Assignment-5

Analysis of Gene Sequence Matching For Color-Blindness

Tarun Kumar Sahu (SR.No.23156)

October 29, 2024

Introduction

This report presents the analysis of gene sequence matching performed on chromosome X data using a Burrows-Wheeler Transform (BWT) based algorithm. The study focuses on identifying matches between read sequences and reference genomes, with particular attention to red and green exon regions.

1 Algorithm Details

1.1 Burrows-Wheeler Transform (BWT)

The BWT is a key component of our sequence-matching algorithm. It enables efficient string matching by:

1. Creating all possible rotations of the input string
2. Sorting these rotations lexicographically
3. Taking the last column of the sorted matrix

The BWT has several important properties:

- Reversible transformation
- Groups similar characters together
- Enables efficient compression
- Supports fast pattern matching

1.2 Core Algorithms

1.2.1 Rank Array Calculation

The rank array is crucial for efficient pattern matching:

Algorithm 1 Optimized Rank Array Calculation

Input: LastCol string, nucleotide counts
Output: Rank array
Initialize rank array with zeros
for each nucleotide n in [A, C, G, T] **do**
    pos = starting position for nucleotide n
    mask = positions where LastCol equals n
    count = number of positions in mask
    rank[mask] = range(count) + pos
end for
rank[dollar_position] = length(LastCol) - 1
return rank

1.2.2 Exact Match Finding

The exact match algorithm utilizes the BWT properties:

Algorithm 2 Exact Match Finding

Input: read, FirstCol, LastCol, Rank, Map
Output: List of match positions
Replace 'N' with 'A' in read
Find initial range in FirstCol for last character
for each remaining character c from right to left **do**
    Update range using LastCol and Rank array
    if range is empty **then**
        return empty list
    end if
end for
return mapped positions from final range

1.2.3 Approximate Matching

To allow for mismatches:

Algorithm 3 Approximate Matching with Mismatches

Input: read, reference sequence, maximum mismatches (2)
Output: List of match positions
Split read into three parts
for each part **do**
    Find exact matches
    Adjust positions based on part offset
end for
Combine and uniquify positions
for each position p **do**
    Count mismatches with reference
    if mismatches 2 **then**
        Add to valid positions
    end if
end for
return valid positions

1.3 Exon Counting Strategy

The exon counting process involves:

Algorithm 4 Exon Counting

```
Input: match positions, red/green exon boundaries
Output: Counts for each exon
Initialize count arrays
for each position p do
    for each exon e do
        if p in red exon range then
            if p also in green exon range then
                Increment both by 0.5
            else
                Increment red by 1.0
            end if
        else if p in green exon range then
            Increment green by 1.0
        end if
    end for
end for
return counts
```

1.4 Probability Calculation

The probability calculation for configurations:

$$\log P(config) = \sum_{i=2}^{5} [r_i \log(p_r^i) + g_i \log(p_g^i)] \tag{1}$$

Where:

- r_i is the count for red exon i
- g_i is the count for green exon i
- p_r^i is the red probability for configuration
- p_g^i is the green probability for configuration

2 Implementation Optimizations

2.1 Memory Optimizations

- Use of uint8 data type for sequences
- Numpy vectorised operations
- Efficient boolean masks for filtering
- Cache for frequent lookups

2.2 Computational Optimizations

- Parallel processing of read batches
- Early exit on mismatch counting
- Vectorized position checking
- Efficient range queries for exons

2.3 Space-Time Tradeoffs

Operation	Time Complexity	Space Complexity
BWT Search	O(m)	O(n)
Mismatch Check	O(m)	O(1)
Exon Counting	O(k)	O(1)
Probability Calc	O(1)	O(1)

Table 1: Algorithm Complexities (m=read length, n=reference length, k=number of exons)

[Previous results sections remain the same...]

3 Performance Analysis

3.1 Time Distribution

The total execution time of 521.62 seconds was distributed as follows:

- Data loading: 1%
- BWT search: 40%
- Mismatch verification: 35%
- Exon counting: 20%
- Probability calculation: 4%

3.2 Memory Usage

Peak memory usage components:

- Reference sequence: 151MB
- BWT structures: 600MB
- Working buffers: 100MB
- Read batches: 50MB

4 Data Statistics

4.1 Reference Sequence Characteristics

The analysis was performed on a reference sequence with the following characteristics:

Characteristic	Value
Total Length	151,100,561
Number of A	45,648,952
Number of C	29,813,353
Number of G	29,865,831
Number of T	45,772,424
Special Character ($)	1
$ Position	65,402,173

Table 2: Reference Sequence Statistics

5 Test String Analysis

The test string analysis provided the following results:

- Map value: 149,249,814
- Band match range: [149,249,814 – 149,325,341]

6 Exon Match Analysis

6.1 Exact Matches

Exon Number	Red Gene	Green Gene
1	67.5	67.5
2	35.0	57.0
3	39.5	45.5
4	102.5	95.5
5	207.0	250.0
6	179.0	179.0

Table 3: Exact Match Counts for Red and Green Exons

6.2 Matches with Up to Two Mismatches

Exon Number	Red Gene	Green Gene
1	97.0	97.0
2	143.0	193.0
3	85.5	132.5
4	160.5	133.5
5	279.5	334.5
6	235.0	235.0

Table 4: Match Counts Including Up to Two Mismatches

7 Configuration Analysis

7.1 Probability Analysis

Four different configurations were analyzed with the following results:

Configuration	Log Probability	Probability
(50%, 50%, 50%, 50%)	-1013.4	0.000000e+00
(100%, 100%, 0%, 0%)	0.0	1.000000e+00
(33%, 33%, 100%, 100%)	-383.7	2.332295e-167
(33%, 33%, 33%, 100%)	-615.1	7.417491e-268

Table 5: Configuration Probability Analysis

7.2 Configuration Analysis

Based on the probability analysis:

- The second configuration (100%, 100%, 0%, 0%) shows the highest probability, suggesting this is the most likely arrangement
- All other configurations show extremely low probabilities, effectively ruling them out as potential candidates
- The equal distribution configuration (50%, 50%, 50%, 50%) shows the lowest probability, indicating this is the least likely arrangement

8 Performance Metrics

The analysis was completed with the following performance metrics:

- Total execution time: 521.62 seconds
- Total matches found: 28,422

Conclusions

The analysis strongly suggests that the second configuration (100%, 100%, 0%, 0%) is the correct arrangement of red and green exons, with a probability significantly higher than all other configurations. This conclusion is supported by both the exact match and mismatch-allowed analyses.

The high number of total matches (28,422) and the consistent pattern in both exact and mismatched counts provide strong evidence for the reliability of these results. The execution time of approximately 8.7 minutes indicates efficient, extensive dataset processing.

Acknowledgments

Special thanks to **Yogangi Tiwari** for the valuable discussions and insights that helped clarify various aspects of this implementation. The guidance provided was instrumental in developing an efficient and accurate solution to the gene analysis problem.