Homework 1: Due Sep 23th

Sushmita Roy

Instructions

There are five problems in all. Question 1 needs to be submitted via the `adhara.biostat.wisc.edu` machine using the accounts we have created for you. Question 1 is a coding question. Please put your code into `/u/medinfo/bmi576/2014-hw//hw1/username` where username is your user name you use to login to this machine. Note, questions 2-5 are due in class on Sep 23rd and 1 is due on Sep 23rd midnight.

Problem 1 (15 points)

The goal here is to find the sequence at given chromosomal positions. Here `http://www.biostat.wisc.edu/bmi576/2014-hw/hw1/data/yeast_chrom.fasta`, you will find a file of the yeast genome, broken up based on each chromosome. Write a program `getSequence`, which takes as input this file, chromosome, start and end positions and the strand, prints out the sequence. Two example usages and output are given below:

```
./getSequence yeast_chrom.fasta chr01  230000 230010 +
GGGGAATGAGA

./getSequence yeast_chrom.fasta t.txt chr01 230000 230010 -
TCTCATTCCCC
```

Note your program should be able to handle strand information. Thus if the last argument is '-' you should be able to print the reverse complement of the sequence. If the last argument is '+' then simply print the sequence corresponding to those positions. Your program should also handle the error case of incorrect positions, that is, if the specified locations are longer than the length of the chromosome the program should print an error message.

You can use either C++, Java, Perl, or Python to implement this program. If you want to use R, please work with the TA to make sure that the program works on the command line on `adhara.biostat.wisc.edu`. We should be able to run it as of the following commands

```
C++: ./getSequence fastafile  chromosome start end strand
Java: java getSequence fastafile chromosome start end strand
Perl: perl getSequence fastafile chromosome start end strand
```

Problem 2. Global alignment (10 points)

You are given two sequences AGATT and AGTT. Assume a match score of 1, a gap penalty of 3 and a substitution score of -1. Using these scores, obtain the global alignment of these two sequences in the following two steps:

(a) Fill in the entries of the F matrix by applying the recurrence relationship for global alignment to these sequences. Please show the back pointers to the matrix entry/entries that give you the maximal score for any entry.

(b) Apply the trace back procedure to obtain an optimal alignment. If there are multiple possible alignments, please show all of them along with their traceback paths.

Problem 3. Local alignment (10 points)

Using the same scores above perform a local alignment for the sequences, GAAGAG and AAGC in the following two steps:

(a) Fill in the entries of the F matrix using the recurrence relationship for the local alignment of these sequences. Show back pointers to matrix entry/entries that give you the maximal score.

(b) Apply the trace back procedure to generate a local alignment.

Problem 4. Concepts in probability (9 points)

Assume we are interested in the distribution of food items, namely, drinks and snacks ordered at a local grocery store. We are also interested in studying if there is any dependency between the food items purchased and the time of day. We will represent these food items and time of day using the random variables D, S and T. D denotes a drink and takes on values {*coffee*, *water*}. S denotes a snack and takes on values $\{cereal, chips\}$. T denotes the time of day, and takes values {*morning*, *noon*}. You are the given the following set of observations for these events, each row in the table below representing a combination for D, S and T. Using these observations, address the questions (a)-(f) below. Show your work for each question by computing the required probability values.

Drink	*Snack*	*TimeOfDay*
coffee	chips	morning
coffee	cereal	morning
coffee	cereal	morning
coffee	cereal	morning
water	cereal	morning
water	chips	noon
coffee	chips	noon
water	cereal	noon
coffee	cereal	noon

(a) Estimate the probability distribution, $P(S)$

(b) Estimate the joint probability distribution, $P(S, D)$

(c) Estimate the marginal probability distribution of $P(S)$ from the joint, $P(S, D)$.

(d) Is *S* independent of *D*?

(e) Is *S* independent of *D* given $T =$*morning*?

(f) Is *S* independent of *D* given $T =$*noon*?

Problem 5. Applying concepts of probability to sequence alignment (6 points)

Recall that the score of a sequence alignment assess the probability of the alignment from a "related model" and an "unrelated model". The table below gives the counts of two bases aligned at a position in a known and validated sequence alignment. The table is symmetric, that is, the counts for {A,T} are the same as

	A	T	G	C
A	10	20	10	5
T		30	20	10
G			15	10
C				15

counts for {T,A}. Further assume that the counts for individual bases are as follows

A	55
T	110
G	70
C	55

(a) Use the above counts to estimate the probabilities of the related model etc for all pairs of matches and substitutions, as well as the probabilities of the unrelated model.

(b) Assume you are given the following DNA sequence alignment.

```
sequence1: ATA
sequence2: GTC
```

Use the probabilities computed in the (a) to assign a probability to this sequence alignment from the related and unrelated models. Using these probabilities, conclude whether the alignment was more likely to be generated by the related or unrelated models.