sequence sizes.
- Return the score along with the aligned versions of the sequences.



## 3. Divide and Conquer:

- To handle large sequences efficiently:
  - Split both sequences into two halves.
  - Align the left halves and right halves separately using the same process.
  - Combine the results to get the total score and aligned sequences.
- If the sequences are short enough (less than 1800 bases), skip the splitting and align them directly.

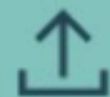
## 4. Finding the Closest Match:

- Compare the target sequence to every sequence in the database:
  - Align them using the divide-and-conquer approach.
  - Calculate scores for all matches.
- Calculate the average score of all alignments and set a *threshold* based on it.
- Identify the best match with the highest score above the threshold.

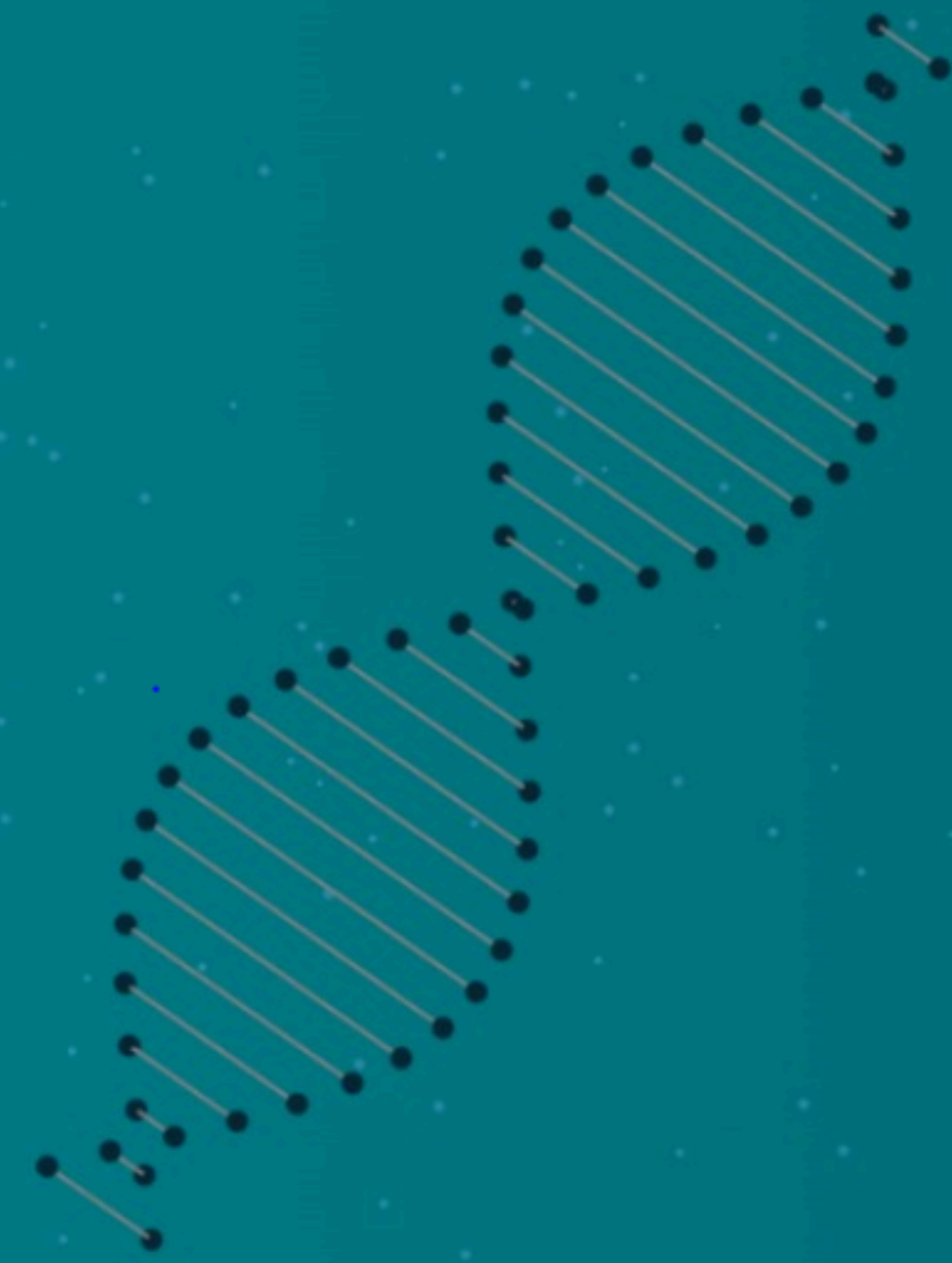
## Virus Ancestry Detection

### Ancestry Virus Identification

Discover the origins of viral DNA with our advanced identification tool. Upload your DNA sequence in FASTA format to analyze its lineage and explore its ancestral connections.

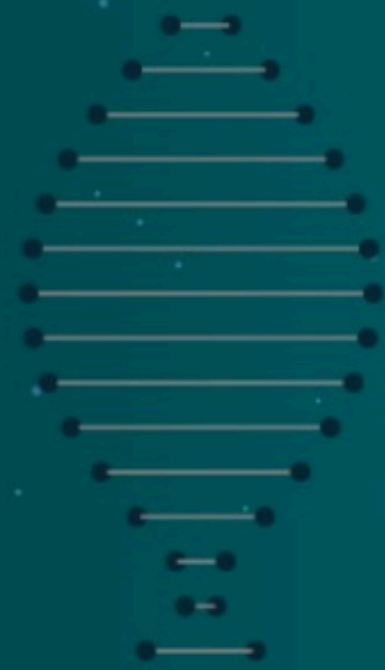


Upload Virus DNA



# UI SNAPSHOTS

Best Match  
SARS

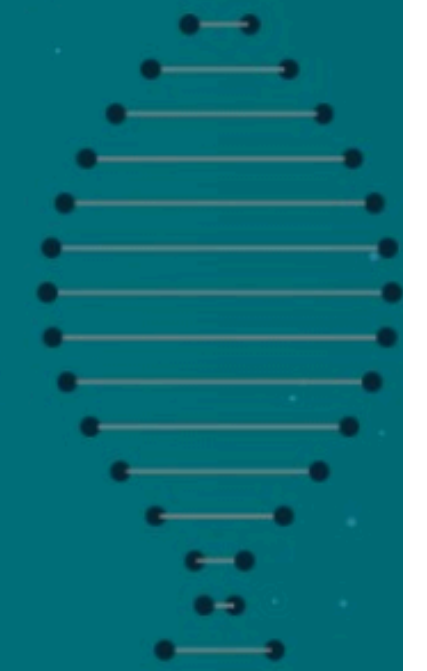


Ancestors Detected



[Download Ancestry Report](#)

Uploaded Virus



Activate

## COMPLEXITY

### Aligning Sequences:

$O(n \times m)$

n: length of small chunk of sequence 1

m: length of small chunk of sequence 2

### Dividing the sequences:

$O(\log(L))$


L: length of longer seq

### Final complexity

$O(n \times m \times \log(L))$


- The total work done across all levels of recursion is proportional to  $n \times m$  at each level.
- There are  $O(\log(L))$  levels in the recursion, since the sequences are halved at each step.

# CONCLUSION



This project showcases how efficient algorithms like divide-and-conquer can accurately identify virus families from genome sequences, aiding in research and pandemic response. It highlights the power of bioinformatics in transforming genetic data into actionable insights.





THANK YOU!!  
Any Questions?