MicroRNA discovery

Due date: May 22, midnight (Wednesday).

About the Lab

Your **BIO 441** partners are searching for microRNAs in the genome fragments they are studying. MicroRNAs are small non-coding fragments of the DNA[1] that have a special biological role: they regulate gene expression, i.e., help increase or decrease production of a specific protein in a cell.

The key feature of microRNAs is that they are formed out of DNA fragments that that have a stem-loop structure, otherwise known as a *hairpin*.

A hairpin is formed, when a single DNA strang can "bend" and bind to itself. For this to be possible, a portion of the DNA fragment (a substring) must be an *exact reverse complement* of some other, *nearby* portion. The nucleotides in between these two portions form the "gap" or the "loop" of the hairpin. Figure 1 illustrates this situation.

Your task is to help your **BIO 448** partners by creating software for microRNA detection in DNA fragments. The overall structure of the lab is as follows. The secondary goal of the assignment is to imrpove the process of building the data structure you are using for microRNA discovery.

Time	CSC 448	CHEM 441
May 14 lab	Requirements/Design	
May 16 lab	Design/Implementation	
May 16 – May 21	Software development	Testing
May 21 lab	Testing	
May 22	Finishing software/ testing	

[1] Technically, microRNAs are small RNA molecules, but **BIO 441** students are looking for their "codes" in the DNA fragments.

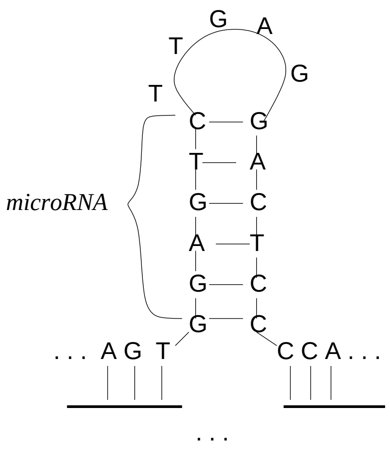

Figure 1: Hairpins, and microRNA formation.

Lab Assignment

You have two goals for this assignment.

Goal 1: microRNA discovery. Your **BIO 441** partners will provide specific requriements for you. Please note the following.

As shown in Figure 1, you are looking for possible hairpin locations in the DNA strings provided to you. There are some "movable parts" to this search: the expected length of the microRNA and the expected size of the hairpin (gap) are two key parameters that your program must be able to use to find appropriate microRNAs.

Goal 2: Suffix tree improvement. It is very likely that to solve the problem of microRNA finding, you will need to use suffix trees. In **Lab 4** you were asked to implement an *ad hoc* method of suffix tree construction, which takes $O(n^2)$ running time, where n is the length of the string inserted into the suffix tree.

This time, you are asked to improve the suffix tree construction and implement *Ukkonen's linear-time suffix tree construction approach* discussed in class on May 7 and May 14.

Note, this improvement should be invisible to your **BIO 441** partners (this is the one time this quarter when we are doing something that does not affect the results of your software), however, it should improve the performance of your software, which is a worth-while goal in and by itself.

Also note, that you may need to change the suffix tree data structure itself to accommodate the information that helps you solve the microRNA discovery problem (as opposed to repeat finding). Otherwise, the instructions on how to implement suffix trees from **Lab 4** remain in force.

Submission Instructions

As usual, two sets of deliverables are needed: one for `handin` and one - for `Piazza`. The core deliverables are the code and two requirements documents: the original document provided to you by your BIO 448 partners and the final version.

handin deliverables. The `handin` deliverables are: *source code*, *compilation/running instructions (README)*, *user documentation* and *both requirements documents*. Submit them using the following command

```
$handin dekhtyar 448-lab5 <files>
```

Piazza delivarables. The final working version of the software (executable) must be delivered to your partners via Piazza. Additionally, the user documentation and the requirements documents need to be available on Piazza as well.

Deadlines. There is only one deadline: submit everything by the end of the calendar day on Wednesday, May 22.