CSI5126. Algorithms in bioinformatics
Fall 2017

Assignment

Deadline: October 30, 2017, 18:00

[PDF]

Learning outcomes

- Conceive a linear time algorithm to solve a string problem
- Imagine an exact algorithm counting the number of possible sequence alignments
- Simulate random sequences and analyze the distribution of the scores for the alignment of randomly generated sequences

1 Exact string matching (10 marks)

Given a **linear time** algorithm for the exact string matching problem, outline a linear time algorithm which determines if the string α is a circular permutation of the string β, *i.e.* α and β have the same length, α consists of a suffix of β followed by a prefix of β. For example, the string `defabc` is a circular permutation of `abcdef`.

1. Give a detailed description of the algorithm. For instance, one should be able to write the actual source code given your description.
2. Analyze the time and space requirement of the algorithm.

2 Counting alignments (5 marks)

This question is about the **edit distance problem**, which consists in finding the minimum number of edit operations that are needed to transform one input string into the other. We have seen in class that although there is an exponential number of alignments, the problem can be efficiently solved in quadratic time and space using dynamic programming. Write a program, involving a dynamic programming solution, **to count the number of possible alignments for two given input sequences of length m and n respectively**.

1. This involves developing a recurrence equation that solves this problem.
2. As well as its implementation.

The input for your program is two numbers, say 280 and 240, which are interpreted as the length of two sequences. The output is the number of possible sequence alignments for two sequences of such lengths. **Note:** Make sure that your implementation works for reasonable sizes of sequences, say 280 and 240.

3 Alignment of random sequences (10 marks)

1. Write a computer program, involving a dynamic programming solution, to compute the **local alignment** (Smith-Waterman) of two DNA sequences of length m and n.
2. Generate 1,000 pairs of random DNA sequences, of length m and n, and plot the distribution of the scores.
3. Does it look like any distribution that you know of?

You can assume that all four nucleotides are equiprobable, i.e. $p_A = p_C = p_G = p_T = 0.25$, and that $m = 280$ and $n = 240$. Your program must find the maximum similarity score, for a local alignment, using the following transition/transversion substitution matrix, and -3 for the cost of each indel.

	A	C	G	T
A	1	-5	-1	-5
C	-5	1	-5	-1
G	-1	-5	1	-5
T	-5	-1	-5	1

Submit your program and plot along with your answer to this question.

4 Testing the significance of an alignment (5 marks)

Cases of food poisoning are unfortunately quite common — "Two people are paralyzed after drinking botulism-contaminated carrot juice (...)" from Botulism-tainted juice paralyzes two in Canada, www.reuters.com, 2006/10/17 — we have been asked to help the investigators.

Background information. Botulism is a rare disease caused by a toxin (a protein) expressed by the bacterium *Clostridium botulinum*. From the above source of information: "Botulism can cause nausea, fatigue, double-vision, paralysis and respiratory failure. In severe cases, the toxin can be fatal."

The investigators suspect that at least one of the following proteins, Unknown A or Unknown B, has a similar domain to that of *Clostridium botulinum*'s toxin.

1. Which protein(s) is (are) related to this toxin?
2. Provide a detailed description of the methodology that you have used to arrive at your conclusions.

```
>gi|27867582 (fragment of the known Clostridium botuninum toxin gene)
GTGAATCAGCACCTGGACTTTCAGATGAAAAATTAAATTTAACTATCCAAAATGATGCTT
ATATACCAAAATATGATTCTAATGGAACAAGTGATATAGAACAACATGATGTTAATGAAC
TTAATGTATTTTTCTATTTAGATGCACAGAAAGTGCCCGAAGGTGAAAATAATGTCAATC
TCACCTCTTCAATTGATACAGCATTATTAGAACAACCTAAAATATATACATTTTTTTCAT
CAGAATTTATTAATAATGTCAATAAACCTGTGCAAGCAGC

> Unknown A
TCTATCAAGTAGATTATTAAATACTACTGCACAAAATGATTCTTACGTTCCAAAATATGA
TTCTAATGGTACAAGTGAAATAAAGGAATATACTGTTGATAAACTAAATGTATTTTTCTA
TTTATATGCACAAAAAGCTCCTGAAGGTGAAAGTGCTATAAGTTTAACTTCTTCAGTTAA
TACAGCATTATTAGATGCATCTAAAGTTTATACGTTTTTTTCTTCAGATTTTATTAATAC

> Unknown B
TCCTGGCTCAGGACGAACGCTGGCGGCGTGTGCTTAACACATGCAAGTCGAGCGATGAAG
CTTCCTTCGGGAAGTGGATTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTGCCTCA
AAGTGGGGGATAGCCTTCCGAAAGGAAGATTAATACCGCATAACATAAGAGAATCGCATG
ATTTTCTTATCAAAGATTTATTGCTTTGAGATGGACCCGCGGCGCATTAGCTAGTTGGTA
```

Directives

- You must do this assignment individually.
- Assignments must be submitted by E-mail.
- Each program must be documented.
- For the programming questions, I should be able to run your programs without downloading additional libraries.
- Clearly identify yourself on every file that you hand in; this includes your name and student id.

A Frequently Asked Questions [FAQ]

1. **"Give us an example of input and output for the program that counts the number of alignments."**

   The input can be as simple as two numbers, m and n, which represent the length of the first and second sequence.

   ```
   > java CountAlignments 10 20
   4,354,393,801
   ```

2. **"What programming language can I use for this assignment?"**

   I should be able to read a wide variety of programming language. I have extensive experience with procedural, object-oriented, logic, and functional programming languages. However, I would like to be able to run your program on my computer and I am using macOS High Sierra. In doubt, E-mail me.

Modified October 12, 2017