Lab 3: Search Algorithms and making sense of data

Due date: September 18 11:59pm

Goals: The objective of this lab is to implement the Sequential search algorithm and to apply it to two datasets, combine the results and plot them. In this lab you will gain experience in working with strings in Matlab and making bar graphs.
Searching through a set of objects (numbers, characters, etc.) is a basic task in computations. For example, we could search a real vector to find where a particular value occurs. Our goals here are to search DNA strings and plot the frequencies of occurrence of microsatellites.
On the website www.peterbeerli.com/classdata/ISC4221/data are two files you will need to download to your account, the fastest way do this is by using the 'wget' command in a unix terminal:

```
wget www.peterbeerli.com/classdata/ISC4221/data/human_chrX.fasta.gz
wget www.peterbeerli.com/classdata/ISC4221/data/human_chrY.fasta.gz
```

You will need to `gunzip` these two files. These two files are very large and you may fail to look at them with an editor. The smaller file (chrY) has 59373566 relevant characters. The larger one (chrX) has 155270560. These are two human chromosomes X and Y from the human reference sequence, respectively. X usually called the female sex chromosome and Y the male sex chromosome. The size difference stems from the fact that the Y chromosome contains very few functional genes and is mostly 'useless', whereas the X contains many vital genes (for more info check http://en.wikipedia.org/wiki/X_chromosome and http://en.wikipedia.org/wiki/Y_chromosome). Most of the human DNA is actually not coding for anything (proteins, hormones etc), but are spacers between these coding regions. Our lab will focus on repeat structures that are present in these two chromosomes and compare them.
The human genome consists of thousands of repeated small fragments such as `CA` or `TA`, you will write matlab code to

1. read one chromosome, make sure to discard the first line that starts with '$>$' and then concatenate all lines after that, the data contains mostly A, C, G, T, and N characters.
2. count the occurrences where the repeat motif `CA` occurs $3, 4, 5, 6, ..., n$ times. You will need to search the chromosome repeatedly to get counts for the occurrences of these repeats (for example in my tests the repeat `CACACACACACA` [we have repeated the motif `CA` 6 times] occurs 345 times in the X chromosome.
3. histogram the occurrences
4. do the same for the other chromosome and discuss similarities or differences.
5. Now pick another repeat `TTC` and compare the two chromosomes
6. Pick another 3 letter motif of your own choice (do not pick any 'N') and show results (Hint: not all 3 letter words will be found in the sequences)
7. Bonus points: What is the largest motif that is repeated at least 3 times?

Matlab Syntax:

You can easily make bar graphs in Matlab. If you have an array such as `y = {4, 6, 2, 10}` and you use the command `bar(y)` then it will make a bar graph with the bars at a height of 4,6,2,10 and it will assume the x-values are $1, 2, 3, 4$. If you have an array of x values, then you can use `bar(x,y)` For labeling bar charts you may ant to consult
`www.mathworks.com/support/solutions/en/data/1-15HXQ/`