CS117 (Fall 2017) Programming Assignment 4 20 pts. Due: Oct. 12 (Th) – Lab1
Oct. 16 (M) – Lab2

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

Task:

Given a data file of biological sequences in which each protein sequence consists of a label and a string of amino acids, find duplicated sequences and rebuild the data file with multiple labels for those duplicated sequences.

Example input data file segment:

```
123456
MLINAGDPSDHLYYIISGSVTVIVEDDEGREIIVAYLNEGDFFGEMGLFDDENEERSAWVKTKTSC
EIAEIAYDTFHELRLSHPEFMLAGTR
134567
MDMKEEAEVQQEEATTDPTLEWFLSHCHIHKYPAKSTLINAGEKAETLYYLIKGSIAVSVKDDEG
KEMILSYLSQGDFFGELGLFEDVKVRSAWVKAKTTCEVAEISYK
234567
MLINAGDPSDHLYYIISGSVTVIVEDDEGREIIVAYLNEGDFFGEMGLFDDENEERSAWVKTKTSC
EIAEIAYDTFHELRLSHPEFMLAGTR
```

Corresponding output data file segment:

```
123456, 234567
MLINAGDPSDHLYYIISGSVTVIVEDDEGREIIVAYLNEGDFFGEMGLFDDENEERSAWVKTKTSC
EIAEIAYDTFHELRLSHPEFMLAGTR
134567
MDMKEEAEVQQEEATTDPTLEWFLSHCHIHKYPAKSTLINAGEKAETLYYLIKGSIAVSVKDDEG
KEMILSYLSQGDFFGELGLFEDVKVRSAWVKAKTTCEVAEISYK
```

- **Please use C++ <u>unordered_map</u> and <u>vector</u> 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.

<u>Submission:</u>

Please submit a zip file containing your source code (don't forget documentation) file and output data file by email to: jpark@csufresno.edu

Please make your zip file name and the email subject field as the following:

CS117-Prog4-yourFirstName-yourLastName.zip