Name: David Southwell

Lab Title: BLAST

Date: 6/8/2017

Before the Lab:

What I expect to learn: How to use my knowledge of databases, indexing and sorting data to display statistics about DNA sequences.

Concepts I think the Lab explores: Indexing databases and utilizing hash tables

Java classes I expect to modify: Location, Validated Input reader?

Where I think there may be problems: Using a hash table as a way of indexing my database.

After the Lab:

What I actually learned: How to utilize a hash table to find matches of substrings within dna sequences, and how to extend the length of the matches.

How well the Lab allowed me to explore the target concepts

What problems I encountered (or none) and how I handled them: I had some issues comparing data read in from the dna data file to the data from the query file, but everything worked after I implemented the QueryReader as a separate class from the DNAReader class.

Things in the write up that confused me or were especially helpful: I found the variety of flowcharts to be very helpful in conceptualizing the steps needed to solve the problem.

Additional Questions:

* What are the objects involved in this project? (May include such things as DNA sequence, query sequence, word, database, hash table, and simulation.)

->Location objects and strings are the most commonly used objects in this program.

* What data structure will you use to hold the virus DNA sequences?

->ArrayLists will be used to store substrings and locations of those substrings within a given sequence.

* Will this data structure be part of a class? If so, what are the responsibilities of this class?

->It is part of the SubStringLocs class which keeps track of the substring location in regards to a given substring.

* What class should include the hash table?

->The simulation class deals with the hash table.

* What information do you need to save when you find a match? Should you create a new type of object for this?

->which viruses contained matches, as well as the starting locations and length of each match within the virus, and the matched substrings.

* Where should the user specify filenames for the DNA sequence data and the query sequence data? Where should they specify *n*, the substring length to use as the words? Where should they specify the length to use for significant matches?

->This should all be specified in the main method of BLASTAppl but it is then passed to the simulation class for analysis.

* Which parts of your Indexing Substrings project might you be able to build from?

->The indexer class will be a good platform to build a simulation class for this project.