**Feedback — Programming Homework 4 Please Note: No Grace Period**

**Question 1**

In a practical, we saw the scs function (copied below along with overlap) for finding the shortest common superstring of a set of strings.

In [1]:

**def** overlap(a, b, min\_length=3):

*""" Return length of longest suffix of 'a' matching*

*a prefix of 'b' that is at least 'min\_length'*

*characters long. If no such overlap exists,*

*return 0. """*

start = 0 *# start all the way at the left*

**while** **True**:

start = a.find(b[:min\_length], start) *# look for b's suffx in a*

**if** start == -1: *# no more occurrences to right*

**return** 0

*# found occurrence; check for full suffix/prefix match*

**if** b.startswith(a[start:]):

**return** len(a)-start

start += 1 *# move just past previous match*

In [2]:

**import** **itertools**

**def** scs(ss):

*""" Returns shortest common superstring of given*

*strings, which must be the same length """*

shortest\_sup = **None**

**for** ssperm **in** itertools.permutations(ss):

sup = ssperm[0] *# superstring starts as first string*

**for** i **in** range(len(ss)-1):

*# overlap adjacent strings A and B in the permutation*

olen = overlap(ssperm[i], ssperm[i+1], min\_length=1)

*# add non-overlapping portion of B to superstring*

sup += ssperm[i+1][olen:]

**if** shortest\_sup **is** **None** **or** len(sup) < len(shortest\_sup):

shortest\_sup = sup *# found shorter superstring*

**return** shortest\_sup *# return shortest*

It's possible for there to be multiple different shortest common superstrings for the same set of input strings. Consider the input strings ABC, BCA, CAB. One shortest common superstring is ABCAB but another is BCABC and another is CABCA.

What is the length of the shortest common superstring of the following strings?

CCT, CTT, TGC, TGG, GAT, ATT

In [3]:

input\_strings = ['CCT', 'CTT', 'TGC', 'TGG', 'GAT', 'ATT']

shortest\_common\_super\_string = scs(input\_strings)

print(shortest\_common\_super\_string, len(shortest\_common\_super\_string))

CCTTGGATTGC 11

**Answer:**

11

**Question 2**

How many different shortest common superstrings are there for the input strings given in the previous question?

Hint 1: You can modify the scs function to keep track of this.

Hint 2: You can look at these examples to double-check that your modified scs is working as expected.

In [4]:

**def** scs\_all(ss):

*""" Returns shortest common superstring of given*

*strings, which must be the same length """*

shortest\_sup = **None**

**for** ssperm **in** itertools.permutations(ss):

sup = ssperm[0] *# superstring starts as first string*

**for** i **in** range(len(ss)-1):

*# overlap adjacent strings A and B in the permutation*

olen = overlap(ssperm[i], ssperm[i+1], min\_length=1)

*# add non-overlapping portion of B to superstring*

sup += ssperm[i+1][olen:]

**if** shortest\_sup **is** **None** **or** len(sup) < len(shortest\_sup):

shortest\_sup = sup *# found shorter superstring*

shortest\_len = len(shortest\_sup)

scs = set()

**for** ssperm **in** itertools.permutations(ss):

sup = ssperm[0] *# superstring starts as first string*

**for** i **in** range(len(ss)-1):

*# overlap adjacent strings A and B in the permutation*

olen = overlap(ssperm[i], ssperm[i+1], min\_length=1)

*# add non-overlapping portion of B to superstring*

sup += ssperm[i+1][olen:]

**if** len(sup) <= shortest\_len:

scs.add(sup)

**return** scs *# return shortest list*

In [5]:

input\_strings = ['CCT', 'CTT', 'TGC', 'TGG', 'GAT', 'ATT']

scs\_list = scs\_all(input\_strings)

len(scs\_list)

Out[5]:

4

**Answer:**

4

**Question 3**

Download this FASTQ file containing synthetic sequencing reads from a mystery virus:

<https://d28rh4a8wq0iu5.cloudfront.net/ads1/data/ads1_week4_reads.fq>

All the reads are the same length (100 bases) and are exact copies of substrings from the forward strand of the virus genome. You don't have to worry about sequencing errors, ploidy, or reads coming from the reverse strand.

Assemble these reads using one of the approaches discussed, such as greedy shortest common superstring. Since there are many reads, you might consider ways to make the algorithm faster, such as the one discussed in the programming assignment in the previous module.

How many As are there in the full, assembled genome?

Hint: the virus genome you are assembling is exactly 15,894 bases long

In [6]:

**def** readFASTQ(filename):

f = open(filename)

reads = []

**while** **True**:

h = f.readline()

**if** len(h) == 0:

**break**

sequence = f.readline().rstrip()

\_ = f.readline()

quality = f.readline().rstrip()

reads.append(sequence)

**return** reads

In [7]:

file = 'ads1\_week4\_reads.fq'

reads = readFASTQ(file)

In [8]:

**def** overlap(a, b, min\_length=3):

*""" Return length of longest suffix of 'a' matching*

*a prefix of 'b' that is at least 'min\_length'*

*characters long. If no such overlap exists,*

*return 0. """*

start = 0 *# start all the way at the left*

**while** **True**:

start = a.find(b[:min\_length], start) *# look for b's suffx in a*

**if** start == -1: *# no more occurrences to right*

**return** 0

*# found occurrence; check for full suffix/prefix match*

**if** b.startswith(a[start:]):

**return** len(a)-start

start += 1 *# move just past previous match*

In [9]:

**def** pick\_maximal\_overlap(reads, k):

*""" Return a pair of reads from the list with a*

*maximal suffix/prefix overlap >= k. Returns*

*overlap length 0 if there are no such overlaps."""*

reada, readb = **None**, **None**

best\_olen = 0

kmer\_dict = {}

**for** read **in** reads:

**for** i **in** range(len(read)-k+1):

kmer\_dict[read[i:i+k]] = set()

**for** read **in** reads:

**for** i **in** range(len(read)-k+1):

kmer\_dict[read[i:i+k]].add(read)

**for** read **in** reads:

current\_kmer\_set = kmer\_dict[read[-1\*k:]]

**for** kmer\_read **in** current\_kmer\_set:

**if** read != kmer\_read:

olen = overlap(read, kmer\_read, min\_length=k)

**if** olen > best\_olen:

reada, readb = read, kmer\_read

best\_olen = olen

**return** reada, readb, best\_olen

In [10]:

**def** greedy\_scs(reads, k):

*""" Greedy shortest-common-superstring merge.*

*Repeat until no edges (overlaps of length >= k)*

*remain. """*

read\_a, read\_b, olen = pick\_maximal\_overlap(reads, k)

**while** olen > 0:

reads.remove(read\_a)

reads.remove(read\_b)

reads.append(read\_a + read\_b[olen:])

read\_a, read\_b, olen = pick\_maximal\_overlap(reads, k)

**return** ''.join(reads)

In [11]:

%%**time**

ss = greedy\_scs(reads, 10)

Wall time: 3min 29s

In [12]:

len(ss)

Out[12]:

15894

In [14]:

i = 0

**for** s **in** ss:

**if** s == 'A':

i +=1

print('As: ', i)

As: 4633

**Answer:**

4633

**Question 4**

How many Ts are there in the full, assembled genome from the previous question?

In [15]:

j = 0

**for** s **in** ss:

**if** s == 'T':

j += 1

print('Ts: ', j)

Ts: 3723

**Answer:**

3723

**Question 5**

This question has no point value. It's just for fun!

What disease is this virus associated with? Give just the name of the disease (one word).

Hint: you might want to use the BLAST database.

In [16]:

ss

Out[16]:

'ACCAAACAAAGTTGGGTAAGGATAGATCAATCAATGATCATATTCTAGTACACTTAGGATTCAAGATCCTATTATCAGGGACAAGAGCAGGATTAGGGATATCCGAGATGGCCACACTTTTGAGGAGCTTAGCATTGTTCAAAAGAAACAAGGACAAACCACCCATTACATCAGGATCCGGTGGAGCCATCAGAGGAATCAAACACATTATTATAGTACCAATTCCTGGAGATTCCTCAATTACCACTCGATCCAGACTACTGGACCGGTTGGTCAGGTTAATTGGAAACCCGGATGTGAGCGGGCCCAAACTAACAGGGGCACTAATAGGTATATTATCCTTATTTGTGGAGTCTCCAGGTCAATTGATTCAGAGGATCACCGATGACCCTGACGTTAGCATCAGGCTGTTAGAGGTTGTTCAGAGTGACCAGTCACAATCTGGCCTTACCTTCGCATCAAGAGGTACCAACATGGAGGATGAGGCGGACCAATACTTTTCACATGATGATCCAAGCAGTAGTGATCAATCCAGGTCCGGATGGTTCGAGAACAAGGAAATCTCAGATATTGAAGTGCAAGACCCTGAGGGATTCAACATGATTCTGGGTACCATTCTAGCCCAGATCTGGGTCTTGCTCGCAAAGGCGGTTACGGCCCCAGACACGGCAGCTGATTCGGAGCTAAGAAGGTGGATAAAGTACACCCAACAAAGAAGGGTAGTTGGTGAATTTAGATTGGAGAGAAAATGGTTGGATGTGGTGAGGAACAGGATTGCCGAGGACCTCTCTTTACGCCGATTCATGGTGGCTCTAATCCTGGATATCAAGAGGACACCCGGGAACAAACCTAGGATTGCTGAAATGATATGTGACATTGATACATATATCGTAGAGGCAGGATTAGCCAGTTTTATCCTGACTATTAAGTTTGGGATAGAAACTATGTATCCTGCTCTTGGACTGCATGAATTTGCTGGTGAGTTATCCACACTTGAGTCCTTGATGAATCTTTACCAGCAAATGGGAGAAACTGCACCCTACATGGTAATCCTAGAGAACTCAATTCAGAACAAGTTCAGTGCAGGATCATACCCTCTGCTCTGGAGCTATGCCATGGGAGTAGGAGTGGAACTTGAAAACTCCATGGGAGGTTTGAACTTTGGTCGATCTTACTTTGATCCAGCATATTTTAGATTAGGGCAAGAGATGGTGAGGAGGTCAGCTGGAAAGGTCAGTTCCACATTGGCATCCGAACTCGGTATCACTGCCGAGGATGCAAGGCTTGTTTCAGAGATTGCAATGCATACTACTGAGGACAGGATCAGTAGAGCGGTCGGACCCAGACAAGCCCAAGTGTCATTTCTACACGGTGATCAAAGTGAGAATGAGCTACCAGGATTGGGGGGCAAGGAAGATAGGAGGGTCAAACAGGGTCGGGGAGAAGCCAGGGAGAGCTACAGAGAAACCGGGTCCAGCAGAGCAAGTGATGCGAGAGCTGCCCATCCTCCAACCAGCATGCCCCTAGACATTGACACTGCATCGGAGTCAGGCCAAGATCCGCAGGACAGTCGAAGGTCAGCTGACGCCCTGCTCAGGCTGCAAGCCATGGCAGGAATCTTGGAAGAACAAGGCTCAGACACGGACACCCCTAGGGTATACAATGACAGAGATCTTCTAGACTAGGTGCGAGAGGCCGAGGACCAGAACAACATCCGCCTACCCTCCATCATTGTTATAAAAAACTTAGGAACCAGGTCCACACAGCCGCCAGCCAACCAACCATCCACTCCCACGACTGGAGCCGATGGCAGAAGAGCAGGCACGCCATGTCAAAAACGGACTGGAATGCATCCGGGCTCTCAAGGCCGAGCCCATCGGCTCACTGGCCGTCGAGGAAGCCATGGCAGCATGGTCAGAAATATCAGACAACCCAGGACAGGACCGAGCCACCTGCAAGGAAGAGGAGGCAGGCAGTTCGGGTCTCAGCAAACCATGCCTCTCAGCAATTGGATCAACTGAAGGCGGTGCACCTCGCATCCGCGGTCAGGGATCTGGAGAAAGCGATGACGACGCTGAAACTTTGGGAATCCCCTCAAGAAATCTCCAGGCATCAAGCACTGGGTTACAGTGTTATCATGTTTATGATCACAGCGGTGAAGCGGTTAAGGGAATCCAAGATGCTGACTCTATCATGGTTCAATCAGGCCTTGATGGTGATAGCACCCTCTCAGGAGGAGACGATGAATCTGAAAACAGCGATGTGGATATTGGCGAACCTGATACCGAGGGATATGCTATCACTGACCGGGGATCTGCTCCCATCTCTATGGGGTTCAGGGCTTCTGATGTTGAAACTGCAGAAGGAGGGGAGATCCACGAGCTCCTGAAACTCCAATCCAGAGGCAACAACTTTCCGAAGCTTGGGAAAACTCTCAATGTTCCTCCGCCCCCGAACCCCAGTAGGGCCAGCACTTCCGAGACACCCATTAAAAAGGGCACAGACGCGAGATTGGCCTCATTTGGAACGGAGATCGCGTCTTTATTGACAGGTGGTGCAACCCAATGTGCTCGAAAGTCACCCTCGGAACCATCAGGGCCAGGTGCACCTGCGGGGAATGTCCCCGAGTGTGTGAGCAATGCCGCACTGATACAGGAGTGGACACCCGAATCTGGTACCACAATCTCCCCGAGATCCCAGAATAATGAAGAAGGGGGAGACTATTATGATGATGAGCTGTTCTCCGATGTCCAAGACATCAAAACAGCCTTGGCCAAAATACACGAGGATAATCAGAAGATAATCTCCAAGCTAGAATCATTGCTGTTATTGAAGGGAGAAGTTGAGTCAATTAAGAAGCAGATCAACAGGCAAAATATCAGCATATCCACCCTGGAAGGACACCTCTCAAGCATCATGATTGCCATTCCTGGACTTGGGAAGGATCCCAACGACCCCACTGCAGATGTCGAACTCAATCCCGACCTGAAACCCATCATAGGCAGAGATTCAGGCCGAGCACTGGCCGAAGTTCTCAAGAAGCCCGTTGCCAGCCGACAACTCCAGGGAATGACTAATGGACGGACCAGTTCCAGAGGACAGCTGCTGAAGGAATTTCAACTAAAGCCGATCGGGAAAAAGGTGAGCTCAGCCGTCGGGTTTGTTCCTGACACCGGCCCTGCATCACGCAGTGTAATCCGCTCCATTATAAAATCCAGCCGGCTAGAGGAGGATCGGAAGCGTTACCTGATGACTCTCCTTGATGATATCAAAGGAGCCAACGATCTTGCCAAGTTCCACCAGATGCTGATGAAGATAATAATGAAGTAGCTACAGCTCAACTTACCTGCCAACCCCATGCCAGTCGACCTAATTAGTACAACCTAAATCCATTATAAAAAACTTAGGAGCAAAGTGATTGCCTCCTAAGTTCCACAATGACAGAGATCTACGATTTCGACAAGTCGGCATGGGACATCAAAGGGTCGATCGCTCCGATACAACCTACCACCTACAGTGATGGCAGGCTGGTGCCCCAGGTCAGAGTCATAGATCCTGGTCTAGGTGATAGGAAGGATGAATGCTTTATGTACATGTTTCTGCTGGGGGTTGTTGAGGACAGCGATCCCCTAGGGCCTCCAATCGGGCGAGCATTCGGGTCCCTGCCCTTAGGTGTTGGTAGATCCACAGCAAAACCCGAGGAACTCCTCAAAGAGGCCACTGAGCTTGACATAGTTGTTAGACGTACAGCAGGGCTCAATGAAAAACTGGTGTTCTACAACAACACCCCACTAACCCTCCTCACACCTTGGAGAAAGGTCCTAACAACAGGGAGTGTCTTCAATGCAAACCAAGTGTGCAATGCGGTTAATCTAATACCGCTGGACACCCCGCAGAGGTTCCGTGTTGTTTATATGAGCATCACCCGTCTTTCGGATAACGGGTATTACACCGTTCCCAGAAGAATGCTGGAATTCAGATCGGTCAATGCAGTGGCCTTCAACCTGCTAGTGACCCTTAGGATTGACAAGGCGATTGGCCCTGGGAAGATCATCGACAATGCAGAGCAACTTCCTGAGGCAACATTTATGGTCCACATCGGGAACTTCAGGAGAAAGAAGAGTGAAGTCTACTCTGCCGATTATTGCAAAATGAAAATCGAAAAGATGGGCCTGGTTTTTGCACTTGGTGGGATAGGGGGCACCAGTCTTCACATTAGAAGCACAGGCAAAATGAGCAAGACTCTCCATGCACAACTCGGGTTCAAGAAGACCTTATGTTACCCACTGATGGATATCAATGAAGACCTTAATCGGTTACTCTGGAGGAGCAGATGCAAGATAGTAAGAATCCAGGCAGTTTTGCAGCCATCAGTTCCTCAAGAATTCCGCATTTACGACGACGTGATCATAAATGATGACCAAGGACTATTCAAAGTTCTGTAGACCGCAGTGCCCAGCAATACCCGAAAACGACCCCCCTCATAATGACAGCCAGAAGGCCCGGACAAAAAAGCCCCCTCCAGAAGACTCCACGGACCAAGCGAGAGGCCAGCCAGCAGCCGACAGCAAGTGTGGACACCAGGCGGCCCAAGCACAGAACAGCCCCGACACAAGGCCACCACCAGCCATCCCAATCTGCGTCCTCCTCGTGGGACCCCCGAGGACCAACCCCGAAGGTCGCTCCGAACACAGACCACCAACCGCATCCCCACAGCTCCCGGGAAAGGAACCCCCAGCAACTGGAAGGCCCCTCCCCCCCTCCCCCAACGCAAGAACCCCACAACCGAACCGCACAAGCGACCGAGGTGACCCAACCGCAGGCATCCGACTCCTTAGACAGATCCTCTCCCCCCGGCATACTAAACAAAACTTAGGGCCAAGGAACACACACACTCGACAGAACCCAGACCCCGGCCCGCGGCACCGCGCCCCCACCCCCCGAAAACCAGAGGGAGCCCCCAACCAAACCCGCCGGCCCCCCCGGTGCCCACAGGTAGGCACACCAACCCCCGACCAGACCCAGCACCCAGCCACCGACAATCCAAGACGGGGGGCCCCCCCCAAAAAAAGGCCCCCAGGGGCCGACAGCCAGCATCGCGAGGAAGCACACCCACCCCACACACGACCACGGCAACCGAACCAGAGTCCAGACCACCCTGGGCCACCAGCTCCCAGACTCGGCCATCACCCCGCAAAAAGGAAAGGCCACAACCCGCGCACCCCAGCCCCGATCCGGCGGGCAGCCACTCAACCCGAACCAGCACCCAAGAGCGATCCCTGGGGGACCCCCAAACCGCAAAAGACATCAGTATCCCACAGCCTCTCCAAGTCCCCCGGTCTCCTCCTCTTCTCGAAGGGACCAAAAGATCAATCCACCACATCCGACGACACTCAATTCCCCACCCCCAAAGGAGACACCGGGAATCCCAGAATCAAGACTCATCCAGTGTCCATCATGGGTCTCAAGGTGAACGTCTCTGCCATATTCATGGCAGTACTGTTAACTCTCCAAACACCCACCGGTCAAATCCATTGGGGCAATCTCTCTAAGATAGGGGTGGTAGGGATAGGAAGTGCAAGCTACAAAGTTATGACTCGTTCCAGCCATCAATCATTGGTCATAAAATTAATGCCCAATATAACTCTCCTCAATAACTGCACGAGGGTAGAGATCGCAGAATACAGGAGACTACTGAGAACAGTTTTGGAACCAATTAGAGATGCACTTAATGCAATGACCCAGAATATAAGACCGGTTCAGAGTGTAGCTTCAAGTAGGAGACACAAGAGATTTGCGGGAGTTGTCCTGGCAGGTGCGGCCCTAGGCGTTGCCACAGCTGCTCAGATAACAGCCGGCATTGCACTTCACCAGTCCATGCTGAACTCTCAAGCCATCGACAATCTGAGAGCAAGCCTGGAAACTACTAATCAGGCAATTGAGGCAATCAGACAAGCAGGGCAGGAGATGATATTGGCTGTTCAGGGTGTCCAAGACTACATCAATAATGAGCTGATACCGTCTATGAACCAACTATCTTGTGATTTAATCGGCCAGAAGCTAGGGCTCAAATTGCTCAGATACTATACAGAAATCCTGTCATTATTTGGCCCCAGCTTACGGGACCCCATATCTGCGGAGATATCCATCCAGGCTTTGAGCTATGCGCTTGGAGGAGATATCAATAAGGTATTAGAAAAGCTCGGATACAGTGGAGGTGATTTACTGGGCATCTTAGAGAGCAGAGGAATAAAGGCCCGGATAACTCACGTCGACACAGAGTCCTACTTCATTGTACTCAGTATAGCCTATCCGACGCTGTCCGAGATTAAGGGGGTGATTGTCCACCGGCTAGAGGGGGTCTCGTACAATATAGGCTCTCAAGAGTGGTATACCACTGTGCCCAAGTATGTTGCAACCCAAGGGTACCTTATCTCGAATTTTGATGAGTCATCGTGTACTTTCATGCCAGAGGGGACTGTGTGCAGCCAAAATGCCTTGTACCCGATGAGTCCTCTGCTCCAAGAATGCCTCCGGGGGTCCACCAAGTCCTGTGCTCGTACACTCGTATCCGGGTCTTTTGGGAACCGGTTCATTTTATCACAAGGGAACCTAATAGCCAATTGTGCATCAATCCTCTGCAAGTGTTACACAACAGGAACGATCATTAATCAAGACCCTGACAAGATCCTAACATACATTGCTGCCGATCACTGCCCGGTGGTCGAGGTGAACGGTGTGACCATCCAAGTCGGGAGCAGGAGGTATCCGGACGCGGTGTACCTGCACAGAATTGACCTCGGTCCTCCCATATCATTGGAGAGGTTGGACGTAGGGACAAATCTGGGGAATGCAATTGCTAAGTTGGAGGATGCCAAGGAATTGTTGGAGTCATCGGACCAGATATTGAGGAGTATGAAAGGTTTATCGAGCACTAGCATAGTTTACATCCTGATTGCAGTGTGTCTTGGAGGGTTGATAGGGATCCCCGCTTTAATATGTTGCTGCAGGGGGCGCTGTAACAAAAAGGGAGAACAAGTTGGTATGTCAAGACCAGGCCTAAAGCCTGATCTTACAGGGACATCAAAATCCTATGTAAGGTCGCTCTGATCCTCTACAACTCTTGAAACACAGATTTCCCACAAGTCTCCTCTTCGTCATCAAGCAACCACCGCATCCAGCATCAAGCCCACCTGAAATTGTCTCCGGCTTCCCTCTGGCCGAACGATATCGGTAGTTAATTAAAACTTAGGGTGCAAGATCATCCACAATGTCACCACAACGAGACCGAATAAATGCCTTCTACAAAGACAACCCACATCCTAAGGGAAGTAGGATAGTTATTAACAGAGAACATCTTATGATTGATAGACCTTATGTTTTGCTGGCTGTTCTATTCGTCATGTTTCTGAGCTTGATCGGGTTGCTAGCCATTGCAGGCATTAGACTCCATCGTGCAGCCATCTACACCGCAGAGATCCATAAGAGCCTCAGCACCAATCTAGATGTAACTAACTCGATCGAGCATCAGGTCAAGGACGTGCTGACACCACTCTTCAAGATCATTGGTGATGAAGTGGGCCTGAGGACACCTCAGAGATTCACTGACCTAGTGAAATTCATCTCTGACAAAATTAAATTCCTTAATCCGGATAGGGAGTACGACTTCAGAGATCTCACTTGGTGTATCAACCCGCCAGAGAGAATCAAATTGGATTATGATCAATACTGTGCAGATGTGGCTGCTGAAGAACTCATGAATGCATTGGTGAACTCAACTCTACTGGAGGCCAGGGCAACCAATCAGTTCCTAGCTGTCTCAAAGGGAAACTGCTCAGGGCCCACTACAATCAGAGGTCAATTCTCAAACATGTCGCTGTCCCTGTTGGACTTGTATTTAAGTCGAGGTTACAATGTGTCATCTATAGTCACTATGACATCCCAGGGAATGTACGGGGGAACTTACCTAGTGGGAAAGCCTAATCTGAGCAGTAAAGGGTCAGAGTTGTCACAACTGAGCATGCACCGAGTGTTTGAAGTAGGGGTTATCAGAAATCCGGGTTTGGGGGCTCCGGTGTTCCATATGACAAACTATTTTGAGCAACCAGTCAGTAATGATTTCAGCAACTGCATGGTGGCTTTGGGGGAGCTTAAATTCGCAGCCCTCTGTCACAGGGAAGATTCTATCACAATTCCCTATCAGGGGTCAGGGAAAGGTGTCAGCTTCCAGCTCGTCAAGCTAGGTGTCTGGAAATCCCCAACCGACATGCGATCCTGGGTCCCCCTATCAACGGATGATCCAGTGATAGATAGGCTTTACCTCTCATCTCACAGAGGTGTTATCGCTGACAATCAAGCAAAATGGGCTGTCCCGACAACACGGACAGATGACAAGTTGCGAATGGAGACATGCTTCCAGCAGGCGTGTAAGGGTAAAAACCAAGCACTCTGCGAGAATCCCGAGTGGGCACCATTGAAGGATAACAGGATTCCTTCATACGGGGTCTTGTCTGTTAATCTGAGTCTGACAGTTGAGCTTAAAATCAAAATTGCTTCAGGATTCGGGCCATTGATCACACACGGTTCAGGGATGGACCTATACAAAACCAACCACAACAATGTGTATTGGCTGACTATCCCGCCAATGAAGAACCTAGCCTTAGGTGTAATCAACACATTGGAGTGGATACCGAGATTCAAGGTTAGTCCCAACCTCTTCACTGTTCCAATCAAGGAAGCAGGCGAGGACTGCCATGCCCCAACATACCTACCTGCGGAGGTGGATGGTGATGTCAAACTCAGTTCCAATCTGGTAATTCTACCTGGTCAGGATCTCCAATATGTTTTGGCAACCTACGATACTTCCAGGGTTGAACATGCTGTGGTTTATTATGTTTACAGCCCAAGCCGCTCATTTTCTTACTTTTATCCTTTTAGGTTGCCTATAAAGGGGGTCCCAATCGAATTACAAGTGGAATGCTTCACATGGGACAAAAAACTCTGGTGCCGTCACTTCTGTGTGCTTGCGGACTCAGAATCTGGTGGACATATCACTCACTCTGGGATGGTGGGCATGGGAGTCAGCTGCACAGTCACTCGGGAAGATGGAACCAATCGCAGATAGGGCTGCCAGTGAACCGATCACATGATGTCACCCAGACATCAGGCATACCCACTAGTGTGAAATAGACATCAGAATTAAGAAAAACGTAGGGTCCAAGTGGTTTCCCGTTATGGACTCGCTATCTGTCAACCAGATCTTATACCCTGAAGTTCACCTAGATAGCCCGATAGTTACCAATAAGATAGTAGCTATCCTGGAGTATGCTCGAGTCCCTCACGCTTACAGCCTGGAGGACCCTACACTGTGTCAGAACATCAAGCACCGCCTAAAAAACGGATTCTCCAACCAAATGATTATAAACAATGTGGAAGTTGGGAATGTCATCAAGTCCAAGCTTAGGAGTTATCCGGCCCACTCTCATATTCCATATCCAAATTGTAATCAGGATTTATTTAACATAGAAGACAAAGAGTCAACAAGGAAGATCCGTGAGCTCCTAAAAAAGGGAAATTCGCTGTACTCCAAAGTCAGTGATAAGGTTTTCCAATGCCTGAGGGACACTAACTCACGGCTTGGCCTAGGCTCCGAATTGAGGGAGGACATCAAGGAGAAAATTATTAACTTGGGAGTTTACATGCACAGCTCCCAATGGTTTGAGCCCTTTCTGTTTTGGTTTACAGTCAAGACTGAGATGAGGTCAGTGATTAAATCACAAACCCATACTTGCCATAGGAGGAGACACACACCTGTATTCTTCACTGGTAGTTCAGTTGAGCTGTTAATCTCTCGTGACCTTGTTGCTATAATCAGTAAGGAGTCTCAACATGTATATTACCTGACGTTTGAACTGGTTTTGATGTATTGTGATGTCATAGAGGGGAGGTTAATGACAGAGACCGCTATGACCATTGATGCTAGGTATGCAGAACTTCTAGGAAGAGTCAGATACATGTGGAAACTGATAGATGGTTTCTTCCCTGCACTCGGGAATCCAACTTATCAAATTGTAGCCATGCTGGAGCCACTTTCACTTGCTTACCTGCAACTGAGGGATATAACAGTAGAACTCAGAGGTGCTTTCCTTAACCACTGCTTTACTGAAATACATGATGTTCTTGACCAAAACGGGTTTTCTGATGAAGGTACTTATCATGAGTTAATTGAAGCCCTAGATTACATTTTCATAACTGATGACATACATCTGACAGGGGAGATTTTCTCATTTTTCAGAAGTTTCGGCCACCCCAGACTTGAAGCAGTAACGGCTGCTGAAAATGTCAGGAAATACATGAATCAGCCTAAAGTCATTGTGTATGAGACTCTGATGAAAGGTCATGCCATATTTTGTGGAATCATAATCAACGGCTATCGTGACAGGCACGGAGGCAGTTGGCCACCCCTGACCCTCCCCCTGCATGCTGCAGACACAATCCGGAATGCTCAAGCTTCAGGTGAAGGGTTAACACATGAGCAGTGCGTTGATAACTGGAAATCATTTGCTGGAGTGAGATTTGGCTGTTTTATGCCTCTTAGCCTGGACAGTGATCTGACAATGTACCTAAAGGACAAGGCACTTGCTGCTCTCCAAAGGGAATGGGATTCAGTTTACCCGAAAGAGTTCCTGCGTTACGATCCTCCCAAGGGAACCGGGTCACGGAGGCTTGTAGATGTTTTCCTTAATGATTCGAGCTTTGACCCATATGATATGATAATGTATGTCGTAAGTGGAGCCTACCTCCATGACCCTGAGTTCAACCTGTCTTACAGCCTGAAAGAAAAGGAGATCAAGGAAACAGGTAGACTTTTCGCTAAAATGACTTACAAAATGAGGGCATGCCAAGTGATCGCTGAAAATCTAATCTCAAACGGGATTGGCAAGTATTTTAAGGACAATGGGATGGCCAAGGATGAGCACGATTTGACTAAGGCACTCCACACTCTGGCTGTCTCAGGAGTCCCCAAAGATCTCAAAGAAAGTCACAGGGGGGGGCCAGTCTTAAAAACCTACTCCCGAAGCCCAGTCCACACAAGTACCAGGAACGTTAAAGCAGAAAAAGGGTTTGTAGGATTCCCTCATGTAATTCGGCAGAATCAAGACACTGATCATCCGGAGAATATAGAAACCTACGAGACAGTCAGCGCATTTATCACGACTGATCTCAAGAAGTACTGCCTTAATTGGAGATATGAGACCATCAGCTTATTTGCACAGAGGCTAAATGAGATTTACGGATTACCCTCATTTTTTCAGTGGCTGCATAAGAGGCTTGAAACCTCTGTCCTCTATGTAAGTGACCCTCATTGCCCCCCCGACCTTGACGCCCATGTCCCGTTATGCAAAGTCCCCAATGACCAAATCTTCATCAAGTACCCTATGGGAGGTATAGAAGGGTATTGTCAGAAGCTGTGGACCATCAGCACCATTCCCTACTTATACCTGGCTGCTTATGAGAGCGGGGTAAGGATTGCTTCGTTAGTGCAAGGGGACAATCAGACCATAGCCGTAACAAAAAGGGTACCCAGCACATGGCCTTACAACCTTAAGAAACGGGAAGCTGCTAGAGTAACTAGAGATTACTTTGTAATTCTTAGGCAAAGGCTACATGACATTGGCCATCACCTCAAGGCAAATGAGACAATTGTTTCATCACATTTTTTTGTCTATTCAAAAGGAATATATTATGATGGGCTACTTGTGTCCCAATCACTCAAGAGCATCGCAAGATGTGTATTCTGGTCAGAGACTATAGTTGATGAAACAAGGGCAGCATGCAGTAATATTGCTACAACAATGGCTAAAAGCATCGAGAGAGGTTATGACCGTTATCTTGCATATTCCCTGAACGTCCTAAAAGTGATACAGCAAATTTTGATCTCTCTTGGCTTCACAATCAATTCAACCATGACCCGAGATGTAGTCATACCCCTCCTCACAAACAACGATCTCTTAATAAGGATGGCACTGTTGCCCGCTCCTATTGGGGGGATGAATTATCTGAATATGAGCAGGCTGTTTGTCAGAAACATCGGTGATCCAGTAACATCATCAATTGCTGATCTCAAGAGAATGATTCTCGCATCACTAATGCCTGAAGAGACCCTCCATCAAGTAATGACACAACAACCGGGGGACTCTTCATTCCTAGACTGGGCTAGCGACCCTTACTCAGCAAATCTTGTATGCGTCCAGAGCATCACTAGACTCCTCAAGAACATAACTGCAAGGTTTGTCCTAATCCATAGTCCAAACCCAATGTTAAAAGGGTTATTCCATGATGACAGTAAAGAAGAGGACGAGAGACTGGCGGCATTCCTCATGGACAGGCATATTATAGTACCTAGGGCAGCTCATGAAATCCTGGATCATAGTGTCACAGGGGCAAGAGAGTCTATTGCAGGCATGCTAGATACCACAAAAGGCCTGATTCGAGCCAGCATGAGGAAGGGGGGGTTAACCTCTCGAGTGATAACCAGATTGTCCAATTATGACTATGAACAATTTAGAGCAGGGATGGTGCTATTGACAGGAAGAAAGAGAAATGTCCTCATTGACAAAGAGTCATGTTCAGTGCAGCTGGCTAGAGCCCTAAGAAGCCATATGTGGGCAAGACTAGCTCGAGGACGGCCTATTTACGGCCTTGAGGTCCCTGATGTACTAGAATCTATGCGAGGCCACCTTATTCGGCGTCATGAGACATGTGTCATCTGCGAGTGTGGATCAGTCAACTACGGATGGTTTTTTGTCCCCTCGGGTTGCCAACTGGATGATATTGACAAGGAAACATCATCCTTGAGAGTCCCATATATTGGTTCTACCACTGATGAGAGAACAGACATGAAGCTCGCCTTCGTAAGAGCCCCAAGTAGATCCTTGCGATCTGCCGTTAGAATAGCAACAGTGTACTCATGGGCTTACGGTGATGATGATAGCTCTTGGAACGAAGCCTGGTTGTTGGCAAGGCAAAGGGCCAATGTGAGCCTGGAGGAGCTAAGGGTGATCACTCCCATCTCGACTTCGACTAATTTAGCGCATAGGTTGAGGGATCGTAGCACTCAAGTGAAATACTCAGGTACATCCCTTGTCCGAGTGGCAAGGTATACCACAATCTCCAACGACAATCTCTCATTTGTCATATCAGATAAGAAGGTTGATACTAACTTTATATACCAACAAGGAATGCTTCTAGGGTTGGGTGTTTTAGAAACATTGTTTCGACTCGAGAAAGATACTGGATCATCTAACACGGTATTACATCTTCACGTCGAAACAGATTGTTGCGTGATCCCGATGATAGATCATCCCAGGATACCCAGCTCCCGCAAGCTAGAGCTGAGGGCAGAGCTATGTACCAACCCATTGATATATGATAATGCACCTTTAATTGACAGAGATGCAACAAGGCTATACACCCAGAGCCATAGGAGGCACCTTGTGGAATTTGTTACATGGTCCACACCCCAACTATATCACATTCTAGCTAAGTCCACAGCACTATCTATGATTGACCTGGTAACAAAATTTGAGAAGGACCATATGAATGAAATTTCAGCTCTCATAGGGGATGACGATATCAATAGTTTCATAACTGAGTTTCTGCTTATAGAGCCAAGATTATTCACCATCTACTTGGGCCAGTGTGCAGCCATCAATTGGGCATTTGATGTACATTATCATAGACCATCAGGGAAATATCAGATGGGTGAGCTGTTGTCTTCGTTCCTTTCTAGAATGAGCAAAGGAGTGTTTAAGGTGCTTGTCAATGCTCTAAGCCACCCAAAGATCTACAAGAAATTCTGGCATTGTGGTATTATAGAGCCTATCCATGGTCCTTCACTTGATGCTCAAAACTTGCACACAACTGTGTGCAACATGGTTTACACATGCTATATGACCTACCTCGACCTGTTGTTGAATGAAGAGTTAGAAGAGTTCACATTTCTTTTGTGTGAAAGCGATGAGGATGTAGTACCGGACAGATTCGACAACATCCAGGCAAAACACTTGTGTGTTCTGGCAGATTTGTACTGTCAACCAGGGACCTGCCCACCGATTCGAGGTCTAAGGCCGGTAGAGAAATGTGCAGTTCTAACCGATCATATCAAGGCAGAGGCTAGGTTATCTCCAGCAGGATCTTCGTGGAACATAAATCCAATTATTGTAGACCATTACTCATGCTCTCTGACTTATCTCCGTCGAGGATCTATCAAACAGATAAGATTGAGAGTTGATCCAGGATTCATTTTTGACGCCCTCGCTGAGGTAAATGTCAGTCAGCCAAAGGTCGGCAGCAACAACATCTCAAATATGAGCATCAAGGATTTCAGACCTCCACACGATGATGTTGCAAAATTGCTCAAAGATATCAACACAAGCAAGCACAATCTTCCCATTTCAGGGGGTAGTCTCGCCAATTATGAAATCCATGCTTTCCGCAGAATCGGGTTAAACTCATCTGCTTGCTACAAAGCTGTTGAGATATCAACATTAATTAGGAGATGCCTTGAGCCAGGGGAAGACGGCTTGTTCTTGGGTGAGGGGTCGGGTTCTATGTTGATCACTTATAAGGAGATACTAAAACTAAACAAGTGCTTCTATAATAGTGGGGTTTCCGCCAATTCTAGATCTGGTCAAAGGGAATTAGCACCCTATCCCTCCGAAGTTGGCCTTGTCGAACACAGAATGGGAGTAGGTAATATTGTCAAGGTGCTCTTTAACGGGAGGCCCGAAGTCACGTGGGTAGGCAGTATAGATTGCTTCAATTTCATAGTCAGTAATATCCCTACCTCTAGTGTGGGGTTTATCCATTCAGATATAGAGACCTTACCTAACAAAGATACTATAGAGAAGCTAGAGGAATTGGCAGCCATCTTATCGATGGCTCTACTCCTTGGCAAAATAGGATCAATACTGGTGATTAAGCTTATGCCTTTCAGCGGGGATTTTGTTCAGGGATTTATAAGCTATGTAGGGTCTCATTATAGAGAAGTGAACCTTGTCTACCCTAGGTACAGCAACTTCATATCTACTGAATCTTATTTAGTTATGACAGATCTCAAAGCTAACCGGCTAATGAATCCTGAAAAGATCAAGCAGCAGATAATTGAATCATCTGTGCGGACTTCACCTGGACTTATAGGTCACATCCTATCCATTAAGCAACTAAGCTGCATACAAGCAATTGTGGGAGGCGCAGTTAGTAGAGGTGATATCAACCCTATTCTGAAAAAACTTACACCTATAGAGCAGGTGCTGATCAGTTGCGGGTTGGCAATTAACGGACCTAAACTGTGCAAAGAATTAATCCACCATGATGTTGCCTCAGGGCAAGATGGATTGCTTAACTCTATACTCATCCTCTACAGGGAGTTGGCAAGATTCAAAGACAACCAAAGAAGTCAACAAGGGATGTTCCACGCTTACCCCGTATTGGTAAGTAGTAGGCAACGAGAACTTGTATCTAGGATCACTCGCAAATTTTGGGGGCATATTCTTCTTTACTCCGGGAACAGAAAGTTGATAAATCGGTTTATCCAGAATCTCAAGTCCGGTTATCTAGTACTAGACTTACACCAGAATATCTTCGTTAAGAATCTATCCAAGTCAGAGAAACAGATTATTATGACGGGGGGTTTAAAACGTGAGTGGGTTTTTAAGGTAACAGTCAAGGAGACCAAAGAATGGTACAAGTTAGTCGGATACAGCGCTCTGATTAAGGATTAATTGGTTGAACTCCGGAACCCTAATCCTGCCCTAGGTAGTTAGGCATTATTTGCAATATATTAAAGAAAACTTTGAAAATACGAAGTTTCTATTCCCAGCTTTGTCTGGT'