

Practice for associative arrays – `unordered_map` in C++ and vector

Task:

Given a data file of biological sequences in which each sequence (a string) follows its label (a string), find duplicated sequences and rebuild the data file with multiple labels for duplicated sequences.

Example input data file segment:

```
Label1
ABCDEFGHIJK
Label2
CDEFGHIJKL
Label3
ABCDEFGHIJK
....
```

Corresponding output data file segment:

```
Label1, Label3
ABCDEFGHIJK
Label2
CDEFGHIJKL
....
```

Note that the output data file does not have to be in a sorted order.

Please use C++ `unordered_map` and vector to accomplish the task.

The `unordered_map` is needed to search duplicated sequences efficiently, and the vector is needed to keep all the labels of a duplicated sequence. For example, if a sequence appears 5 times with different labels, the output file should have that sequence preceded by those five labels (separated by a comma each – see the example above).

Input: “Prog4-data” – placed in the Blackboard;

Output: The resulting data file after manipulating the duplicated sequences.

Sequences in the resulting file do not have to be in a sorted order nor the same order used in the input data file.

Submission:

Hard copy of your source code, and submit the output data file (in text file format) by email to T.A.