# BJYZ

Bowen Yao, Jingwu Wang,
Yuqi Chen, Ken Chen

# Team Member And Role

Bowen Yao:  Project Manager

Jingwu Wang: Documentation Lead

Yuqi Chen: Testing Lead

Zikang Chen: Orchestration/IO Lead

# Content

- Algorithm
- Testing
- Result and Metrics
- Future Goals

# Good to know :)

- Python Version
    - Python 3.6 or later version
- Pip
    - Default feature in python environment
- Input Library
    - biopython Library: Fasta/ Fastq file reader

# Algorithm

# Some Algorithms We Tried

Reference Genome Length - n

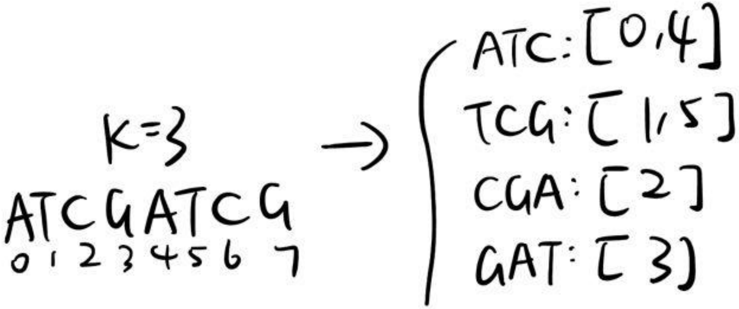
Read Genome Length - m

Seed Length - k (kmer)

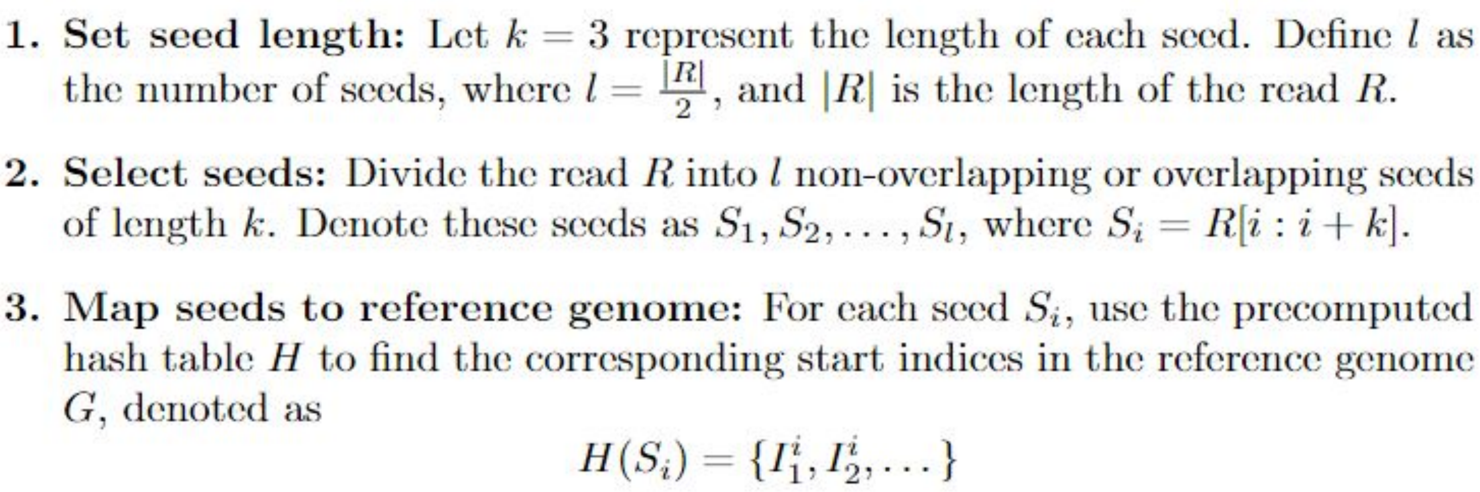
- Suffix Array
  - Time complexity is O(m log n)
  - Takes several minutes for 1000 reads
  - Memory Intensive - Space O(n)
- Smith-Waterman Algorithm
  - Gives the optimal approximate matching
  - Low Efficiency - Time Complexity O(nm)
  - Takes > 1 hour for 1000 reads
- Hash table + Seeds-and-Extends + Smith-Waterman

# Reference Genome

1. Decide a k (we choose 30)
2. Find all k-mers in the reference genome and store them along with their starting indices to a hash table



# Split Reads - retrieve potential candidate



1. **Set seed length:** Let $k = 3$ represent the length of each seed. Define $l$ as the number of seeds, where $l = \frac{|R|}{2}$, and $|R|$ is the length of the read $R$.

2. **Select seeds:** Divide the read $R$ into $l$ non-overlapping or overlapping seeds of length $k$. Denote these seeds as $S_1, S_2, \ldots, S_l$, where $S_i = R[i : i + k]$.

3. **Map seeds to reference genome:** For each seed $S_i$, use the precomputed hash table $H$ to find the corresponding start indices in the reference genome $G$, denoted as

$$H(S_i) = \{I_1^i, I_2^i, \ldots\}$$

)

seed length=3, select seeds for every 2 index

ATCG ATCG →[(ATC,0),(CGA,2),(ATC,4)]

0 1 2 3 4 5 6 7

ATC: [0,4]

TCG: [1,5]

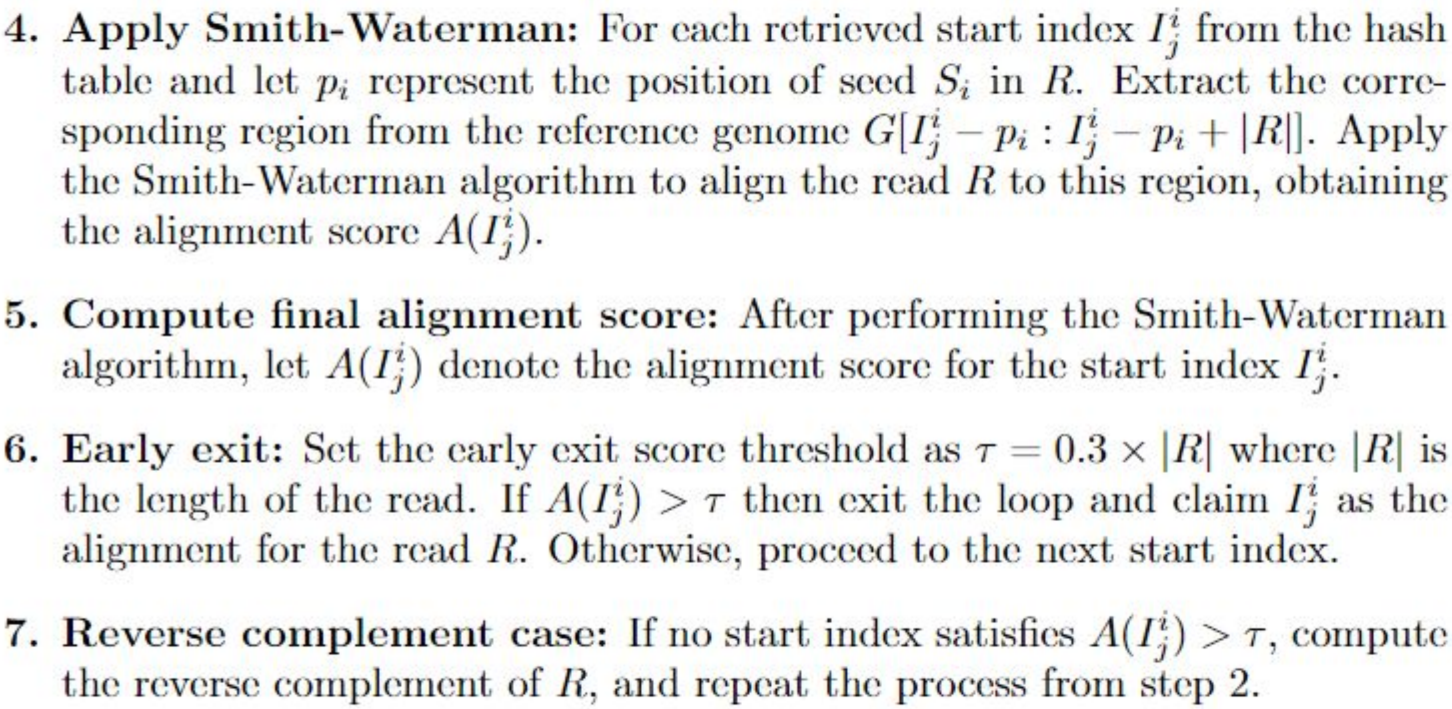
CGA: [2]

GAT: [3]

→ [([0,4],0),([2],2),([0,4],4)]

→[[0,4],[0],[0]]

# Split Reads - apply SW algorithm



4. **Apply Smith-Waterman:** For each retrieved start index $I_j^i$ from the hash table and let $p_i$ represent the position of seed $S_i$ in $R$. Extract the corresponding region from the reference genome $G[I_j^i - p_i : I_j^i - p_i + |R|]$. Apply the Smith-Waterman algorithm to align the read $R$ to this region, obtaining the alignment score $A(I_j^i)$.

5. **Compute final alignment score:** After performing the Smith-Waterman algorithm, let $A(I_j^i)$ denote the alignment score for the start index $I_j^i$.

6. **Early exit:** Set the early exit score threshold as $\tau = 0.3 \times |R|$ where $|R|$ is the length of the read. If $A(I_j^i) > \tau$ then exit the loop and claim $I_j^i$ as the alignment for the read $R$. Otherwise, proceed to the next start index.

7. **Reverse complement case:** If no start index satisfies $A(I_j^i) > \tau$, compute the reverse complement of $R$, and repeat the process from step 2.

# Parameters

- k: Seed Length [significantly impact the speed]
    - The larger k, the faster
    - The larger k, the lower possibility to find a match
- Seed_num: The number of seeds for each read
    - The larger, the greater possibility of finding a match
    - The larger, the more comparison that slows speed
- DP threshold: Matching similarity

# Testing

# Testing on I/O

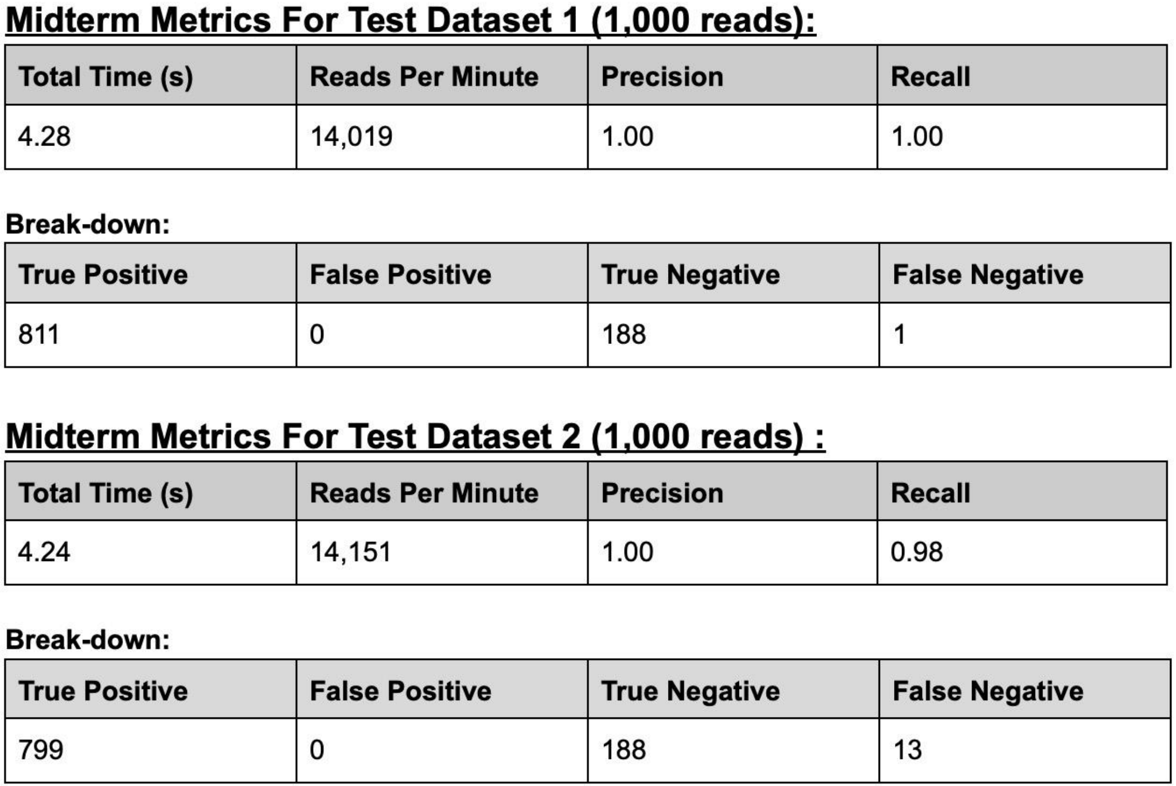
- ReadFasta and ReadFastq
    - Test on empty file
    - Test incorrect input path
    - Test on malformed file
    - Test on a simple correct file
- SAMWriter
    - Validate the SAM file output
    - Check correct header
    - Check correct mapping

# Testing on Sequence Mapping

- Comparison of Expected Start Position
    - For each read in the dataset, compare the match index returned by the mapping algorithm with the corresponding expected start position from the ground truth data
- Performance Metrics Calculation:
    - Precision
    - Recall
    - Runtime

# Result

# Result



**Midterm Metrics For Test Dataset 1 (1,000 reads):**

| Total Time (s) | Reads Per Minute | Precision | Recall |
|---|---|---|---|
| 4.28 | 14,019 | 1.00 | 1.00 |

**Break-down:**

| True Positive | False Positive | True Negative | False Negative |
|---|---|---|---|
| 811 | 0 | 188 | 1 |

**Midterm Metrics For Test Dataset 2 (1,000 reads) :**

| Total Time (s) | Reads Per Minute | Precision | Recall |
|---|---|---|---|
| 4.24 | 14,151 | 1.00 | 0.98 |

**Break-down:**

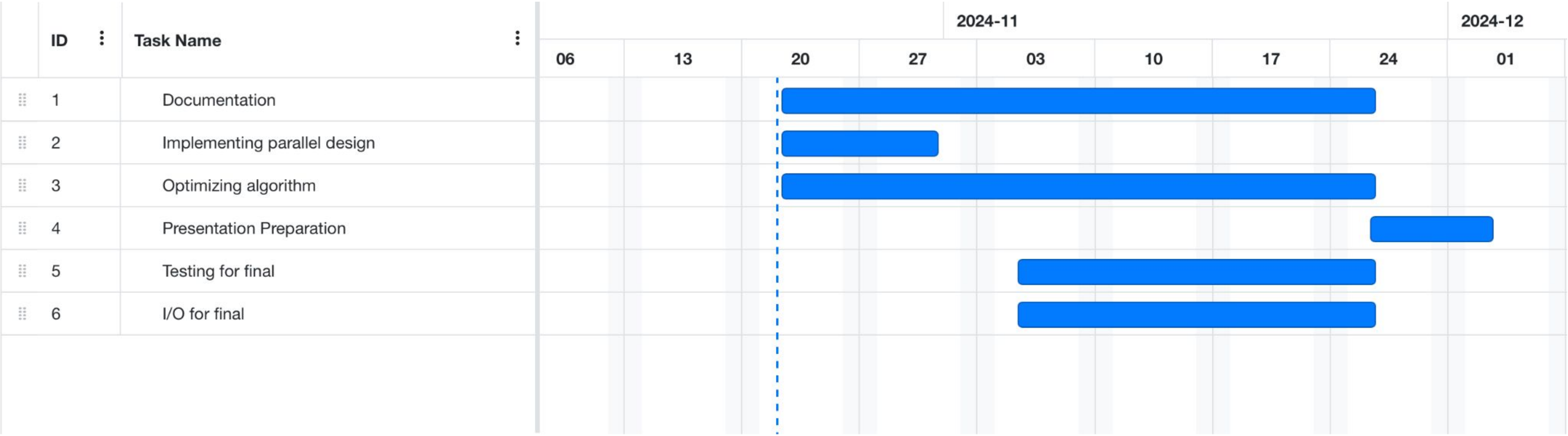
| True Positive | False Positive | True Negative | False Negative |
|---|---|---|---|
| 799 | 0 | 188 | 13 |

# Future Goals

# Next Steps

- Parameter Tuning
  - Develop and refine methods to determine the optimal number of seeds and the ideal k-value for improved alignment accuracy.
- Implement Parallel Design
  - Leverage parallel processing techniques to accelerate the algorithm, efficiently handle larger datasets, and reduce runtime.
- Optimize Algorithm Performance
  - Explore optimization strategies to enhance throughput and overall algorithm efficiency.

# Gantt Chart



| ID | Task Name |
|---|---|
| 1 | Documentation |
| 2 | Implementing parallel design |
| 3 | Optimizing algorithm |
| 4 | Presentation Preparation |
| 5 | Testing for final |
| 6 | I/O for final |