Comparison of Performance Across Suffix Tries, Trees, and Arrays

Jason Hunter

March 6, 2025

1 Introduction

2 Results

2.1 Runtime Comparison

3 Methods

3.1

3.2

3.3

3.4

4 Conclusion

5 Dependencies

- numpy
- subprocess
- psutil
- pandas
- seaborn
- matplotlib
- tqdm
- argparse
- gzip
- os
- glob
- GraphicViz

5.1 Installing GraphicViz

To install GraphicViz, the package used to generate the trees and tries in this document, run the following command:

```
$ pip3 install git+https://github.com/cjdrake/ipython-magic.git
```

6 Reproducibility

6.1 Commands to collect this data and reproduce the results

I ran these commands to generate and collect the data for the figures in this report.
NOTE: You may need to install some dependencies specified in the README.md file and section 4 of this document.
ALSO NOTE: These commands are shell scripts that can take up to 30 minutes to run, depending on how fast your CPU is.

```
$ git clone https://github.com/cu-compg-spring-2025/assignment-5-k-mer-index-JasonHunter95.git
$ cd assignment-5-k-mer-index-JasonHunter95
$ for k in 12 15 18 21 24; do
    for error_rate in 0.01 0.033 0.066 0.1; do
        python3 src/kmer_idx.py \
          --kmer $k \
          --reference data/chr22.fa.gz \
          --read_size 100 \
          --num_reads 50 \
          --error_rate $error_rate \
          --max_mismatches 3 \
          --gap -2 --miss -1 --match 1 \
          --show_progress
    done
done

$ for k in 12 15 18 21 24; do
    for error_rate in 0.01 0.033 0.066 0.1; do
        python3 src/kmer_idx.py \
          --kmer $k \
          --reference data/chr22.fa.gz \
          --read_size 100 \
          --num_reads 50 \
          --error_rate $error_rate \
          --max_mismatches 3 \
          --gap -2 --miss -1 --match 1 \
          --use_extension \
          --show_progress
    done
done
```

To reproduce these, clone the repository, ensure that you are in the root directory and run the above shell commands in the repository root.