DNA Mapping allows researchers to identify Single Nucleotide Variants in subject data. The experiment will normally consist of several million reads randomly drawn from the subjects genome. The mapping software will identify where each of these reads matches best to the genome. Your implementation must be very efficient to complete this task in a reasonable amount of time. Minimal implementations of the project must map reads without errors to a reference genome. Full implementations will be able to map reads with errors (See rosalind 7o).

Given the following genome from a fasta file (example.fasta):

```
>Chr1

actacccgattcagggaaattcatacaggaatatttg
```

And the following reads from another fasta file:

```
>R1

gattcaggg

>R2

attcataca

>R3

ggaatattt

>R4

cagggaaat
```

The following mapping would result:

```
00000000011111111112222222223333333

12345678901234567890123456789012345678

actacccgattcagggaaattcatacaggaatatttg

R1     gattcaggg

R2                attcataca
```

```
R3                          ggaatattt

R4          cagggaaat
```

R1 would be mapped to location 8, R2 to location 19, R3 to location 28 and R4 to location 12.

The output of a mapping algorithm will often be represented in Sequence Alignment/Map (SAM) format. Although the format is complex and can be used to specify gaps, mismatches and rearrangements, you will be able to use a subset of the format to describe your mapping. The general format of a SAM file entry is:

```
Col Field Type    Brief description

1   QNAME String  Query template NAME

2   FLAG  Int     bitwise FLAG

3   RNAME String  Reference sequence NAME

4   POS   Int     1-based leftmost mapping POSition

5   MAPQ  Int     MAPping Quality

6   CIGAR String  CIGAR string

7   RNEXT String  Ref. name of the mate/next read

8   PNEXT Int     Position of the mate/next read

9   TLEN  Int     observed Template LENgth

10  SEQ   String  segment SEQuence

11  QUAL  String  ASCII of Phred-scaled base QUALity+33
```

An example SAM entry for R1 shows a zero flag, Chr1 reference, 8 for position (remember it is 1 based), mapping quality 255 unknown, 9 characters matched for the CIGAR string, no paired read (*) and 0 for the pair name and position, gattcaggg for the sequence, * for no base quality. You can put this in a file example.sam (Be sure to put tab characters between the fields).

```
R1      0       Chr1    8       255     9M      *       0       0
gattcaggg       *
```

Now, to visualize the mapping, you will need to convert the SAM file into a binary BAM file and sort it. Run the following commands:

```
samtools view -bT example.fasta example.sam > example.bam

samtools sort example.bam example.sorted

samtools index example.sorted.bam
```

This will create a example.sorted.bam file along with its index. In order to view the alignment, you can run:

```
samtools tview example.sorted.bam example.fasta
```

Approaches

You can use any of the following approaches to implement your mapping algorithm:

1. Suffix Trees or Suffix Arrays - In order to deal with reads with errors you will have to implement backtracking and speculative path extending. But you should be able to easily map reads without errors.
2. Burroughs Wheeler - Problem 7o illustrates how to deal with errors using BW mapping.
3. Hash based kmer mapping - Hash all of the kmers in the genome to create an index for all kmers. Slide a window across the read and grab the index values for each kmer and put it in a list. Then compute the number of differences for each location and map to the best location. If you pick a kmer value that is less than half of the read size, you are guaranteed to be able to tolerate at least one error.

Data Sets

The following data sets can be used to help you to determine if your algorithm is working correctly.

1. (10 Points)The example data set from this lab writeup. Include a screenshot of the whole region in samtools tview. The data set has no errors.
   - example.fasta Download
   - example.reads.fasta Download
2. (30 Points) A simulated data set of RNASEQ reads. You should expect to see areas of the genome where there are no reads corresponding to non-coding regions. You should see other regions that have different depths corresponding to differential expressoin. You should report the region that has the highest depth or expression (give starting and ending positions for the gene). Include a screenshot of the depth throughout the

chromosome (you can generate this in excel using output from pysam). There are no errors in this data set.
    - You can find the data on schizo: at /users/faculty/clement/public_html/cs418/Ch7-project/rnasim.fa and the reference genome at /users/faculty/clement/public_html/cs418/Ch7-project/rnaseqChr.fa
3. (60 Points) Shotgun data with SNVs corresponding to regions of variations between the reads and the genome. You should report on locations for the SNVs you found and should show a screenshot of one of the SNVs you feel you are most confident in. This dataset will have errors and will not be able to be mapped with normal suffix tree or BW algorithms. You should map the following data sets and determine where the SNVs are.
    - A large, but not crazy large (100,000 reads) dataset in fastq format of real data with errors, forward and reverse reads, and the accompanying quality scores. You can find the data on schizo: at /users/faculty/clement/public_html/cs418/Ch7-project/19.small.fastq and the reference genome at /users/faculty/clement/public_html/cs418/Ch7-project/chr19.fa
    - (+20 Extra Credit)A very large (~23.5 million reads) dataset in fastq format of real data with errors, forward and reverse reads, and the accompanying quality scores. You can find the data on schizo: at /users/faculty/clement/public_html/cs418/Ch7-project/19.fastq and the reference genome at /users/faculty/clement/public_html/cs418/Ch7-project/chr19.fa

Deliverables:

1. A written report including:
    a. Method(s) used for mapping
    b. Method(s) used for handling errors
    c. An analysis of the quality of your mapping for each data set
    d. Comparison to another mapping algorithms like bowtie2 and bwa
    e. Ideas on how your mapping algorithm might be improve
    f. include 2 recommendations for improving this project for next semester
2. Attached files:
    a. Samtools tview screen shot for a region of each data set that you find interesting
    b. Include your source code as an attachment, include all files necessary to run it as well as instructions on how to run it

Good luck!