# MIDTERM STUDY GUIDE

COM SCI/HUM GEN CM122/222
CHEM CM160B/260B
BIOINFO M260B

Date: Wednesday, February 1, 2017
Instructor : Eleazar Eskin

**Name (Print):** _____
**UID :** _____

Undergraduate ☐ Graduate ☐

Guidelines for midterm.

1. Please write down your name, UID, and whether you are un undergraduate or a graduate student.
2. Closed book and no calculator.
3. Your answers should not require much more space than is provided. Be concise.
4. There is an extra blank sheet of paper at the end of exam. Please use it if you need more space for your answer.

| Question | Score |
|---|---|
| 1 | /15 |
| 2 | /25 |
| 3 | /15 |
| 4 | /25 |
| 5 | /20 |
| Total | |

# 1 Biology Basics (15 pts)

*3 points per question. On the midterm there will only be 5 questions similar to the ones below.*

1. How long is the human genome?
2. What are the nitrogenous bases that can occur at each position in the genome?
3. If you were to encode the entire human reference genome in a binary array that supports random access in constant time, approximately how much memory would it occupy in bytes? Show your work.
4. What is the definition of "coverage" as it relates to sequencing?
5. List one advantage and disadvantage of using 30x coverage versus using 5x coverage.
6. What is the reference sequence? the consensus sequence? the donor sequence? a read.
7. A 150-base sequence in the reference genome has been duplicated in the donor. How can this be observed when aligning reads to the reference?
8. Is it easy or difficult to map reads uniquely to long repetitive portions of the genome?
9. Why is it difficult to call SNPs accurately in long repetitive regions of the genome?
10. Why can a 30 base pair long read perfectly align to multiple locations in the genome?
11. What is assemply?
12. What is re-sequencing?
13. Why is assembly much harder than re-sequencing?

## 2 Trivial Aligner

### Part (a)

Write (pseudo-)code that aligns a single read `read` to a reference genome `ref` using the trivial alignment algorithm with up to `m` mismatches. You should return the index in the genome of an acceptable alignment, if it exists.

### Part (b)

Describe one way to improve the speed of the trivial aligner without fundamentally changing the algorithm (i.e you're not allowed to say "use hashing instead").

### Part (c)

Assume that the only operation that takes time in the trivial aligner is comparisons between the read and the reference. If the computer takes $t$ seconds to perform a single comparison, how much time will be required to align $N$ reads of length $k$ to a genome of length $L$ with up to $m$ mismatches?

### Part (d)

Assuming no read errors, when will the trivial aligner not be able to map reads?

## 3 Alignment by Hashing

Write an algorithm that aligns reads by hashing. How does your algorithm find the correct alignments for reads that span SNPs? Why is it advantageous to divide your read into 3 even sections? How many read errors or SNPs can exist on a read and have the read still be alignable to the correct position? Does it matter where the read errors and SNPs occur? How does the hash-based aligner compare to the trivial aligner and the BWT-based aligner with regards to memory and runtime?

# 4 Pileup

1. What is the consensus sequence generated by these reads?

   ```
   -Ref:GCATAGGCATG
   Read:...TAGGCATG
   Read:G.........
   Read:GCATTA.....
   Read:GCATTAG....
   Read:CATTAGGCA..
   Read:GAATTAGGCAT
   Read:GCATTAGGCAT
   ```

2. What is the most likely change from the reference genome that generated the reads?

3. What is the most likely donor sequence that generated these reads?

4. Show evidence for this being the actual change.

5. If a pileup algorithm does not call insertions and deletions, how will this effect the type 1 and type 2 error of detecting SNPs.

# 5 Burrows-Wheeler Transform

1. Why do you need a "$" (or other special character) at the end of a string when performing a Burrows-Wheeler transform?

2. Is the BWT faster or slower than hashing methods? Why is it the most used algorithm for aligning reads?

3. What is the Burrows-Wheeler Transform of the string "ROCKANDROLL"?

4. Un-permute the string represented by the BWT: LWNLEOD$E