Omics Big Data

Spring 2019

Homework 3, week 3

1. Problem 1 (theory)
   Given the single-letter scoring system and sequence shown below:

   A = +2, C = −1, G = −4, T = +2

   TTACTGCGCCTTATAGCTATACGCTGTCGATCTGCGCAATTCCCCCCAATATCCCTCGGTTGATATTAC

   A. What is the maximum segment score?
   B. What are the start and end points of the maximum-scoring segment (MSS)?
   C. Can the A+T composition of the maximum-scoring segment be deduced from its score alone, or from its score and length?
   D. Suggest scoring systems for which the answer to 1C would be different.

2. Problem 2 (theory)
   Using the crude "hydrophobicity" scoring system shown below, if the 20 common amino acid codes are equiprobable in the sequences being searched, are Karlin-Altschul statistics applicable? Why or why not?

   AGILVFHPW = + 3
   CSTMY = 0
   NQ = −1
   RKDE = −2

3. Problem 3 (theory)
   A. What P-values (probabilities) are associated with the E-values (expected frequencies of chance occurrence): 100., 10., 1., 0.1, 0.01?
   B. What trend is observed as E approaches infinity?
   C. What trend is observed as E approaches 0?
      [Note: P = 1 – exp(-E)]

4. Problem 4 (theory)
   A protein sequence of length 600 with typical amino acid composition was compared against the "NR" protein sequence database (70.028 billion amino acids in 192.341 million sequences on March 1, 2019). The scoring matrix was BLOSUM62 and the gap penalties were infinite. BLASTP computed precise values for λ, K and H of 0.304, 0.171, and 0.587, respectively. Clearly state any assumptions you make in answering the following questions.
   [Note: values for λ and H are expressed here in units of "nats", not bits.]
   A. What is the highest alignment score to be expected between our query sequence and the unrelated sequences in the database? [Hint: the expected high score corresponds to the situation where E=1].
   B. How many bits of information are associated with expected high score?
   C. What is the the expected length of the highest scoring alignment between our query and

unrelated sequences?

D. Answer the above questions ABC again, but this time using values for λ, K and H appropriate for a second database search with the same query sequence, but using affine gap penalties u=10 and v=2. [Note: as used here, the penalty for a gap of length n is u+(n-1)v]. There are at least two sources to which you might refer for these values of λ, k and H, including:

   The output from an appropriately parameterized BLASTP run, be it WU-BLASTP or NCBI-BLASTP.

   Table V. in Altchul and Gish, 1996.

   Be sure to indicate how you obtained the values you used.

5. Problem 5 (theory)

   The first search in problem 4 uncovered a few high-scoring segment pairs (HSPs) having only marginal significance against the N-terminal portion of our query sequences. The search was repeated using just the N-terminal 100 residues (1/6th of the length of our original query sequence). This time, the same HSPs were reported, but their statistical significance was 12- to 40-fold higher than before (P-value were 12 to 40-fold lower), making the alignments appear to be somewhat significant.

   From the Karlin-Altschul equation, we expected the significance of the HSPs to increase (their P-values to decrease) when a shorter query sequence was used, but why did the statistical significance increase 12- to 40-fold, when the length of our query sequence was reduced by only 6-fold?

6. Problem 6 (Perl/Python or any programming language)

   This computational experiment is intended to illustrate how the score and length expected for the maximal scoring segment(MSS) vary with the length and residue composition of the input model sequences. We'll also see how the experimentally determined values for the expected score and length affect the values of K and relative entropy H.

   For these experiments, you'll need to write a Perl script that:

   - Generates random model sequences;
   - Finds the MSS in each model sequence;
   - Gathers statistics about the score and length of the MSSes;
   - Gathers statistics on the frequency of occurrence of the letter A in the MSS;
   - Outputs just three numbers: the expected score, expected length and expected frequency of A(q_A) in the MSS.

   The script must be capable of generating model sequences of arbitrary length and residue composition (although for any given experiment these will remain fixed). The process for selecting residues in the sequences should be *i.i.d.* Use Perl's rand function as the random number generator. Using the model sequences so generated, the same script will need to search each sequence for the MSS, noting its score, length and fraction that is A(%A).

The alphabet for the model sequences should consist of the four letters A, C, G and T. You might find it convenient to use instead the digits 0, 1, 2 and 3 to represent these letters in your Perl script. (Maybe a numerical alphabet will even speed up your Perl script. You can test it.) Use any residue composition you wish for the nonuniform distribution, subject to the constraints that

$P_i >= 0.05$ for all i

All P_i differ from the uniform distribution by at least 0.1

The conditions for using Karlin-Altschul statistics still hold.

The scoring system to use in all experiments is A=+1, C=G=T=-1. For all experimental results, have your script perform 10^4 iterations (i.e., collect statistics for MSSes from 10^4 model sequences). After iterating over the 10^4 sequences, have the script report just the expected (arithmetic mean) values for the score, length and %A.

Your fully contained Perl script (random sequence generator, MSS finder, statistics gather and reporter) should utilize 3 command line arguments in this order:

a. The model sequence length;
b. Whether to use a nonuniform distribution (0=> no, non-zero => yes);
c. The number of model sequences to generate and examine (usually 10000).

Note: full credit for this homework will require implementation of an O(N) (i.e., linear in the length of the model sequences) algorithm for finding the MSS.

You are forewarned that, even using an O(N) algorithm, if implemented inefficiently the lengthiest simulations performed here may take several hours to complete. To speed things up a good deal, it is acceptable to capitalize on the specialized conditions of the simulations and the limited output your program is required to produce, to create an efficient script that produces correct results far faster than a more general script. (For instance, no model sequences are to be reported by the program – only the results of computation on these sequences are to be reported – so it is unnecessary to expend time actually storing a randomly generated sequence, only to have to read it again to find the MSS. Why not scan a sequence for the MSS simultaneously with its generation?) In fact, if you program this problem in the C programming language instead of Perl, you might create a much faster program. However, there are no extra points for faster code or code written in C versus Perl, so don't necessarily put your time into optimizing code if you haven't already finished the requirements for answering all the questions. On the other hand, if you wait until the last day to start work on this problem, you may be forced to optimize your code in order to finish all simulations by the due date.

Independently of the experiments, you should populate the table below with values in units of nats computed to 2 decimal digits of accuracy. (Hint: identify an equation that λ solves.) Compute as well the target frequencies for the letter A in the MSS, reporting the value of qA for each distribution.

With the results of your experiments, fill in the table with the expected score, E(S), expected length, E(L), and expected fraction of A residues in the MSS, E(%A), when model sequence length of 10^2, 10^3, and 10^4, letters are used. Report values with 3 digits of precision for E(S), E(L) and E(%A). Using the experimental results, compute values for H (in units of nats) and K to 2 digits of precision. (Hint: solve $E=Knme^{-\lambda S}$ for K in the case of the expected high score.)

Distribution	Uniform			Nonuniform		
λ						
qA						
Model Length, N	10^2	10^3	10^4	10^2	10^3	10^4
E(S)						
E(L)						
E(%A)						
H						
K						

Questions:

(1) What nonuniform distribution did you use?

A=

C=

G=

T=

(2) Which model sequence length yielded values for E(%A) closest to the theoretical target frequency q_A? Why?

(3) Which pair of model sequence lengths yielded results for H and K that are most similar? Why?

(4) For both the uniform and nonuniform distributions and a single model sequence of length 10^4, what is the expected frequency of occurrence of a MSS having a score twice the expected high score?

If you are unable to complete the programming portion of this weeks' exercise, let the instructor know well in advance of the due date.

Turning in your work

Submit your homework to http://cgm.sjtu.edu.cn/test/obd/index.html. Please submit it before 10:00AM on March 20th, 2019. You are also required to submit a hard copy before the class start on March 20th, 2019.

Ask **TA Huimin Lu: linuslu6@outlook.com** in case you have any question about the homework submission webpage.

----------cut-----here----------

独立作业承诺：（请选择一个， 并签名）

1. 本人，_____，保证本次作业由自己独立完成。

   签名

   时间　　年　月　日

或者

2. 本人，_____，保证本次作为和_____同学讨论后，由自己独立完成。
   讨论内容包括_____

   签名
   时间　　年　月　日