ITU Computer Engineering Department
BLG 223E Data Structures, Spring 2023
Homework #2

1 Problem Definition

Two files will be used **(predict.txt, gt.txt)** in this homework. These files contain three columns for CHROM (chromosome), POS (position), and ALT_BASE (alternative base). ”predict.txt” file contains variant predictions that have been made. ”gt.txt” file contains the correct variants. Using the data in these two files, the performance of **three data structures (Binary Search Tree, AVL Tree, Unsorted Singly Linked List)** will be compared for the following operations.

- Reading the data from ”gt.txt” file and adding it to the specified data structure (Binary Search Tree, AVL Tree, Unsorted Singly Linked List).
- Reading the data from ”predict.txt” file and adding it to the specified data structure.

  Note: Predictions (”predict.txt” data) and ground truth (”gt.txt” data) should be kept in separate data structures. When adding, the insertion should be done first according to the chromosome order, then according to the position order for tree structures. When adding to a unsorted Single Linked List, you can add to the end.
- After adding the data from ”predict.txt” file to the specified data structure, adding a prediction variant with CHROM, POS, and ALT_BASE information to the data structure where predictions are stored.
- After adding the data from ”predict.txt” file to the specified data structure, deleting a prediction variant with CHROM, POS, and ALT_BASE information from the data structure where predictions are stored.
- After adding the data from ”predict.txt” file to the specified data structure, listing the prediction variants with CHROM, POS, and ALT_BASE information from the data structure where predictions are stored. For the linked list, you should start from the head and go to the tail. For the trees, you should perform inorder traversal.
- After adding the data from ”predict.txt” file to the specified data structure, searching a prediction variant with CHROM, POS, and ALT_BASE information from the data structure where predictions are stored.
- After adding the data from ”predict.txt” and ”gt.txt” files, calculating true positive variant count. True positive variant count indicates how many of the predicted variants are correctly predicted. Both the predictions and ground truth data structures should be used for this calculation. In addition to the specified data structure, ”std::vector” can be used for this calculation.

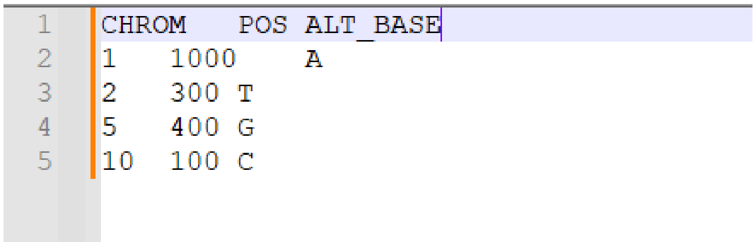
Note: All operations except true positive variant count operation should be performed on the specified data structure using specified data structure (For example, when searching in Binary Search Tree data structure, data from the Binary Search Tree should not be transferred to an array and searched on the array). An additional data structure can only be used for true positive variant count operation. ”std::vector” data structure can be used for this operation.

Especially when performing addition or deletion operations, Memory leaks should be avoided. Memory Leak check will be performed using **valgrind**.

You can use chrono library to measure execution time.

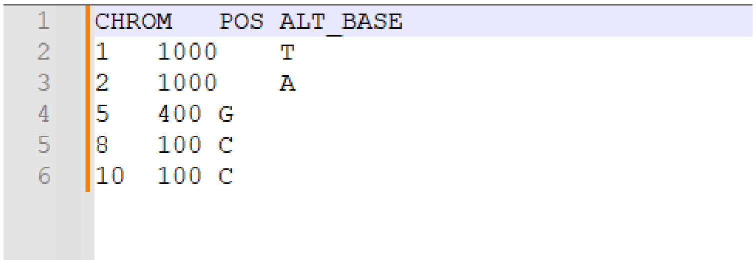
An example scenario is shown on the next page.

2 Example



```
CHROM   POS ALT_BASE
1   1000    A
2   300 T
5   400 G
10  100 C
```

Figure 1: Example gt.txt file.



```
CHROM   POS ALT_BASE
1   1000    T
2   1000    A
5   400 G
8   100 C
10  100 C
```

Figure 2: Example prediction.txt file.

Terminal Screen

```
Choose a data structure
1: Binary Search Tree
2: AVL Tree
3: Unsorted Singly Linked List
Enter a choice 1,2,3:1
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:1
Ground truth data structure was created from file in 10 ms
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:2
Prediction data structure was created in 10 ms
```

Example gt.txt and prediction.txt files are as shown in Figure 1 and Figure 2, respectively. Generated Binary Search Trees for prediction and ground truth after the operations performed in the terminal are as shown in Figure 3.

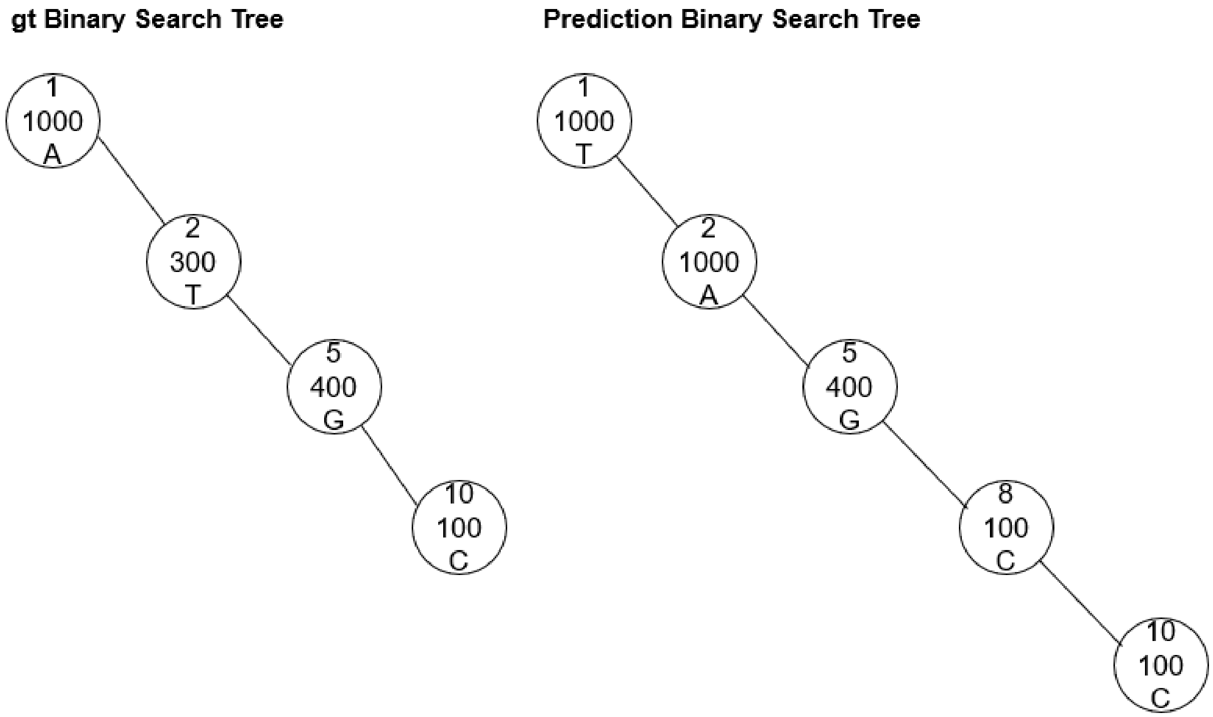


Figure 3: Created Prediction and gt Binary Search Trees from files.

```
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:3
Enter the CHROM POS and ALT_BASE information with a space in between: 2 300 T
2 300 T was added in 7ms
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:4
Enter the CHROM POS and ALT_BASE information with a space in between: 1 1000 T
1 1000 T was deleted in 7ms
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:4
Enter the CHROM POS and ALT_BASE information with a space in between: 1 1000 A
1 1000 A could not be deleted because it could not be found
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:5
2 300 T,2 1000 A,5 400 G,8 100 C,10 100 C
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:6
Enter the CHROM POS and ALT_BASE information with a space in between: 10 100 C
10 100 C was found in 5ms
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:6
Enter the CHROM POS and ALT_BASE information with a space in between: 1 1000 A
1 1000 A could not be found
```

```
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:7
True positive variant count is 3. It took 7 ms to calculate.
Choose an operation
1: Create ground truth data structure from file
2: Create prediction data structure from file
3: Add a variant prediction
4: Delete a variant prediction
5: List predictions
6: Search a prediction variant from predictions
7: Calculate true positive variant count
0: Exit
Enter a choice 1,2,3,4,5,6,7,0:0
```

3 Report

Write a report about your homework. This report should include the following:

- How to run the program you will submit.
- Report should include the execution times of adding from file, manual adding, deletion, searchig, listing, and calculate true positive operations for Unsorted Single Linked List, Binary Search Tree, and AVL Tree.
- Advantages and disadvantages of the data structures for the specified operations. For example, deletion operations are slower in linked lists compared to other data structures.

4 Submission Rules

- Make sure you write your name and number in all of the files of your project, in the following format:

  /* @Author
  Student Name: <student_name>
  Student ID : <student_id>
  Date: <date> */
- Use comments wherever necessary in your code to explain what you did.
- Do not share any code or text that can be submitted as a part of an assignment (discussing ideas is okay).
- Only electronic submissions through Ninova will be accepted no later than deadline.
- You may discuss the problems at an abstract level with your classmates, but you should not **share or copy code** from your classmates or from the Internet. You should submit your **own, individual** homework.
- Academic dishonesty, including cheating, plagiarism, and direct copying, is unacceptable.

- If you have any question about the recitation, you cand send e-mail to Yunus Emre Cebeci(cebeci16@itu.edu.tr).
- Note that **YOUR CODES WILL BE CHECKED WITH THE PLAGIARISM TOOLS!**