BIL 105E – Introduction to Scientific and Engineering Computing

Spring 2016-17

Assignment 2: Disease Detection

Posting Date: 03.03.2017
Due Date and Time: 12.03.2017 – 23.59

Prof. Bad has inserted random characters into a DNA sequence represented by four letters (adenine (A), cytosine (C), guanine(G), thymine(T)) to create a corrupted DNA sequence. He handed this new sequence over to his assistant Dr. Good to check the order of nucleotides whether the original sequence indicates a disease or not. Dr. Good should pre-process the sequence to produce the original sequence and check the existence of the disease. Nucleotides are grouped in pairs: A, T and C, G.A healthy DNA sequence consists of nucleotides A coupled with T and G coupled with C. The order of the appearance of these pairs is not important and pairs may be interlaced with each other. Repeated nucleotides have to be ignored before the expected nucleotide appears in the sequence. Check Table 1 for examples.

Write a C program to help Dr. Good to detect a disease through DNA analysis. Dr. Good should enter a corrupted DNA sequence of characters as an input. It should be pre-processed to produce the DNA sequence to detect the disease, and to produce an output to prompt Dr. Good about the outcome.

Table 1: Healthiness for a given DNA sequence

DNA Sequence	Status	Pairs (indices)
	No Data	`None`
`T`	Not Healthy	`0 not paired`
`ACGTACGTT`	Not Healthy	`0-3,1-2,4-7,5-6, 8 not paired`
`GATCACGGTC`	Healthy	`0-3,1-2,4-8,5-6,7-9`
`GACGCT`	Healthy	`0-2,1-5,3-4`
`TATGCC`	Not Healthy	`0-1,3-4, 2 and 5 not paired`
`AGTCATGACT`	Healthy	`0-2,1-3,4-5,6-8,7-9`

Some of the essential operations to implement:

a) Input must be entered after the user is prompted with the text: "**`Enter a seq.:`**" and output must be printed after the user is prompted with the text: "**`DNA seq.:`**". Note that these texts to be prompted have to be exactly the same in your program for automatic testing purposes. (Example: Sample run at the end of this text.)

b) Although there is no upper limit for the size of the input, only the first 25 characters will be processed. (Example: Table 2, line1)

c) Input string will be evaluated/checked character by character. Thus, input string will not be stored in the memory.

d) Filter out any character in the input sequence except for 'A','C','G','T','a','c','g', and 't' and convert all of the accepted characters into upper case.(Example: Table 2, lines 1, 2, 3, 4)

e) A pre-processed DNA sequence may have at most 10 characters representing nucleotides. The evaluation of the DNA sequence will start immediately once a sequence 10 charactersis produced. (Example: Table 2, line4)

f) Check whether all nucleotides are paired with regard to the healthiness definition given in the first paragraph of the assignment text.

g) "**`Disease: `*x*` nucleotide.`**" or "**`Disease: `*x*` and `*y*` nucleotides.`**" should be printed on screen on which ***x*** and ***y*** represent the nucleotides that the program fails to find a pair. "**`Healthy!`**" should be printed if all pairs are matched."**`No data to test!`**" should be printed if DNA sequence has no elements. Note that the output text to be printed on screen has to be exactly the same in your program for automatic testing purposes.(Example: Sample run at the end of this text.)

h) User should enter **`X`** to terminate. **`X`** may also be succeeded by other characters. Therefore upon entering "**`XATGC`**" or "**`xAsfl123`**" as input, the program will terminate after prompting the user the message "**`Terminated!`**"

Table 2: Sample inputs, pre-processed DNA sequence and Operations

Corrupted Sequence	Pre-processed DNA Sequence	Operations
`Qwertyuiopasdfghjklzcvbnm1234567890ATCG`	**`TAGC`**	**`Filter out, convert to upper case, finish at 25th character on the corrupted sequence`**
`Ac2gtQwErA5CGT`	**`ACGTACGT`**	**`Filter out, convert to upper case`**
`GaaCGaCaYt`	**`GACGCT`**	**`Filter out, convert to upper case, ignore repitation`**
`AGTCAAAAAATQWEGAAACTATGCATGC`	**`AGTCATGACT`**	**`Filter out, ignore repitation, finish at 10th nucleotide`**

Important:

Submissions have to be completed and uploaded in time through Ninova pages. No other ways of submission will be accepted.

Note that all submissions will be compiled on ssh.itu.edu.tr machine using the gcc compiler for evaluation. Make sure that your code to submit can be compiled on that machine.

You are expected to utilise the lecture contents taught until the posting data of this assignment. Therefore you are not expected to use advanced topics such as pointers, strings, arrays.

Do not use **`scanf(…)`** function.

Use a **`for`** statement at least for once while implementing the above mentioned essential operations.

If you have not fulfilled any of the above mentioned issues, you will be considered as "not submitted" for this assignment.

All the programs will be tested automatically, the input and output messages of your program have to be exactly **the same** with the messages provide in the sample program.

Plagiarism:

All the assignments are considered individual assignments and you are expected to do it by yourself. Any form of plagiarism, even partial, will not be tolerated. It is subject to serious disciplinary actions. Note that professional help in any form or shape is considered as an act of plagiarism.

A Sample Run:(User inputs are underlined.)

```
Enter a seq.:  Ac2gtQwErA5CGT
DNA seq.:      ACGTACGT
Healthy!
Enter a seq.:  acgtacg
DNA seq.:      ACGTACG
Disease: A nucleotide.
Enter a seq.:  GaaCGaCaYt
DNA seq.:      GACGCT
Healthy!
Enter a seq.:  uvuvwevwevwe onyewenyevwe ugwemubwem ossas
DNA seq.:
No data to test!
Enter a seq.:  AAAAAAAAAAAAATGttttttttttttTTTTCCCC
DNA seq.:      ATGT
Disease: G and T nucleotides.
Enter a seq.:  c
DNA seq.:      C
Disease: C nucleotide.
Enter a seq.:
DNA seq.:
No data to test!
Enter a seq.:  AAAAAAAAAAAAAAAAAAAAAAAATTTTTTTTTTTTTTTTTTTT
DNA seq.:      A
Disease: A nucleotide.
Enter a seq.:  AAAAAAAAAAAAAAAAAAAAAAAATTTTTTTTTTTTTTTTTTTT
DNA seq.:      AT
Healthy!
Enter a seq.:  AAAAAAAAAAAAAAAAAAAAAAATTTTTTTTTTTTTTTTTTTT
DNA seq.:      ATT
Disease: T nucleotide.
Enter a seq.:  AGxCT
DNA seq.:      AG
Terminated!
Enter a seq.:  Qwertyuiopasdfghjklzcvbnm1234567890ATCG
DNA seq.:      TAGC
Healthy!
Enter a seq.:  XGT
DNA seq.:
Terminated!
```