**Homework #5**

**Name**: Purnabhishek Sripathi  
**Email**: [ps747@nau.edu](mailto:ps747@nau.edu)  
**User id**: 6274051

**Problem #1 (of 1): Prefix Trie**

Create a class called ***Prefix\_Trie.***  The purpose of the class will be to contain a dataset of genomic sequences (queries) and all of the functions needed to operate on this set.  Use the **prefix trie** data structure to store the genomic fragments of a given size.  Here you will be performing fuzzy matching, tolerating up to 1 mismatch.

At a minimum, the class must contain(1**5pts**):

**A default constructor**

**A screen shot of a computer

Description automatically generated**

**At least one custom constructor to build a trie from a set of queries (of size n)**

**A screenshot of a computer

Description automatically generated**

**A function to traverse (*search*) the trie using a genome of size G. Note that you can assume that G >> n.  You will need to implement a *fuzzy search tolerating up to 1 mismatch*(substitutions only).  Hint: use a stack to keep track of branches in the tree that need to be explored.**

A screenshot of a computer

Description automatically generated

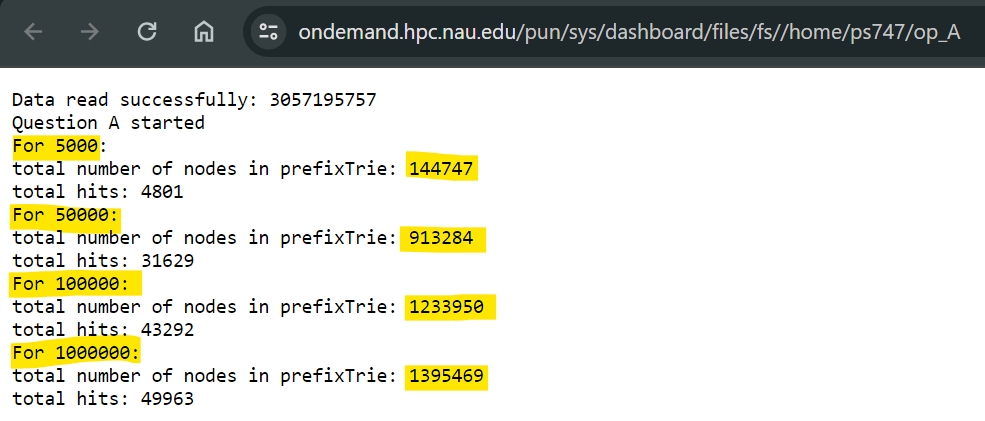
**A destructor**

A screenshot of a computer

Description automatically generated

**Part A. (20pts) Basic prefix trie**

**For each of the 36-mer datasets, what are the sizes of the trie (# of nodes)? Explain the pattern that you observed.**

****

**Iterate through all possible 36-mers in the segment, using each to search/traverse the prefix trie with up to 1 mismatch. How many of your 36-mers had a match?  Does it make sense?  Explain why.**

**A screenshot of a computer

Description automatically generated**

.

**Part B (20pts) Impact of error rate on trie structure:**

**For each of the 36-mer datasets, what are the sizes of the trie (# of nodes)? Explain differences (if any) between the trie sizes in part A and part B.**

**A screenshot of a computer

Description automatically generated**

**Iterate through all possible 36-mers in the segment, using each to search/traverse the prefix trie with up to 1 mismatch. How many of your 36-mers had a match?  Does it make sense?  Explain why.**

**A screenshot of a computer

Description automatically generated**.

**Part C (20pts) Full prefix trie experience:**

**How long did it take you to find all 32-mers of 100K, 1M, and 100M character segments within the prefix trie? Estimate how long it would take to search the entire human genome.**

**A screenshot of a computer

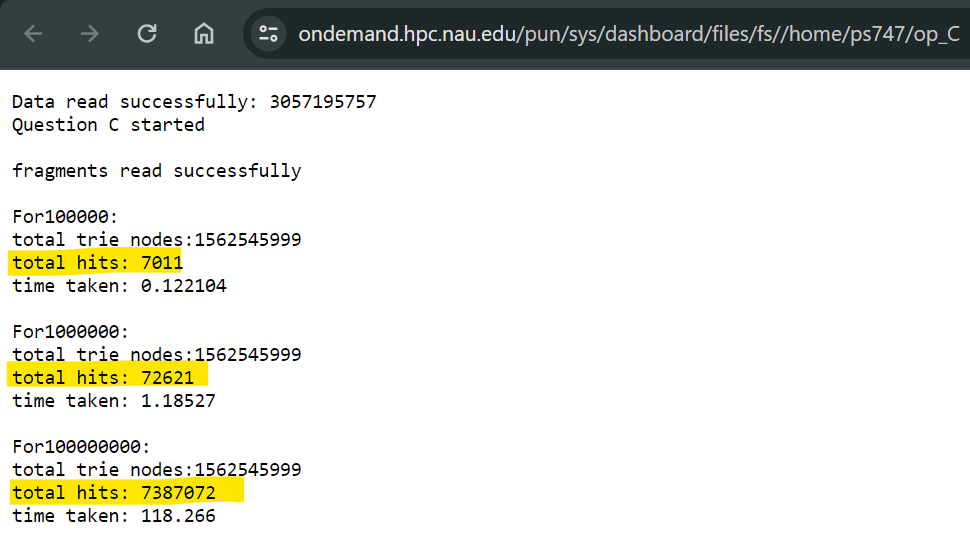
Description automatically generated**

**Estimated time for entire genome =** 30.57195757×118.266

**Estimated time for the entire genome =** 3613.303 seconds.

Therefore, to search the full human genome, it would take around 1 hour and 0.22 minutes, or about 1 hour and 13 minutes.

**How many ‘hits’ did you find for the 100K, 1M, and 100M segments? Estimate how many you would find in the full genome.**

****

Therefore, the estimated number of hits for the entire human genome would be around 214474 to 22556071 hits, depending on the assumed hits per fragment ratio**.**