7.36/7.91/20.390/20.490/6.802/6.874
PROBLEM SET 2. BWT, Library complexity, RNA-seq, Genome assembly, Motifs, Multiple hypothesis testing (31 Points)

Due: Thursday, March 13th at noon in the dropbox labeled with 7.36/7.91 outside of the Biology Education Office on the ground floor of Building 68.

Python Scripts
All Python scripts must work on athena using /usr/athena/bin/python. You **may not assume availability of any third party modules** unless you are explicitly instructed so. You are advised to test your code on Athena before submitting. Please only modify the code between the indicated bounds, with the exception of adding your name at the top, and remove any print statements that you added before submission.

Electronic submissions are subject to the same late homework policy as outlined in the syllabus and submission times are assessed according to the server clock. **All python programs must be submitted electronically, as .py files** on the course website using appropriate filename for the scripts as indicated in the problem set or in the skeleton scripts provided on Stellar. To submit a file electronically:
1. Go to
http://stellar.mit.edu/S/course/7/sp14/7.36/homework/index.html
2. Click on the corresponding problem set.
3. On the Assignment Details page, click the Add Submission link.
4. On the Add Submission page, select the appropriate file on your
computer. **Do not use the paste box, do not zip files**
5. Click Submit when ready.

Problem 1. Aligning reads to a genome using a Burrows Wheeler Transform and FM Index (9 points)

For this exercise you will be implementing the core of a genome search function utilizing the Burrows Wheeler transform (BWT) and an FM-index. We have provided scaffolding code so that you can focus on the core of the algorithm. Please do not use Internet search tools to try to solve this problem – the point is for you to understand how the algorithm works.

You will need the coding and testing files from the course website (keep them in the same folder). This includes scaffold code, a 10kb segment of the yeast genome, reads which you will map to the genome, and an index for testing with correct output.

(A) (7 pt.) Complete the LF mapping code in the _lf(self, idx, qc) function and the search code in the bounds(self, q) function (both in fmindex.py).

To test your implementations we have provided the FM-index of an abbreviated version of the yeast genome in test.index. Running

```
% python fm-search.py test.index yeast_chr1_reads.txt out.txt
```

will place the mapped reads in out.txt and test your implementation. The correct output of this command is given in test_mapped.txt for you to check the correctness of your implementation. Your implementations of the _lf(self, idx, qc) and bounds(self, q) functions are the answer to 1.1. **Submit** fmindex.py.

```
  def _lf(self, idx, qc):
      """ get the nearset lf mapping for letter qc at position idx """
      o = self._occ(qc)
      c = self._count(idx, qc)
      return o + c

  def bounds(self, q):
      """ find the first and last suffix positions for query q """
      """These are positions in the BWT string"""
      """This is the meat of the FM search algorithm"""
      top = 0
      bot = len(self.data)
      for i, qc in enumerate(q[::-1]):#iterate over letters in query string q in reverse
          top = self._lf(top, qc)#returns occ(qc)+c(idx, qc), which maps the position in the last
column to the position in the first column
          bot = self._lf(bot, qc)
          if top == bot: return (-1,-1)#since bottom is non-inclusive, top==bot implies that the
string was not found.
      return (top,bot)
```

(B) (2 pt.) Now let's make sure your implementation works on a larger genome. First you will build the FM-index. To build the index use the command:

```
% python fm-build.py yeast_chr1_10k.txt yeast_chr1_10k.index
```

Now let's search using the FM index with the code you wrote, and use it to map some reads

To map the reads:

```
% python fm-search.py yeast_chr1_10k.index yeast_chr1_reads.txt
mapped_reads.txt
```

View the output:

```
% more mapped_reads.txt
example mapped_reads.txt (you will not have this same read):
ATGGGTATCGATCACACTTCCAAGCAACAC count:1 matches:[561]
…
```

Submit your mapped_reads.txt on Stellar as the answer to 1.2.

```
AATAGAATAACAGTTGTATGGGTCACCTGG      count:1   matches:[8024]
GGAAATTTATATATAAACTTCATTTACGTC      count:1   matches:[7126]
TGTATTCGTATGCGCAGAATGTGGGAATGC      count:1   matches:[2453]
ACTGCCAAATTTTTCTTGCTCATTTATAAT count:1   matches:[3086]
GTACTTTGAAACCTGATTTATATATTGCAG      count:1   matches:[6565]
ACTTACCCTACTCTCAGATTCCACTTCACT count:1   matches:[459]
TTCAGGACTTGCAAAAAGAATCTAACTGAT      count:1   matches:[6980]
AAATATTTGATTCATTATTCGTTTTACTGT count:1   matches:[4516]
TAATATAACTTATCAGCGGCGTATACTAAA      count:1   matches:[1181]
AACATTGCAGCATAAATGCAAACCATTTGG      count:1   matches:[7594]
CTAGTTACAGTTACACAAAAAACTATGCCA      count:1   matches:[1293]
TTATGATATTTTTTTTTTATAGTAGTAGTG count:1   matches:[6940]
TATTTTATTTTGTTCGTTAATTTTCAATTT  count:1   matches:[1741]
GCCTTATAAAACCCTTTTCTGTGCCTGTGAcount:1   matches:[2504]
TTTTCCACACCATGTTTAGAGTTATAAAGC      count:1   matches:[7284]
AAGTTAATATTATGGTGGTAGTATCTCAAA      count:1   matches:[4660]
TACTTACTACCACTCACCCACCGTTACCCT      count:1   matches:[195]
TCCATTCCCATATGCTAACCGCAATATCCT      count:1   matches:[1017]
AGTTTGGTACCATGACTTGTAACTCGCACT      count:1   matches:[1375]
TACAAATATATATTAAAGAAATCCAAACAA      count:1   matches:[10326]
GTTTTTTTAGTAATTTCTTTGTAAATACAG count:1   matches:[3634]
GAAAAATACATGAATGACAGGTAAAAATAT      count:1   matches:[3687]
TACTACTTTGTAAACCAGTGGATTTTTGCTcount:1   matches:[5939]
TAGCAGTTGTTATAACGACAAATACAGGCC      count:1   matches:[4209]
GCCACTACATGACAAGCAACTCATAATTTA      count:1   matches:[4324]
TATATCATCAAAAAAAAGTAGTTTTTTTAT      count:1   matches:[1714]
GGTCACTAATGAGAACTTAAATAGTTTTCA      count:1   matches:[5203]
CACACCCACACACCCACACACCACACCACA      count:1   matches:[6]
TGTAGCGAATGTCCATTCATCATAACAGGT      count:1   matches:[9081]
TCTTAATTTCAATTTCATGCCCTCCTTCAC count:1   matches:[5141]
```

Problem 2. Library Complexity (5 points)

Imagine you are responsible for sequencing DNA samples for your lab's latest important experiment. Using extensive simulations, you know that you need to observe at least 12 million unique molecules in order to test your current hypothesis. From previous experience, you know that each time a DNA library is constructed from a sample, it will contain exactly 40 million unique molecules (selected perfectly at random). You also know that C. elegans, your model organism, has a genome size of approximately 100 million base pairs.

You can have your sample sequenced in units called lanes. Each lane gives you 10 million reads, and a library can be sequenced on as many lanes as you want. However, ever-protective of your grant money, you want to achieve your experimental goals in the most efficient way possible. Suppose that each sample collection and library preparation step costs $500 and that each sequencing lane costs $1000.

(A) (2 pt.) Assume that each molecule in the library had equal probability of being sequenced. What is the most cost-effective experimental design (number of libraries and lanes sequenced for each library) for achieving your goal of observing 12 million unique molecules? Show your work.

Consider one library. Compute M = K(L)*C = (1-exp(-N/40M))*40M.

One lane: 8.8M unique reads ($1500) (1-poisspdf(0,.25))*40

** Two lanes: 15.7M unique reads ($2500) (1-poisspdf(0,.5))*40

Consider two libraries:

When considering the 2nd library, all calculations for the number of unique reads within the 2nd library are the same (the same number of reads are coming off the sequencer and the average coverage L for these reads is the same). However, we only add 60% of them to the unique reads from the 1st library to get the total unique reads over both libraries since on average 40% of the reads in the 2nd library will have already been covered by the 1st library. Therefore:

One lane each: 8.8M unique reads from first library, 8.8M*0.6=5.3M from second = 14.1M ($3000)

Two lanes from 1st library, one lane from the 2nd: too expensive - ($4000)

(B) (3 pt.) Now suppose that there is variation in the selection probabilities across each molecule, which follows a negative binomial distribution with rate lambda = 0.25 (10 million reads divided by 40 million molecules) and variance factor k = 2 (estimated from previous experiments). What is the most cost-effective experimental design for this situation? Show your work and comment on any differences between the two cases.

Hint: A more common formulation of the negative binomial distribution is in terms of failures n and a success probability p. This conversion is found in the lecture slides.

Same as (A) but we compute K(L) with the NB distribution.
We know that p=L/(L + 1/k) and vary L with the number of lanes.
Note: if using the Matlab or Mathematica implementations of the NegBin, you should actually use 1-*p* if calculating *p* as mentioned in lecture.

One library:

($1500) One lane: 7.34M unique reads (1-nbinpdf(0,0.5,2/3))*40
($2500) Two lanes: 11.7M unique reads (1-nbinpdf(0,0.5,0.5))*40
*($3500) Three lanes: 14.7M unique reads (1-nbinpdf(0,0.5,0.4))*40

Two libraries:

($3000) One lane each: 7.34M + 4.4M = 11.74M unique reads (not enough)
First library: 7.34M unique reads as above
Second library: 4.4M unique reads (1-nbinpdf(0,0.5,2/3))*40*0.6
Two lanes (one library) then one (second library): over 12 million unique reads but too expensive

Problem 3. Differential gene expression (4 points)

You are analyzing RNA-seq data to identify differentially expressed genes between two treatment conditions. You have three biological replicates in each of the two conditions for a total of 6 samples, and you process and sequence each of the samples separately.

(A) (1 pts) Imagine you first pool the sequencing results for each of the conditions, resulting in two pools. What kind of variation have you lost the ability to observe, and why might this variation be important?

This is performing analysis without any replicates. If we observe a difference between the conditions, we are unable to know if this difference is due to differential expression between the different conditions or due to baseline variation between the replicates (just due to technical or biological variation).

(B) (3 pts) Devise an improved analysis strategy for these six samples and identify the sources of variation it can detect. Identify how you would estimate the mean-dispersion function for use in a negative binomial model of variation.

Don't pool the sequencing results together. Now, since we have replicates for each condition, we can compute an empirical dispersion value per gene, rather than estimating dispersion from genes which have similar expression levels across conditions (under the assumption that the condition effect is minimal for these genes). This allows us to detect technical/biological variation among samples within the same condition as well as variation in excess of that due to differential expression between the two conditions.

Problem 4. RNA Isoform quantification (3 points)

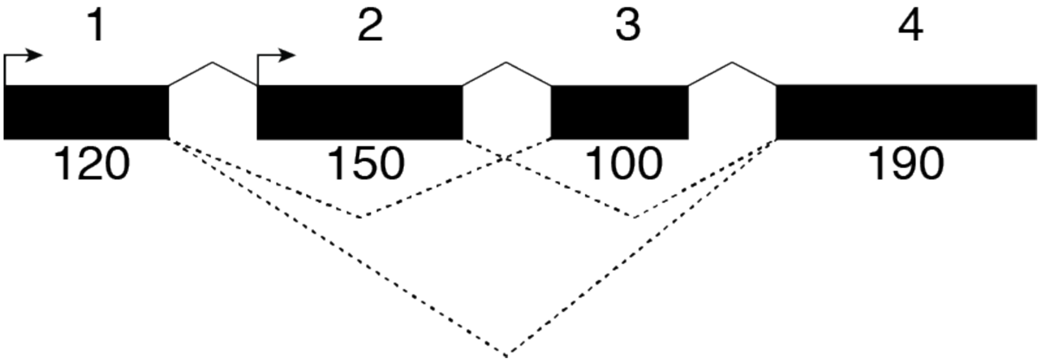


Fig. 1

Consider the gene structure in the above figure.

Exon numbers and sizes in nucleotides are indicated. The transcript can initiate at either of the arrows shown, and exons 2 and/or 3 can be spliced out.

(A) (1 pt.) How many possible isoforms of this gene could exist?

6 isoforms

(B) (1 pt.) For each isoform, list the junction spanning RNA-seq reads that would support it.

Isoform	**Reads**
1-4	Only 1-4 spanning reads
1-2-4	1-2, and 2-4 spanning reads
1-3-4	1-3 and 3-4 spanning reads
1-2-3-4	1-2, 2-3, and 3-4 spanning reads
2-4	Only 2-4 spanning reads
2-3-4	2-3 and 3-4 spanning reads

(C) **(1 pt.)** Assuming single ended reads, what is the shortest read length that would guarantee the ability to unambiguously identify all isoforms of this gene if we require that a junction read must have minimum overlap of 5bp with each exon?

260bp – the 150bp of exon 2, 100bp of exon 3, and 5bp overlap with exons 1 and 4

Problem 5. de Bruijn graphs (5 points)

Suppose you are interested in sequencing a particular RNA sequence. You opt to take a next generation sequencing approach and submit your sample to your local sequencing facility. You receive the following set of 6 bp reads in return, which are all in the same orientation.

AGCTGT, CAGCTG, TTCTGC, GCTGTA, TCAGCT, CTGTAT, TGTAGC, TTCAGC, CTGTAG, TTTCAG

(A) **(1 pt.)** Construct the corresponding de Bruijn graph with k = 5

(B) **(1 pt.)** Simplify any chains in the graph. Remove any tips present in the graph.

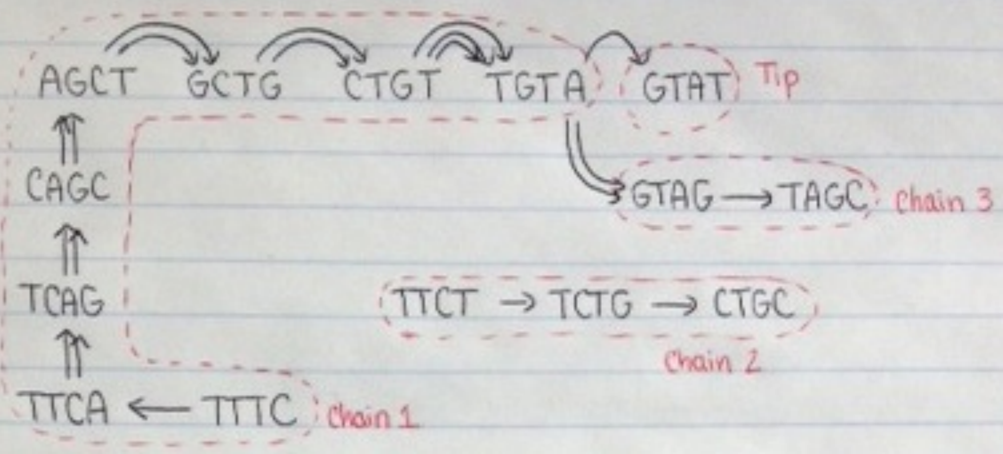
(C) **(1 pt.)** Identify any bubbles in the graph. Resolve the bubbles by removing the path most likely to be caused by a sequencing error.

(D) **(1 pt.)** Which read(s) contain sequencing errors? Identify the error(s).

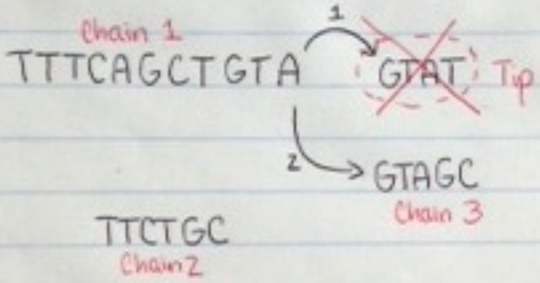
(E) **(1 pt.)** Write the sequence represented by the de Bruijn graph after the error correction steps.

For K=5:

(A)



(B)



(C) No bubbles

(D) TTC$\overset{A}{\text{T}}$GC CTGTA$\overset{C}{\text{T}}$

(E) TTTCAGCTGTAGC

Problem 6. Modeling and information content of sequence motifs (5 points).

To analyze gene evolution in three phylogenetic groups of protists, you collect samples of three different protist species, A, B, and C, that represent these lineages. You conduct both genome sequencing and cDNA sequencing from each and use spliced alignment of cDNAs to genomes to obtain sets of 10,000 confirmed 3' splice site (3'SS) sequences from each species. In all three species the invariant AG at the end of each intron is preceded by an 8 base polypyrimidine tract (PPT), with frequencies $f_C = f_T = \frac{1}{2}$ at each position. Your goal is to develop probabilistic models of the PPT motif in each species for use in exon-intron prediction. Throughout this problem, unless instructed otherwise, you should describe the simplest possible model (fewest parameters) that accurately models the frequencies of all 8mers in the training data (and should therefore give good predictive accuracy). Information content of models should be calculated using the formula given in lecture: I = 2w – H(model), in bits, where w is the width of the motif and H(model) is the Shannon entropy of the model. The abbreviation Y_8 refers to 8mers that consist exclusively of pyrimidine (C or T) nucleotides.

(A) (1 pt.) In species A, all four dinucleotides CC, CT, TC, and TT occur equally often $\left(f_{CC} = f_{CT} = f_{TC} = f_{TT} = \frac{1}{4}\right)$ at each of the seven pairs of positions (1,2), (2,3),…,(7,8), and each 8mer of the form Y_8 occurs with frequency 2^{-8}. In one sentence, describe a model for the PPT of species A. What is the information content of this model?

The simplest model is a weight matrix model, with P(C) = P(T) = ½ at each position.

H(model) = 8 x [-((½ $\log_2$ (½) + ½ $\log_2$ (½))] = 8 bits.

Information = (2 x 8) – 8 = 8 bits.

(B) (1 pt.) In species B, all four dinucleotides CC, CT, TC, and TT are equally likely $\left(f_{CC} = f_{CT} = f_{TC} = f_{TT} = \frac{1}{4}\right)$ at each of the seven pairs of positions (1,2), (2,3),…,(7,8), but examining the frequencies of 8mers reveals that $f_{T_8} = f_{C_8} = f_{(TC)_4} = f_{(CT)_4} = \frac{1}{4}$. In one sentence, describe a model for the PPT of species B. What is the information content of this motif?

The simplest model is one that assigns $f_{T_8} = f_{C_8} = f_{(TC)_4} = f_{(CT)_4} = \frac{1}{4}$. (4 nonzero probabilities)

H(model) = -[4 x (¼ $\log_2$ (¼))] = 2 bits (using the fact that $0 \log_2(0)$ is defined to be 0 (by continuity) in information theory).

Therefore, Information = (2 x 8) – 2 = 14 bits.

(C) (3 pt.) In species C, $f_{CC} = f_{TT} = \frac{3}{8}$, $f_{TC} = f_{CT} = \frac{1}{8}$ at each of the seven pairs of consecutive positions (1,2), (2,3),…,(7,8), and the frequencies of all 8mers of the form Y_8 are equal to $3^{a+b}/Z$ where a is the number CC dinucleotides in the 8mer and b is the number of TT dinucleotides in the 8mer, and Z is the normalization constant that causes the frequencies to sum to 1. In one sentence, describe a model for the PPT of species B. What is the information content of this motif?

Recognize that this distribution can be achieved by use of a first-order Markov model with parameters $f_C = f_T = \frac{1}{2}$ at position 1, and conditional probabilities P(C|C) = P(T|T) = ¾ and P(C|T) = P(T|C) = ¼ at all subsequent positions.

To calculate the information content of the model, let $k = 7 - (a + b)$ be the number of CT and TC dinucleotides. The probability of generating an 8mer sequence with k CT and TC dinucleotides is $P_i(k) = \frac{1}{2}\left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{7-k}$ (the factor of ½ is because there is a ½ probability of the first nucleotide – C or T). The total number of sequences with k CT and TC dinucleotides is $2\binom{7}{k}$ (the factor of 2 is for the two possible first nucleotides – C or T). You can check that the total probability of all sequences $\sum_{k=0}^{7} P_{total}(k) = \sum_{k=0}^{7} 2\binom{7}{k}\frac{1}{2}\left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{7-k} = 1$ as required.

The Shannon entropy of the motif is:

$$-\sum_{i=1}^{2^8} p_i \log_2(p_i) = -\sum_{k=0}^{7} P_{total}(k) * \log_2(P_i(k))$$

$$= -\sum_{k=0}^{7} 2\binom{7}{k}\frac{1}{2}\left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{7-k} * \log_2\left(\frac{1}{2}\left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{7-k}\right) = 6.68\ bits$$

Therefore, Information = (2 x 8) –6.68= 9.32 bits.

(Extra 6.874 Problem) Multiple Hypothesis Testing (4 points)

Differential expression analysis of RNA-seq data involves testing thousands of hypotheses in a single experiment. To limit false positives, it is necessary to adjust P-values. Two popular methods for doing so are Bonferroni correction and Benjamini-Hochberg.

Consider the following uncorrected p-values of 20 genes from a gene expression study in which we wish to identify differentially expressed genes, say using DEseq.

Gene	P-value
1	0.0002
2	0.0005
3	0.0040
4	0.0060
5	0.0070
6	0.0080
7	0.0090
8	0.0110
9	0.0120
10	0.0120
11	0.01500
12	0.02300
13	0.02400
14	0.03400
15	0.03900
16	0.04700
17	0.05000
18	0.05800
19	0.06000
20	0.09800

(A) (1 pt.) Apply Bonferroni correction and list the genes that would be reported as differentially expressed at alpha = 0.05. Show how you obtain the list.

Genes 1 and 2. Use cutoff 0.05/20 = 0.0025

(B) (1 pt.) List the genes that would be reported as differentially expressed using Benjamini-Hochberg correction at alpha = 0.05. Show how you obtain the list.

Genes 1-14 are significant. Show threshold calculation at least near cutoff.

```
      pvals threshold
 [1,] 0.0002    0.0025 1
 [2,] 0.0005    0.0050 1
 [3,] 0.0040    0.0075 1
 [4,] 0.0060    0.0100 1
 [5,] 0.0070    0.0125 1
 [6,] 0.0080    0.0150 1
 [7,] 0.0090    0.0175 1
 [8,] 0.0110    0.0200 1
 [9,] 0.0120    0.0225 1
[10,] 0.0120    0.0250 1
[11,] 0.0150    0.0275 1
[12,] 0.0230    0.0300 1
[13,] 0.0240    0.0325 1
[14,] 0.0340    0.0350 1
[15,] 0.0390    0.0375 0
```

```
[16,] 0.0470    0.0400 0
[17,] 0.0500    0.0425 0
[18,] 0.0580    0.0450 0
[19,] 0.0600    0.0475 0
[20,] 0.0980    0.0500 0
```

(C) (2 pt.) How do the two lists differ in composition? What does this show about the stringency of these corrections?

Bonferroni results in far fewer significant genes and is more stringent