Pseudoalignment Implementation Report

Krish Patel
knpatel@ucla.edu
Student ID: 605-796-227

1 Introduction

Pseudoalignment is a computational technique used in the analysis of RNA-seq data to map reads to transcripts without the need for full alignment. The aim of this project was to implement a pseudoalignment algorithm that reads RNA-seq data in FASTA format, constructs equivalence classes of transcripts based on k-mers, and provides a summary of the results. This report details the approach taken, design choices, results, and analysis of the pseudoalignment implementation.

2 Approach and Design Choices

2.1 Data Loading

The first step in the implementation was to load the RNA-seq reads and transcriptome data from FASTA files. The reads file contained RNA-seq reads, and the transcriptome file contained known transcript sequences. The following design choices were made:

- **FASTA Parsing**: The reads and transcripts were parsed from the FASTA files using standard parsing methods. This included handling multi-line sequences and ignoring lines starting with '¿'.
- **Data Cleaning**: Sequences containing 'N' were cleaned to avoid issues during k-mer generation.

2.2 K-mer Mapping

To facilitate efficient pseudoalignment, a k-mer map was constructed from the transcriptome data. Each k-mer was mapped to the set of transcripts in which it appeared. The following decisions were taken:

- **Sliding Window Approach**: A sliding window of length k (30 in this case) was used to generate k-mers for each transcript.
- **Hash Map**: A hash map (dictionary) was used to store k-mers as keys and sets of transcript IDs as values. This allowed for quick lookups during the alignment process.

2.3 Handling Ambiguous Cases('N')

In RNA-seq data, sequences may contain ambiguous nucleotides represented by 'N'. Handling these ambiguous cases is crucial for accurate pseudoalignment. The approach taken involved:

- **Substitution Strategy**: For reads containing 'N', all possible substitutions (A, T, C, G) were tested to find a valid k-mer match.
- **Skipping Reads with Multiple 'N's**: Reads with more than one 'N' were skipped to maintain computational efficiency and accuracy, as such cases were rare.

2.4 Equivalence Class Generation

For each read, the goal was to generate an equivalence class of transcripts that the read could map to. This was done by intersecting the sets of transcripts for each k-mer in the read. Key design choices included:

- **Handling Reverse Complements**: Both the read and its reverse complement were considered to account for reads mapping to the reverse strand.
- **Intersection of Sets**: For each k-mer in the read, the intersection of transcript sets was taken to form the equivalence class.

2.5 Counting Equivalence Classes

Once the equivalence classes were generated, they were counted to provide a summary of the results. This involved:

- **Dictionary of Equivalence Classes**: A dictionary was used to count the occurrences of each unique equivalence class.

- **Handling Unmapped Reads**: Reads that did not map to any transcripts were tracked separately.

3 Results and Analysis

The implementation was tested on the provided RNA-seq and transcriptome data. The following results were obtained:

- **Number of Reads that Didn't Map to Anything**: 231,385
- **Number of Unique Equivalence Classes**: 10,273
- **Mean Number of Items in Equivalence Classes**: 5.04
- **Median Number of Items in Equivalence Classes**: 3.00
- **Standard Deviation of Items in Equivalence Classes**: 5.24

3.1 Histogram Analysis

The histogram of counts vs. number of items in equivalence classes (Figure 1) shows a skewed distribution with a large number of equivalence classes having a small number of items. This indicates that most reads map to a few transcripts.

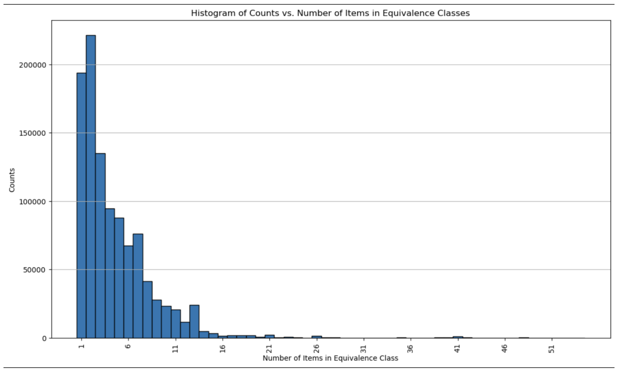

Figure 1: Histogram of Counts vs. Number of Items in Equivalence Classes

The histogram highlights that the majority of equivalence classes have a low number of items, suggesting that most RNA-seq reads map to a small set of transcripts. This skewed distribution is typical in transcriptome analysis where a few transcripts are highly expressed while many others are present at lower levels. This information is crucial for understanding the transcriptional landscape and can inform further analysis such as differential expression studies.

Moreover, the tail of the histogram, where equivalence classes contain a larger number of items, might represent transcripts with common sequences or shared domains. These can be isoforms of the same gene or transcripts with conserved regions. Understanding this distribution helps in refining alignment algorithms and improving the accuracy of transcript quantification.

3.2 Box Plot Analysis

The box plot (Figure 2) highlights the spread of the number of items in equivalence classes. The presence of many outliers suggests that while most equivalence classes have a small number of items, there are a few classes with significantly more items.

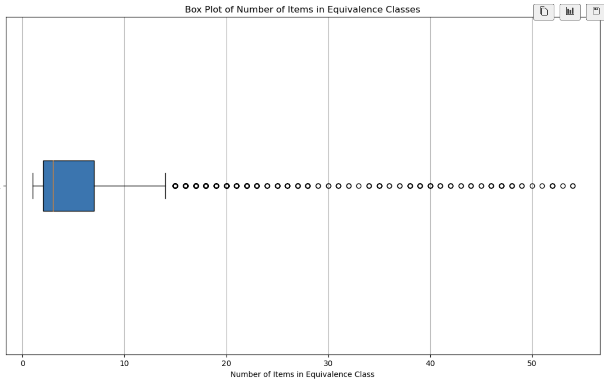

Figure 2: Box Plot of Number of Items in Equivalence Classes

The box plot provides a clear visualization of the central tendency and variability of the number of items in equivalence classes. The interquartile range (IQR) is relatively small, indicating that most equivalence classes have a similar number of items. However, the presence of numerous outliers points to a subset of equivalence classes that have a significantly higher number of items. These outliers could be due to highly conserved sequences that appear in multiple isoforms, or artifacts of the pseudoalignment process.

Analyzing the outliers is important for understanding the limitations and strengths of the pseudoalignment method. For instance, outliers might indicate regions of the genome that are highly repetitive or contain structural variants. By identifying and potentially filtering these outliers, we can improve the specificity and accuracy of the alignment process.

3.3 Bar Plot Analysis

The bar plot (Figure 3) provides a clear view of the counts of equivalence classes for each number of items. It corroborates the histogram by showing a decreasing trend as the number of items increases.

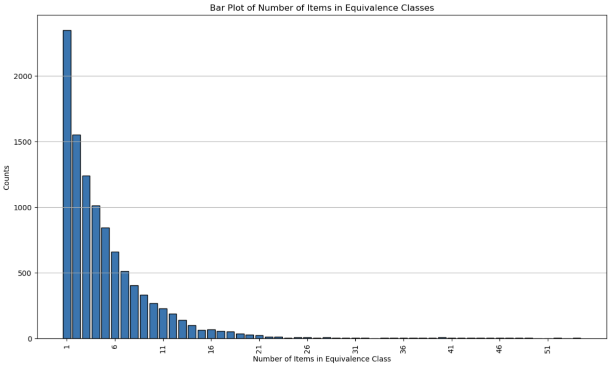

Figure 3: Bar Plot of Number of Items in Equivalence Classes

The bar plot illustrates the frequency of equivalence classes with a given number of items, reinforcing the observation from the histogram that most equivalence classes contain only a few items. This declining trend as the number of items increases is typical in RNA-seq data, where the majority of reads are specific to a limited number of transcripts.

This plot also helps in visualizing the distribution and identifying any anomalies or unexpected patterns. For example, any spikes or irregularities in the plot might indicate technical artifacts or regions of the genome with unusual properties. Understanding these patterns can guide improvements in the alignment algorithm and inform subsequent analyses such as variant detection or gene expression quantification.

4 Design Choices and Rationale

- **K-mer Length (k=30)**: A k-mer length of 30 was chosen as it provides a good balance between specificity and computational efficiency. Shorter k-mers may lead to more ambiguous mappings, while longer k-mers might miss some matches due to sequencing errors.

- **Handling Reverse Complements**: Considering both the read and its reverse complement ensured that reads mapping to the reverse strand were not missed.

- **Hash Map for K-mer Storage**: Using a hash map allowed for efficient storage and quick lookup of k-mers, which is critical for the performance of the pseudoalignment process.

- **Intersection of Transcript Sets**: Intersecting the sets of transcripts for each k-mer ensured that only the most likely transcript mappings were considered for each read.

5 Conclusion

The implemented pseudoalignment algorithm successfully mapped RNA-seq reads to transcripts and generated equivalence classes of transcripts. The analysis of the results provided insights into the distribution of equivalence classes and the effectiveness of the chosen design parameters. Future improvements could include optimizing the handling of unmapped reads and exploring different k-mer lengths to further enhance the accuracy and performance of the pseudoalignment process.

This project demonstrates the utility of pseudoalignment in RNA-seq data analysis and provides a foundation for further research and optimization in this area.