Bioinformatics Mini Project

Aim: Motif Finding

Code:

# Motif discovery is one of the sequence analysis

# problems under the application layer and it is one

# of the significant difficulties in bioinformatics applications.

# A DNA sequence motif is a subsequence of DNA sequence that is a

# short similar recurring pattern of nucleotides, and it has many

# biological functions

# motif finding

import random

from Bio import Seq

from Bio import SeqUtils

I = int(input("enter the length of motif:-"))

file = open("h1n1.txt", "r")

r = file.read()

print("sequence:\n", r)

size = len(r)

print("size of the sequence:", size)

pos = random.randint(0, len(r)-I)

print("position", pos)

motif = r[pos:pos+I]

print("motif", motif)

results = SeqUtils.nt\_search(str(r), motif)

print("Match motif:", results)

i = pos+1

while(i < size-1):

    if(motif == r[i:i+I]):

        str1 = r[i:i+I]

        print("match motif:", str1)

        file1 = open("motoutput.txt", "a")

        file1.write(str1+"")

    i += 1

enter the length of motif:-4

sequence:

GACAACTTCCAACTTCCAACTTCCCGTCCCAACTTCACAACTTCGGCCCAACTTCCATGCAACTTCACCATCAACTTCGCTCGAAGCTGCCTTCCACTCCAACTTCACAACTTCCTCAACTTCCTCACCAACTTCAGCAACTTCTCTAGGGCCAACTTCCAACTTCTCAACTTCTCAACTTCCAACTTCCGACAACTTCTCCTGGCAACTTCCAACTTCCAACTTCAATACAACTTCGCAGACAACTTCCGCAACTTCGAACAACTTCCAACTTCCCCAACTTCCAACTTCCAACTTCGCCAACTTCCAACTTCCAACTTCCCAACTTCAGATAGCAACTTCGATCTTACACAACTTCACGCAACTTCTCCAACTTCCAACTTCTGTGCAACTTCTCTGAACAACTTCCTCAACTTCCAACTTCGCAACTTCCCCAACTTCCTCAACTTCATGCAACTTCGAGGCAACTTCCCAACTTCGCAACTTCCTATTCCCAACTTCTGTGGCAACTTCTCAACTTCTGGACAACTTCTATGCCCAACTTCACAACTTCCCCAACTTCTTTACAACTTCGACAACTTCATCAACTTCTAGTCAACTTCTGGTCCAACTTCCAACTTCCCCAACTTCCAAAGTGCCGCAACTTCGTAACAACTTCACGCGCTCAACTTCAACCAACTTCTTTTCCCGCAACTTCGCAACTTCACAACTTCTAATCAACTTCCAACTTCGGATCAACTTCCAACTTCGCCAACTTCCAACTTCCAACTTCTCCAGGGACAACTTCAAGTACAACTTCCAACTTCGCAACTTCACAACTTCCCAACTTCGCAACTTCTACACGCAACTTCCAACTTCTGGTCCCAACTTCATCAACTTCAGTCAACTTC

CAACTTCCA

ATTGTGTATCGCTATGTATCGTGTATCGGATTTTGTATCGTGTATCGGTGTATCGTGTATCGTGTATCGTATCACTGTATCGTTGTATCGTAGCGTTGTATCGTAATGTATCGCTCTGTATCGTGTATCGGGTTTGTATCGATGTGTATCGCTGTATCGGTGTATCGGTGTATCGCTTGTATCGACAGCGCTTGTATCGTGTATCGACCTGTATCGGTGTATCGTGTATCGAATGTATCGTTGTATCGAATTGTGTATCGTGTATCGTGTATCGTGTATCGTATGTATCGACTGTATCGCTGTATCGTAGCCTGTATCGTGTATCGGTGTATCGGTGTATCGGCGTGTATCGAATGTATCGTTGTATCGCTGTATCGTGTGCTGTGTATCGGTGTATCGACCCGTTGTATCGTGTATCGTGGGTAAGTGTATCGTTGTATCGTAGTGTATCGTGTATCGTGTATCGTCATGGTATTGTATCGTTATAGCTGTATCGCTCGGCTGTATCGTATGTATCGTGTATCGTGTATCGCATGTATCGGTGTATCGCTGTATCGACCATGTATCGTGTATCGGTGTATCGATGTATCGCTTCCATTAGAAATGTATCGTATGTATCGTGTATCGCTGTATCGTTTGTATCGCATGTATCGATTGTGTATCGTGTATCGTTGTATCGTGTATCGTGTATCGTTGTATCGTAATAACGATGTATCGAATGATGTATCGTTGTGATGTATCGTGTATCGACTGTATCGATGTATCGGGTTGTATCGTGTATCGTGTATCGTAGATGTATCGAGAGCCATTGTATCGCCTTGTATCGCGTGTATCGGTGTGTATCGTTTGTATCGTGTTTGAATGTATCGCTGTATCG

TGTATCGTG

TTTCAGCGTGTTTGGCTTTCAGCACTACTTTCAGCGGTTTCAGCTTTTTCAGCTTTTCAGCTATTTTCAGCTTTCAGCCATCTTTCAGCCCAGGTTTCAGCTGTTTCAGCTTTCAGCGAGTTTCAGCTTTCAGCGCGGGGATTTCAGCGTATTTCAGCGACATTTCAGCAGTTTCAGCTGGAGTGAAGCCGTTTCAGCGCTTTTCAGCACCATTTCAGCTAGTAGGTTTCAGCTTTTCAGCCTTTTCAGCTTTCAGCTATTTCAGCTCTAGAAGTCGTTTCAGCTTTCAGCTTCCTTTTCAGCTTTCAGCTTTCAGCCTTTCAGCTTTCAGCAAACTTTCAGCGATTTCAGCTTTCAGCTTTCAGCTTTCAGCTTTCAGCATTTCAGCTTTCAGCCGTTTTCAGCGACTCTTTCAGCGTTTCAGCTTTCAGCTGATCGTTTTCAGCTTTTCAGCCGGTTTCAGCGAAAGTTGGTCTTTTCAGCCAAATTTTCAGCTTTCAGCTTTCAGCGTTTCAGCTTTTCAGCCATTTCAGCCTATTTTCAGCAATCTTTCAGCAATTTTCAGCGCAATTTCAGCCAAAATTTCAGCATTTCAGCTTTCAGCGTAGTTTCAGCTTTCAGCTGAGCCTTTCAGCTTTTCAGCTCTCTTTTCAGCTTTCAGCGATTTCAGCCTTTTCAGCGCTTTCAGCACCACGCCTTTCAGCTTTCAGCTTTTCAGCCTTTCAGCCTATTTTCAGCGGTTTCAGCGTTTCAGCTTTCAGCACTTTCAGCTTTCAGCTTTTCAGCGGAAGGTTTTCAGCA

TTTCAGCTT

AGCATCCGACTGCATCCGTGCATCCGACGGCATCCGGCATCCGCAGCATCCGCACAGCCGCATCCGAGCCCGCATCCGAAGCATCCGGCATCCGGGCATCCGGCCGCATCCGGCATCCGGGCATCCGTGCATCCGGCATCCGTGTGCATCCGGCATCCGGGGTGCTTGCATCCGCGCATCCGTGCATCCGCGCATCCGGCATCCGCAGTCCGCATCCGCGCATCCGGCATCCGAGCATCCGTATTCGCATCCGTCGAGCATCCGTACGTCGCATCCGTGCATCCGGGAGAGGCATCCGGGCATCCGCGCATCCGGTGGACTATAACGCGCATCCGGCGAGCATCCGCCGCGGCATCCGTGCATCCGGTACGGCATCCGGGCTGAGGCATCCGCTCCCGGGGCATCCGCGGCGCATCCGCACGCGCATCCGGCATCCGAGCATCCGTTGCATCCGGAGCATCCGTAGCATCCGATAGCATCCGCAAGCTTGCATCCGGCATCCGCTACAGCATCCGGTGGCATCCGAAGCATCCGCGCATCCGCGGATTGCATCCGCACACGGCATCCGTAGCATCCGTGCATCCGGAGTGCATCCGCCGGGCGCATCCGCTGCATCCGCAGCGCATCCGAGCATCCGATCTTGCATCCGGCATCCGGGAATGCATCCGGGACTGCATCCGGTCTTAAGGGTGCATCCGAATGCATCCGCTGTAGCATCCGGCATCCGAGTAAGCATCCGAGTTCTGCATCCGGAAGCAGCATCCGGCATCCGGACACCAGCATCCGCGCATCCGGGGCGAGCATCCGAGCATCCGGCATCCGGGGCAAGTGGCATCCGGCATCCGTCGTGCATCCGGGCATCCGAA

GCATCCGGC

GACGAGCTGACGAGCGTGACGAGCAGGCGACGAGCGACGAGCATGCGGTGACGAGCGGACGAGCGACGACGAGCGCGGACGAGCGACGAGCGGACGAGCTCCGACGAGCGACGAGCACGACGAGCGCCCGACGAGCGACGAGCATGATGACGAGCGACGAGCTTAGACGAGCATGCTGACGAGCTGACGAGCTGACGAGCTCGTACATCGACGAGCGGACGAGCAGCCCGATAAGCCTTCGACGAGCTGACGAGCCACGACGAGCGACGAGCGACGAGCGGACGAGCGGGACGAGCGTCGGACGAGCGGACGAGCAGGACGAGCACCTCAATCGACGAGCTGACGAGCGGACGAGCCGACGAGCTGAACTGAAGGACGAGCCTACGTTCTAACGTGCCGTCACTGACGAGCGACGAGCGGACGAGCGATGGACGAGCTGACGAGCGCAGGACGAGCGAGAAGGCTGACGAGCAGACGAGCTGACGAGCTCGACGAGCAGACGAGCGGACGAGCTTTGATAAGACGAGCACGTCGACGAGCCTCGTGACGAGCTGACGAGCGACGAGCTAGACGAGCGACGAGCTGACGAGCGACGAGCACTAAACGCGACGAGCTCGACGACGAGCATGGACGAGCGACGAGCTAGGACGAGCGGCGACGAGCAGACGAGCGACGAGCTGACGAGCTTATAGACGAGCCTGACGACGAGCCAAGACGAGCGACGAGCGGACGAGCGACGAGCACGACGAGCCGACGAGCCGCGGACGAGCCGTAGACGAGCCAATCATTAAGACGAGCCGACGAGCCATTTGGGGACGAGCCTCCTCGACGAGCCATAAGACGAGCTGACGAGCCATGACGAGCATGCCCGACGAGC

GACGAGCGA

GAGGACCCCGGACCCCTGGACCCCAGGACCCCGACGGTGGGGCGGACCCCGAGAAGGACCCCTGGACCCCCGCAGGACCCCTTTATCGGACCCCGGACCCCGGACCCCCGGACCCCGGCTGGACCCCGTCGTAAGGACCCCGAATCGGACCCCTAGGACCCCAGGGACCCCTCCCCGGTTGGACCCCCTGGACCCCCTTGAGAGGGACCCCGGACCCCACGCCGTGCTTAAGGACCCCACTATGGACCCCATACGGACCCCGGACCCCGATCAGAGCGACCAGGACCCCCGGCCTGGACCCCTCGGACCCCAGAAGGACCCCTCGGACCCCTGGACCCCGGACCCCACAGGGACCCCGGACCCCTAGCCGCGGTGGACCCCCGGACCCCCGCAGATGGGGACCCCTCATGCGGACCCCCTGACGGACCCCTGGACCCCCGGGACCCCCAGGGACCCCATTTCGAGGACCCCATGGGGACCCCAAGCTGGACCCCGGACCCCTGGGACCCCCGGACCCCGGACCCCGGACCCCATAGGACCCCCGTTTGTTGCCATGGACCCCTTGGGACCCCGAGTCGGACCCCGGACCCCCAGGACCCCACAGGACCCCGGGACCCCGGACCCCATCGATCGGACCCCGGACCCCGGACCCCTACGGACCCCGCTAAGGGACCCCGGACCCCGTGGACCCCCACGGACCCCTAGGACCCCGGGACCCCGGACCCCGGACCCCAGTTGGACCCCCCTTAGGACCCCAGGACCCCACATGAGGACCCCGGGACCCCGGGACCCCTGGACCCCTGGACCCC

GGACCCCGG

AAACCTTCGGACCTTCGTACCTTCGACCTTCGTGAAACCTTCGGACCTTCGTGACCTTCGGGTGACCTTCGGGTACTCACCTTCGGCACCTTCGTATACCTTCGAGACCTTCGCTGGACTGTAAACCTTCGATCGTTATTTTCTACCTTCGCACACCTTCGACCTTCGAGCCGAACCTTCGGACTGGCCTCACCTTCGATCACCTTCGACCTTCGAACCTTCGACCTTCGACCATACCTTCGACCTTCGGACCTTCGGACCTTCGCACCTTCGAACCTTCGCCACCTTCGACCTTCGAACCTTCGGTGGGCCACCTTCGTTACCTTCGAACCTTCGGGCACCTTCGACCTTCGGCTTACACCTTCGGACCTTCGATGACTGACCTTCGAAACCTTCGACCTTCGACCTTCGATTCCACACCTTCGACCTTCGACACCTTCGACCTTCGCGTTCACCTTCGACCTTCGACCTTCGGGCCAATACCTTCGACCTTCGGCCTAATACCAACCTTCGAACCTTCGAACCTTCGGGTGACCTTCGGACCTTCGCGCACACCTTCGCACTTACCTTCGACCTTCGGTGACCTTCGACCTTCGTGCACCTTCGCGTCTCACCTTCGCACCTTCGTTACCTTCGAACCTTCGACCTTCGACAACCTTCGTACCTTCGCTACCTTCGGGAACCTTCGATGTCCTACCTTCGACCTTCGCGACCTTCGACCTTCGTAACCTTCGAACCTTCGACCACCTTCGAACCTTCGAAGGAAGGACCTTCGAATCATTACCTTCGTTACCTTCGTAGTACCTTCGCCGACCTTCGGACCTTAACCTTCGTGA

ACCTTCGAC

AAGTGGTATTCCCAAAAAGATATTCCCGTTATTCCCGCTATAAATATTCCCAACTGTTTATTCCCACTTATTCCCGTATTCCCGTATTCCCCGGGCCTTTTATTCCCCATTCTATTCCCGTATTCCCATATTCCCACTTATTCCCCTATTCCCATATTCCCTATTCCCTATTCCCGTAAGTATTCCCCCCCTTATTCCCCCATTTGTCCATATTCCCTCAATATTCCCTATTCCCCTATTCCCGCTTCCCCAGTCTATCCCAAATTGTATTCCCAATATTCCCCTTTATTCCCTTTATTCCCGTATTCCCTATTCCCGTGGCTATTCCCCCGACGTTATTCCCTCTATATTCCCGTCTCAAAGTATTCCCTTATTCCCAGTATTCCCATTCGCCTGGTATTCCCGTATTCCCGGTATTCCCGTTTATTCCCATCTATTCCCATTATTCCCTGGTATTCCCGCTATTCCCGCAACGGTATTCCCTATTCCCTATTCCCGTATTCCCAATATTCCCTTCCTATTCCCGTATTCCCTAGCTATTCCCATATTCCCGATATTCCCTATTCCCGTATTCCCTATTCCCTTATTCCCTATTCCCTAGTTATTCCCAAACAGGCTATTCCCTATTCCCTATTCCCATATTCCCTTATTCCCTACGTATTCCCTCTTAAAAATATTATTCCCCTATTCCCTATTCCCGTATTCCCCGAAAGTATTCCCCCATATTCCCAATTATATTCCCGGTATTCCCCCGTATTCCCTATTCCCTTATTCCCATATTCCCATATTCCCAAAGTTTATTCCCGTATTCCCCACTATTCCCATAGATTATTCCCTATTCCCTATTCCCTTATTCCCCTATTCCCTATTCCCTTATTCCCGTATTCCCAGTTTATTCCC

TATTCCCTA

size of the sequence: 6958

position 3758

motif GACG

Match motif: ['GACG', 2632, 3486, 3494, 3503, 3514, 3521, 3535, 3543, 3550, 3553, 3563, 3570, 3578, 3588, 3595, 3604, 3615, 3622, 3634, 3641, 3651, 3663, 3671, 3679, 3695, 3703, 3726, 3734, 3744, 3751, 3758, 3766, 3775, 3786, 3794, 3803, 3819, 3827, 3835, 3843, 3860, 3890, 3897, 3905, 3916, 3924, 3935, 3951, 3959, 3967, 3976, 3984, 3992, 4007, 4019, 4031, 4039, 4046, 4055, 4062, 4070, 4077, 4093, 4102, 4105, 4115, 4122, 4132, 4142, 4150, 4157, 4165, 4177, 4186, 4189, 4199, 4206, 4214, 4221, 4230, 4238, 4249, 4260, 4277, 4285, 4300, 4313, 4325, 4333, 4343, 4356, 4365, 4408, 4796, 6378]

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

match motif: GACG

PS E:\bioinformatics>