Biomedical Data Mining, Fall 2022

> ### Term Project#1 Motif Finding

310551178 穆冠蓁

Code Explanation

1. Read input file

```
i = 0
before_len = len("motif width is ")
k = 0
base = ['A', 'T', 'C', 'G']
dna = []
with open('Q1.txt') as f:
    for line in f:
        if not line.isspace():
            temp = line.strip()
            if i == 0 :
                k = int(temp[before_len:])
            elif temp[0] in base:
                dna.append(temp)
            i += 1
```

2. Main function

   Start to call gibbsSampler function and do it by the setting iterations

```
n = 1000 # iteration
t = len(dna)
best = gibbsSampler(dna, k, t, n)
s = score(best)

for x in tqdm(range(200)): # for each iteration of the gibbs sampler
    sample = gibbsSampler(dna, k, t, n)
    if score(sample) < s: # if the score of the sample is less than the score of the best
        s = score(sample) # set s to the score of the sample
        best = sample[:] # set best to a copy of the sample
```

3. gibbsSampler

```
def gibbsSampler(dna, k, t, n): # dna is a list of strings, k,t and n are integers
    bestMotifs = []
    motifs = []
    for x in range(len(dna)): # for each string in dna
        i = random.randint(0, len(dna[x])-k) # i is a random integer between 0 and the length of the string minus k
        motifs.append(dna[x][i:i+k]) # add the k-mer starting at index i to the motifs list
    bestMotifs = motifs[:] # bestMotifs is a copy of motifs
    for i in range(n): # for each iteration
        j = random.randint(0,t-1) # j is a random integer between 0 and t-1
        profile = profileForm(motifs[:j] + motifs[j+1:]) # profile is a list of lists of floats
    # print("j = ",j)
        r = profileRandom(k, profile, dna[j]) # r is a random k-mer from the profile
        motifs[j] = dna[j][r:r+k] # set the jth element of motifs to the k-mer starting at index r
        if score(motifs) < score(bestMotifs): # if the score of motifs is less than the score of bestMotifs
            bestMotifs = motifs[:] # set bestMotifs to a copy of motifs
    return bestMotifs
```

4. Score: to calculate the score of the candidate motif

```
def score(motifs): # motifs is a list of strings
  profile = profileForm(motifs) # profile is a list of lists of floats
  cons = consensus(profile) # cons is a string
  score = 0
  for x in motifs: # for each string in motifs
    for i in range(len(x)): # for each symbol in the string
      if cons[i] != x[i]: # if the symbol in the consensus is different from the symbol in the string
        score += 1 # add 1 to the score
  return score
```

5. profileForm: accumulate the weighting of each base in each position

```
def profileForm(motifs): # motifs is a list of strings
  k = len(motifs[0]) # k is the length of the first string in motifs
  profile = [[1 for i in range(k)] for j in range(4)] # initialize profile to a list of lists of length k  with all elements initialize
  for x in motifs: # for each string in motifs
    for i in range(len(x)): # for each symbol in the string
      j = symbolToNumber(x[i]) # j is the index of the symbol in the profile
      profile[j][i] += 1 # add 1 to the count of the symbol in the profile
  for x in profile: # for each list in profile
    for i in range(len(x)): # for each element in the list
      x[i] = x[i]/len(motifs) # divide the count of the symbol by the number of strings in motifs
  return profile
```

6. profileRandom:calculate the probability by using the findings in the profileForm

```
def profileRandom(k, profile, text): # k is an integer, profile is a list of lists of floats, text is a string
    probs = []
    for i in range(0,len(text) - k +1): # for each kmer in text
        prob = 1.0
        pattern = text[i:i+k] # pattern is the kmer
        for j in range(k): # for each symbol in the kmer
            l = symbolToNumber(pattern[j]) # l is the index of the symbol in the profile
            prob *= profile[l][j] # prob is the product of the probabilities of the symbols in the kmer
        probs.append(prob) # add the probability of the kmer to the list of probabilities
    r = myRandom(probs) # r is the index of the kmer with the highest probability
    return r
```

7. symbolToNumber & numberToSymbol: tranformation between symbol and number

```
def symbolToNumber(symbol): # symbol is a string of length 1
  if symbol == "A":
    return 0
  if symbol == "C":
    return 1
  if symbol == "G":
    return 2
  if symbol == "T":
    return 3

def numberToSymbol(x): # x is an integer
  if x == 0:
    return "A"
  if x == 1:
    return "C"
  if x == 2:
    return "G"
  if x == 3:
    return "T"
```

8. Find the key and find the index of the motif of each DNA sequence

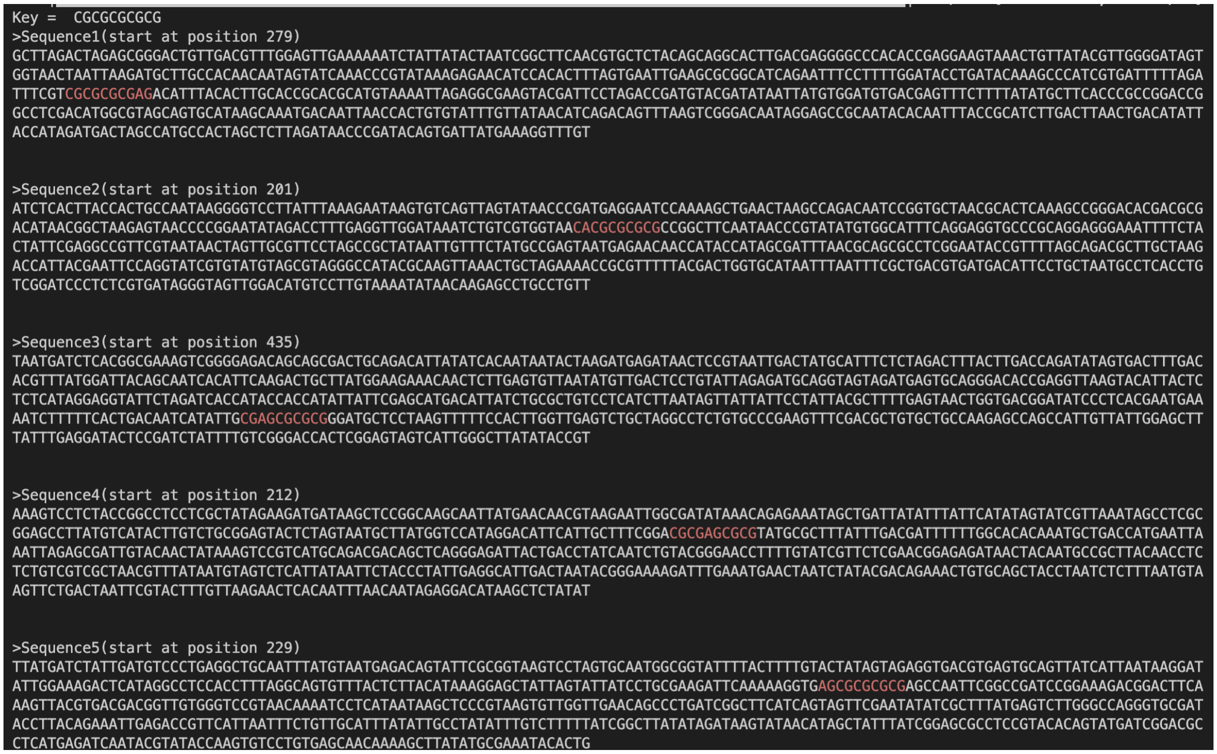
```
# Find the key
Key = str()
for i in range(k):
    count = [0] * 4 # ACGT
    for j in range(len(best)):
        b = best[j][i]
        count[symbolToNumber(b)] += 1
    max_symbol = numberToSymbol(count.index(max(count)))
    Key += max_symbol
output.write(f"Key = {Key}\n")
# Find the index of every motif
index = []
```

```
for i in range(len(dna)):
    b = best[i]
    index.append(dna[i].find(b))
# Print motif result
for i in range(len(dna)):
    start_index = index[i]
    output.write(f">Sequence{i + 1}(start at position {start_index + 1}, motif is {best[i]})")
    output.write("\n")
    output.write(dna[i][:start_index]+"'"+dna[i][start_index:start_index+k]+"'"+dna[i][start_index+k:])
    output.write("\n")
# Print the best motif
for b in best:
  print(b)
output.close()
```

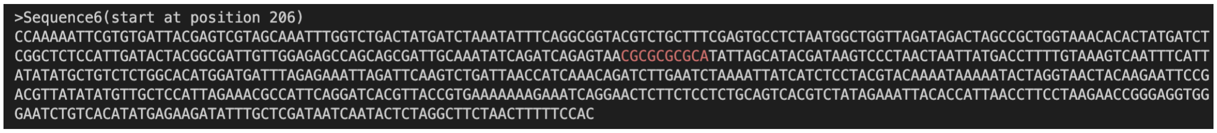
Result

The red bold sequence means the motif we found by the program.Also, the starting position is included.

1. Testing data 1:

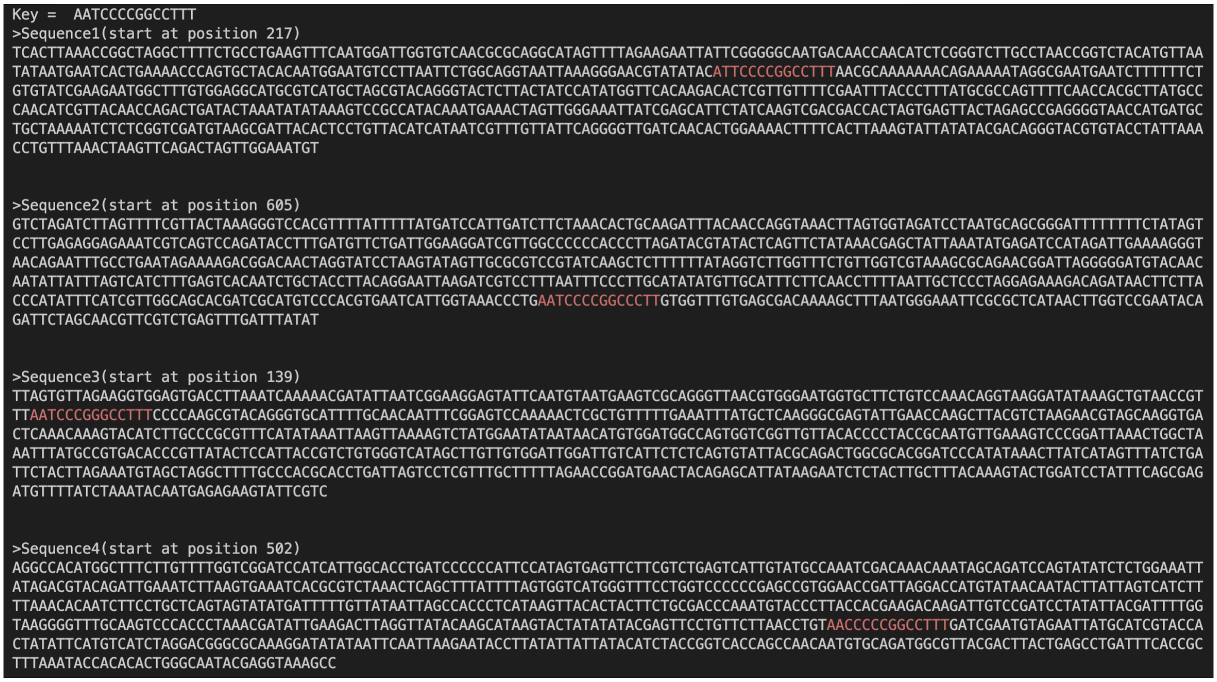


```
Key =  CGCGCGCGCG
>Sequence1(start at position 279)
GCTTAGACTAGAGCGGGACTGTTGACGTTTGGAGTTGAAAAAATCTATTATACTAATCGGCTTCAACGTGCTCTACAGCAGGCACTTGACGAGGGGCCCACACCGAGGAAGTAAACTGTTATACGTTGGGGATAGT
GGTAACTAATTAAGATGCTTGCCACAACAATAGTATCAAACCCGTATAAAGAGAACATCCACACTTTAGTGAATTGAAGCGCGGCATCAGAATTTCCTTTTGGATACCTGATACAAAGCCCATCGTGATTTTTAGA
TTTCGTCGCGCGCGAGACATTTACACTTGCACCGCACGCATGTAAAATTAGAGGCGAAGTACGATTCCTAGACCGATGTACGATATAATTATGTGGATGTGACGAGTTTCTTTTATATGCTTCACCCGCCGGACCG
GCCTCGACATGGCGTAGCAGTGCATAAGCAAATGACAATTAACCACTGTGTATTTGTTATAACATCAGACAGTTTAAGTCGGGACAATAGGAGCCGCAATACACAATTTACCGCATCTTGACTTAACTGACATATT
ACCATAGATGACTAGCCATGCCACTAGCTCTTAGATAACCCGATACAGTGATTATGAAAGGTTTGT


>Sequence2(start at position 201)
ATCTCACTTACCACTGCCAATAAGGGGTCCTTATTTAAAGAATAAGTGTCAGTTAGTATAACCCGATGAGGAATCCAAAAGCTGAACTAAGCCAGACAATCCGGTGCTAACGCACTCAAAGCCGGGACACGACGCG
ACATAACGGCTAAGAGTAACCCCGGAATATAGACCTTTGAGGTTGGATAAATCTGTCGTGGTAACACGCGCGCGCCGGCTTCAATAACCCGTATATGTGGCATTTCAGGAGGTGCCCGCAGGAGGGAAATTTTCTA
CTATTCGAGGCCGTTCGTAATAACTAGTTGCGTTCCTAGCCGCTATAATTGTTTCTATGCCGAGTAATGAGAACAACCATACCATAGCGATTTAACGCAGCGCCTCGGAATACCGTTTTAGCAGACGCTTGCTAAG
ACCATTACGAATTCCAGGTATCGTGTATGTAGCGTAGGGCCATACGCAAGTTAAACTGCTAGAAAACCGCGTTTTTACGACTGGTGCATAATTTAATTTCGCTGACGTGATGACATTCCTGCTAATGCCTCACCTG
TCGGATCCCTCTCGTGATAGGGTAGTTGGACATGTCCTTGTAAAATATAACAAGAGCCTGCCTGTT


>Sequence3(start at position 435)
TAATGATCTCACGGCGAAAGTCGGGGAGACAGCAGCGACTGCAGACATTATATCACAATAATACTAAGATGAGATAACTCCGTAATTGACTATGCATTTCTCTAGACTTTACTTGACCAGATATAGTGACTTTGAC
ACGTTTATGGATTACAGCAATCACATTCAAGACTGCTTATGGAAGAAACAACTCTTGAGTGTTAATATGTTGACTCCTGTATTAGAGATGCAGGTAGTAGATGAGTGCAGGGACACCGAGGTTAAGTACATTACTC
TCTCATAGGAGGTATTCTAGATCACCATACCACCATATTATTCGAGCATGACATTATCTGCGCTGTCCTCATCTTAATAGTTATTATTCCTATTACGCTTTTGAGTAACTGGTGACGGATATCCCTCACGAATGAA
AATCTTTTTCACTGACAATCATATTGCGAGCGCGCGGGATGCTCCTAAGTTTTTCCACTTGGTTGAGTCTGCTAGGCCTCTGTGCCCGAAGTTTCGACGCTGTGCTGCCAAGAGCCAGCCATTGTTATTGGAGCTT
TATTTGAGGATACTCCGATCTATTTTGTCGGGACCACTCGGAGTAGTCATTGGGCTTATATACCGT


>Sequence4(start at position 212)
AAAGTCCTCTACCGGCCTCCTCGCTATAGAAGATGATAAGCTCCGGCAAGCAATTATGAACAACGTAAGAATTGGCGATATAAACAGAGAAATAGCTGATTATATTTATTCATATAGTATCGTTAAATAGCCTCGC
GGAGCCTTATGTCATACTTGTCTGCGGAGTACTCTAGTAATGCTTATGGTCCATAGGACATTCATTGCTTTCGGACGCGAGCGCGTATGCGCTTTATTTGACGATTTTTTGGCACACAAATGCTGACCATGAATTA
AATTAGAGCGATTGTACAACTATAAAGTCCGTCATGCAGACGACAGCTCAGGGAGATTACTGACCTATCAATCTGTACGGGAACCTTTTGTATCGTTCTCGAACGGAGAGATAACTACAATGCCGCTTACAACCTC
TCTGTCGTCGCTAACGTTTATAATGTAGTCTCATTATAATTCTACCCTATTGAGGCATTGACTAATACGGGAAAAGATTTGAAATGAACTAATCTATACGACAGAAACTGTGCAGCTACCTAATCTCTTTAATGTA
AGTTCTGACTAATTCGTACTTTGTTAAGAACTCACAATTTAACAATAGAGGACATAAGCTCTATAT


>Sequence5(start at position 229)
TTATGATCTATTGATGTCCCTGAGGCTGCAATTTATGTAATGAGACAGTATTCGCGGTAAGTCCTAGTGCAATGGCGGTATTTTACTTTTGTACTATAGTAGAGGTGACGTGAGTGCAGTTATCATTAATAAGGAT
ATTGGAAAGACTCATAGGCCTCCACCTTTAGGCAGTGTTTACTCTTACATAAAGGAGCTATTAGTATTATCCTGCGAAGATTCAAAAAGGTGAGCGCGCGCGAGCCAATTCGGCCGATCCGGAAAGACGGACTTCA
AAGTTACGTGACGACGGTTGTGGGTCCGTAACAAAATCCTCATAATAAGCTCCCGTAAGTGTTGGTTGAACAGCCCTGATCGGCTTCATCAGTAGTTCGAATATATCGCTTTATGAGTCTTGGGCCAGGGTGCGAT
ACCTTACAGAAATTGAGACCGTTCATTAATTTCTGTTGCATTTATATTGCCTATATTTGTCTTTTTATCGGCTTATATAGATAAGTATAACATAGCTATTTATCGGAGCGCCTCCGTACACAGTATGATCGGACGC
CTCATGAGATCAATACGTATACCAAGTGTCCTGTGAGCAACAAAAGCTTATATGCGAAATACACTG
```

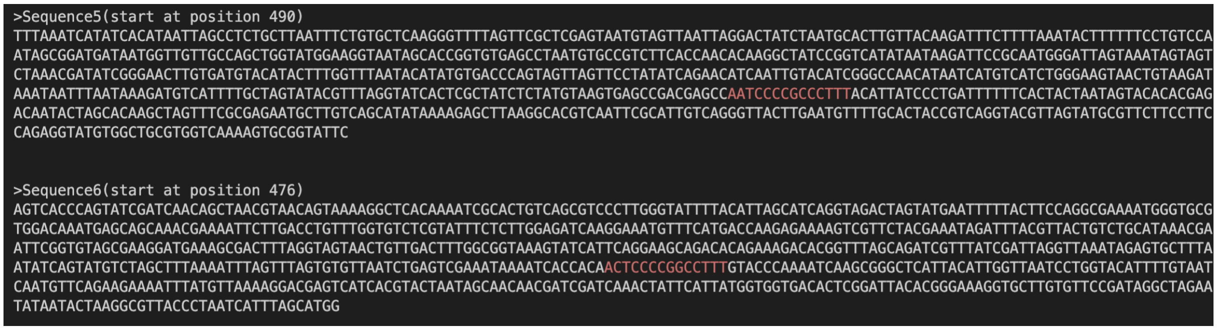


```
>Sequence6(start at position 206)
CCAAAAATTCGTGTGATTACGAGTCGTAGCAAATTTGGTCTGACTATGATCTAAATATTTCAGGCGGTACGTCTGCTTTCGAGTGCCTCTAATGGCTGGTTAGATAGACTAGCCGCTGGTAAACACACTATGATCT
CGGCTCTCCATTGATACTACGGCGATTGTTGGAGAGCCAGCAGCGATTGCAAATATCAGATCAGAGTAACGCGCGCGCATATTAGCATACGATAAGTCCCTAACTAATTATGACCTTTTGTAAAGTCAATTTCATT
ATATATGCTGTCTCTGGCACATGGATGATTTAGAGAAATTAGATTCAAGTCTGATTAACCATCAAACAGATCTTGAATCTAAAATTATCATCTCCTACGTACAAAATAAAAATACTAGGTAACTACAAGAATTCCG
ACGTTATATATGTTGCTCCATTAGAAACGCCATTCAGGATCACGTTACCGTGAAAAAAAGAAATCAGGAACTCTTCTCCTCTGCAGTCACGTCTATAGAAATTACACCATTAACCTTCCTAAGAACCGGGAGGTGG
GAATCTGTCACATATGAGAAGATATTTGCTCGATAATCAATACTCTAGGCTTCTAACTTTTTCCAC
```

2. Testing data 2:

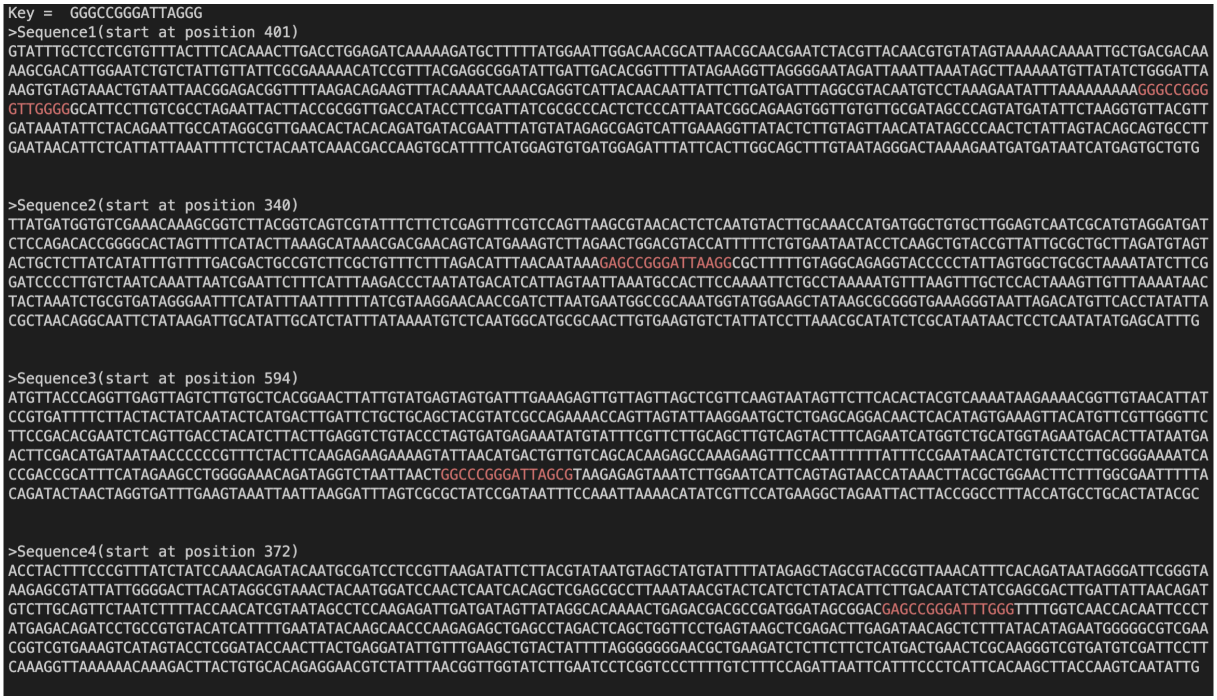


```
Key =  AATCCCCGGCCTTT
>Sequence1(start at position 217)
TCACTTAAACCGGCTAGGCTTTTCTGCCTGAAGTTTCAATGGATTGGTGTCAACGCGCAGGCATAGTTTTAGAAGAATTATTCGGGGGCAATGACAACCAACATCTCGGGTCTTGCCTAACCGGTCTACATGTTAA
TATAATGAATCACTGAAAACCCAGTGCTACACAATGGAATGTCCTTAATTCTGGCAGGTAATTAAAGGGAACGTATATACATTCCCCGGCCTTTAACGCAAAAAAACAGAAAAATAGGCGAATGAATCTTTTTTCT
GTGTATCGAAGAATGGCTTTGTGGAGGCATGCGTCATGCTAGCGTACAGGGTACTCTTACTATCCATATGGTTCACAAGACACTCGTTGTTTTCGAATTTACCCTTTATGCGCCAGTTTTCAACCACGCTTATGCC
CAACATCGTTACAACCAGACTGATACTAAATATATAAAGTCCGCCATACAAATGAAACTAGTTGGGAAATTATCGAGCATTCTATCAAGTCGACGACCACTAGTGAGTTACTAGAGCCGAGGGGTAACCATGATGC
TGCTAAAAATCTCTCGGTCGATGTAAGCGATTACACTCCTGTTACATCATAATCGTTTGTTATTCAGGGGTTGATCAACACTGGAAAACTTTTCACTTAAAGTATTATATACGACAGGGTACGTGTACCTATTAAA
CCTGTTTAAACTAAGTTCAGACTAGTTGGAAATGT


>Sequence2(start at position 605)
GTCTAGATCTTAGTTTTCGTTACTAAAGGGTCCACGTTTTATTTTTATGATCCATTGATCTTCTAAACACTGCAAGATTTACAACCAGGTAAACTTAGTGGTAGATCCTAATGCAGCGGGATTTTTTTTCTATAGT
CCTTGAGAGGAGAAATCGTCAGTCCAGATACCTTTGATGTTCTGATTGGAAGGATCGTTGGCCCCCCACCCTTAGATACGTATACTCAGTTCTATAAACGAGCTATTAAATATGAGATCCATAGATTGAAAAGGGT
AACAGAATTTGCCTGAATAGAAAAGACGGACAACTAGGTATCCTAAGTATAGTTGCGCGTCCGTATCAAGCTCTTTTTTATAGGTCTTGGTTTCTGTTGGTCGTAAAGCGCAGAACGGATTAGGGGGATGTACAAC
AATATTATTTAGTCATCTTTGAGTCACAATCTGCTACCTTACAGGAATTAAGATCGTCCTTTAATTTCCCTTGCATATATGTTGCATTTCTTCAACCTTTTAATTGCTCCCTAGGAGAAAGACAGATAACTTCTTA
CCCATATTTCATCGTTGGCAGCACGATCGCATGTCCCACGTGAATCATTGGTAAACCCTGAATCCCCGGCCCTTGTGGTTTGTGAGCGACAAAAGCTTTAATGGGAAATTCGCGCTCATAACTTGGTCCGAATACA
GATTCTAGCAACGTTCGTCTGAGTTTGATTTATAT


>Sequence3(start at position 139)
TTAGTGTTAGAAGGTGGAGTGACCTTAAATCAAAAACGATATTAATCGGAAGGAGTATTCAATGTAATGAAGTCGCAGGGTTAACGTGGGAATGGTGCTTCTGTCCAAACAGGTAAGGATATAAAGCTGTAACCGT
TTAATCCCGGGCCTTTCCCCAAGCGTACAGGGTGCATTTTGCAACAATTTCGGAGTCCAAAAACTCGCTGTTTTTGAAATTTATGCTCAAGGGCGAGTATTGAACCAAGCTTACGTCTAAGAACGTAGCAAGGTGA
CTCAAACAAAGTACATCTTGCCCGCGTTTCATATAAATTAAGTTAAAAGTCTATGGAATATAATAACATGTGGATGGCCAGTGGTCGGTTGTTACACCCCTACCGCAATGTTGAAAGTCCCGGATTAAACTGGCTA
AATTTATGCCGTGACACCCGTTATACTCCATTACCGTCTGTGGGTCATAGCTTGTTGTGGATTGGATTGTCATTCTCTCAGTGTATTACGCAGACTGGCGCACGGATCCCATATAAACTTATCATAGTTTATCTGA
TTCTACTTAGAAATGTAGCTAGGCTTTTGCCCACGCACCTGATTAGTCCTCGTTTGCTTTTTAGAACCGGATGAACTACAGAGCATTATAAGAATCTCTACTTGCTTTACAAAGTACTGGATCCTATTTCAGCGAG
ATGTTTTATCTAAATACAATGAGAGAAGTATTCGTC


>Sequence4(start at position 502)
AGGCCACATGGCTTTCTTGTTTTGGTCGGATCCATCATTGGCACCTGATCCCCCCATTCCATAGTGAGTTCTTCGTCTGAGTCATTGTATGCCAAATCGACAAACAAATAGCAGATCCAGTATATCTCTGGAAATT
ATAGACGTACAGATTGAAATCTTAAGTGAAATCACGCGTCTAAACTCAGCTTTATTTTAGTGGTCATGGGTTTCCTGGTCCCCCCGAGCCGTGGAACCGATTAGGACCATGTATAACAATACTTATTAGTCATCTT
TTAAACACAATCTTCCTGCTCAGTAGTATATGATTTTTGTTATAATTAGCCACCCTCATAAGTTACACTACTTCTGCGACCCAAATGTACCCTTACCACGAAGACAAGATTGTCCGATCCTATATTACGATTTTGG
TAAGGGGTTTGCAAGTCCCACCCTAAACGATATTGAAGACTTAGGTTATACAAGCATAAGTACTATATATACGAGTTCCTGTTCTTAACCTGTAACCCCCGGCCTTTGATCGAATGTAGAATTATGCATCGTACCA
CTATATTCATGTCATCTAGGACGGGCGCAAAGGATATATAATTCAATTAAGAATACCTTATATTATTATACATCTACCGGTCACCAGCCAACAATGTGCAGATGGCGTTACGACTTACTGAGCCTGATTTCACCGC
TTTAAATACCACACACTGGGCAATACGAGGTAAAGCC
```

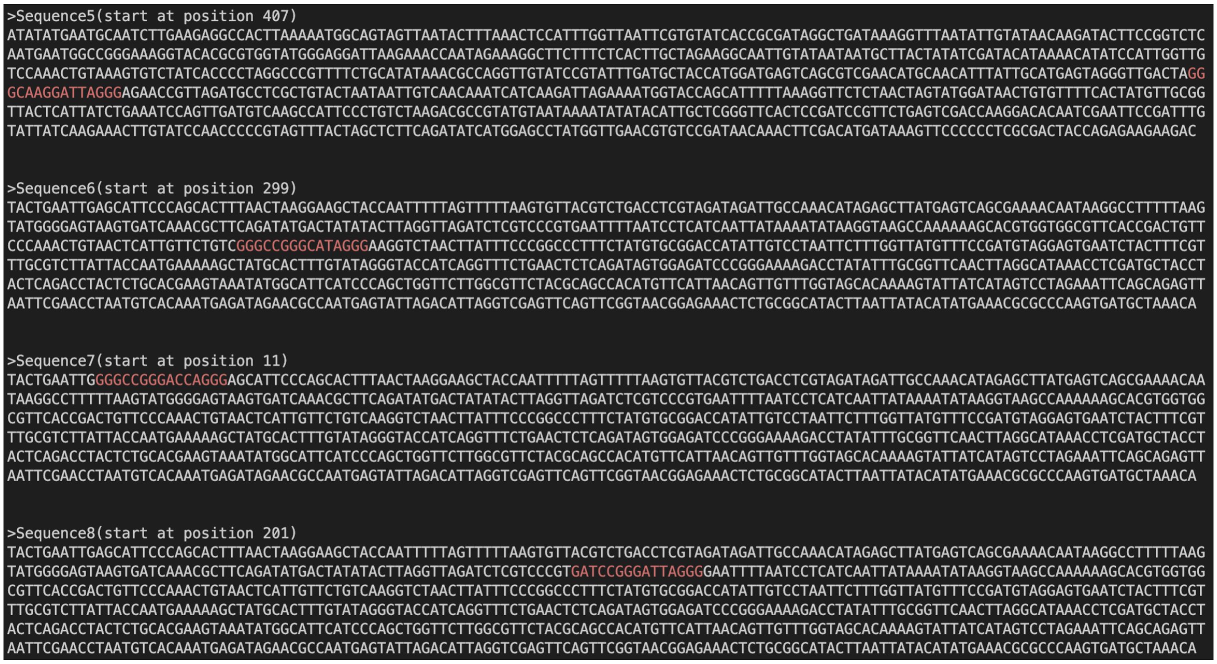


```
>Sequence5(start at position 490)
TTTAAATCATATCACATAATTAGCCTCTGCTTAATTTCTGTGCTCAAGGGTTTTAGTTCGCTCGAGTAATGTAGTTAATTAGGACTATCTAATGCACTTGTTACAAGATTTCTTTTAAATACTTTTTTCCTGTCCA
ATAGCGGATGATAATGGTTGTTGCCAGCTGGTATGGAAGGTAATAGCACCGGTGTGAGCCTAATGTGCCGTCTTCACCAACACAAGGCTATCCGGTCATATAATAAGATTCCGCAATGGGATTAGTAAATAGTAGT
CTAAACGATATCGGGAACTTGTGATGTACATACTTTGGTTTAATACATATGTGACCCAGTAGTTAGTTCCTATATCAGAACATCAATTGTACATCGGGCCAACATAATCATGTCATCTGGGAAGTAACTGTAAGAT
AAATAATTTAATAAAGATGTCATTTTGCTAGTATACGTTTAGGTATCACTCGCTATCTCTATGTAAGTGAGCCGACGAGCCAATCCCCGCCCTTTACATTATCCCTGATTTTTTCACTACTAATAGTACACACGAG
ACAATACTAGCACAAGCTAGTTTCGCGAGAATGCTTGTCAGCATATAAAAGAGCTTAAGGCACGTCAATTCGCATTGTCAGGGTTACTTGAATGTTTTGCACTACCGTCAGGTACGTTAGTATGCGTTCTTCCTTC
CAGAGGTATGTGGCTGCGTGGTCAAAAGTGCGGTATTC


>Sequence6(start at position 476)
AGTCACCCAGTATCGATCAACAGCTAACGTAACAGTAAAAGGCTCACAAAATCGCACTGTCAGCGTCCCTTGGGTATTTTACATTAGCATCAGGTAGACTAGTATGAATTTTTACTTCCAGGCGAAAATGGGTGCG
TGGACAAATGAGCAGCAAACGAAAATTCTTGACCTGTTTGGTGTCTCGTATTTCTCTTGGAGATCAAGGAAATGTTTCATGACCAAGAGAAAAGTCGTTCTACGAAATAGATTTACGTTACTGTCTGCATAAACGA
ATTCGGTGTAGCGAAGGATGAAAGCGACTTTAGGTAGTAACTGTTGACTTTGGCGGTAAAGTATCATTCAGGAAGCAGACACAGAAAGACACGGTTTAGCAGATCGTTTATCGATTAGGTTAAATAGAGTGCTTTA
ATATCAGTATGTCTAGCTTTAAAATTTAGTTTAGTGTGTTAATCTGAGTCGAAATAAAATCACCACAACTCCCCGGCCTTTGTACCCAAAATCAAGCGGGCTCATTACATTGGTTAATCCTGGTACATTTTGTAAT
CAATGTTCAGAAGAAAATTTATGTTAAAAGGACGAGTCATCACGTACTAATAGCAACAACGATCGATCAAACTATTCATTATGGTGGTGACACTCGGATTACACGGGAAAGGTGCTTGTGTTCCGATAGGCTAGAA
TATAATACTAAGGCGTTACCCTAATCATTTAGCATGG
```

3. Testing data 3:



```
Key =  GGGCCGGGATTAGGG
>Sequence1(start at position 401)
GTATTTGCTCCTCGTGTTTACTTTCACAAACTTGACCTGGAGATCAAAAAGATGCTTTTTATGGAATTGGACAACGCATTAACGCAACGAATCTACGTTACAACGTGTATAGTAAAAACAAAATTGCTGACGACAA
AAGCGACATTGGAATCTGTCTATTGTTATTCGCGAAAAACATCCGTTTACGAGGCGGATATTGATTGACACGGTTTTATAGAAGGTTAGGGGAATAGATTAAATTAAATAGCTTAAAAATGTTATATCTGGGATTA
AAGTGTAGTAAACTGTAATTAACGGAGACGGTTTAAGACAGAAGTTTACAAAATCAAACGAGGTCATTACAACAATTATTCTTGATGATTTAGGCGTACAATGTCCTAAAGAATATTTAAAAAAAAAGGGCCGGG
GTTGGGGGCATTCCTTGTCGCCTAGAATTACTTACCGCGGTTGACCATACCTTCGATTATCGCGCCCACTCTCCCATTAATCGGCAGAAGTGGTTGTGTTGCGATAGCCCAGTATGATATTCTAAGGTGTTACGTT
GATAAATATTCTACAGAATTGCCATAGGCGTTGAACACTACACAGATGATACGAATTTATGTATAGAGCGAGTCATTGAAAGGTTATACTCTTGTAGTTAACATATAGCCCAACTCTATTAGTACAGCAGTGCCTT
GAATAACATTCTCATTATTAAATTTTCTCTACAATCAAACGACCAAGTGCATTTTCATGGAGTGTGATGGAGATTTATTCACTTGGCAGCTTTGTAATAGGGACTAAAAGAATGATGATAATCATGAGTGCTGTG


>Sequence2(start at position 340)
TTATGATGGTGTCGAAACAAAGCGGTCTTACGGTCAGTCGTATTTCTTCTCGAGTTTCGTCCAGTTAAGCGTAACACTCTCAATGTACTTGCAAACCATGATGGCTGTGCTTGGAGTCAATCGCATGTAGGATGAT
CTCCAGACACCGGGGCACTAGTTTTCATACTTAAAGCATAAACGACGAACAGTCATGAAAGTCTTAGAACTGGACGTACCATTTTTCTGTGAATAATACCTCAAGCTGTACCGTTATTGCGCTGCTTAGATGTAGT
ACTGCTCTTATCATATTTGTTTTGACGACTGCCGTCTTCGCTGTTTCTTTAGACATTTAACAATAAAGAGCCGGGATTAAGGCGCTTTTTGTAGGCAGAGGTACCCCCTATTAGTGGCTGCGCTAAAATATCTTCG
GATCCCCTTGTCTAATCAAATTAATCGAATTCTTTCATTTAAGACCCTAATATGACATCATTAGTAATTAAATGCCACTTCCAAAATTCTGCCTAAAAATGTTTAAGTTTGCTCCACTAAAGTTGTTTAAAATAAC
TACTAAATCTGCGTGATAGGGAATTTCATATTTAATTTTTTATCGTAAGGAACAACCGATCTTAATGAATGGCCGCAAATGGTATGGAAGCTATAAGCGCGGGTGAAAGGGTAATTAGACATGTTCACCTATATTA
CGCTAACAGGCAATTCTATAAGATTGCATATTGCATCTATTTATAAAATGTCTCAATGGCATGCGCAACTTGTGAAGTGTCTATTATCCTTAAACGCATATCTCGCATAATAACTCCTCAATATATGAGCATTTG


>Sequence3(start at position 594)
ATGTTACCCAGGTTGAGTTAGTCTTGTGCTCACGGAACTTATTGTATGAGTAGTGATTTGAAAGAGTTGTTAGTTAGCTCGTTCAAGTAATAGTTCTTCACACTACGTCAAAATAAGAAAACGGTTGTAACATTAT
CCGTGATTTTCTTACTACTATCAATACTCATGACTTGATTCTGCTGCAGCTACGTATCGCCAGAAAACCAGTTAGTATTAAGGAATGCTCTGAGCAGGACAACTCACATAGTGAAAGTTACATGTTCGTTGGGTTC
TTCCGACACGAATCTCAGTTGACCTACATCTTACTTGAGGTCTGTACCCTAGTGATGAGAAATATGTATTTCGTTCTTGCAGCTTGTCAGTACTTTCAGAATCATGGTCTGCATGGTAGAATGACACTTATAATGA
ACTTCGACATGATAATAACCCCCCGTTTCTACTTCAAGAGAAGAAAAGTATTAACATGACTGTTGTCAGCACAAGAGCCAAAGAAGTTTCCAATTTTTTATTTCCGAATAACATCTGTCTCCTTGCGGGAAAATCA
CCGACCGCATTTCATAGAAGCCTGGGGAAACAGATAGGTCTAATTAACTGGCCCGGGATTAGCGTAAGAGAGTAAATCTTGGAATCATTCAGTAGTAACCATAAACTTACGCTGGAACTTCTTTGGCGAATTTTTA
CAGATACTAACTAGGTGATTTGAAGTAAATTAATTAAGGATTTAGTCGCGCTATCCGATAATTTCCAAATTAAAACATATCGTTCCATGAAGGCTAGAATTACTTACCGGCCTTTACCATGCCTGCACTATACGC


>Sequence4(start at position 372)
ACCTACTTTCCCGTTTATCTATCCAAACAGATACAATGCGATCCTCCGTTAAGATATTCTTACGTATAATGTAGCTATGTATTTTATAGAGCTAGCGTACGCGTTAAACATTTCACAGATAATAGGGATTCGGGTA
AAGAGCGTATTATTGGGGACTTACATAGGCGTAAACTACAATGGATCCAACTCAATCACAGCTCGAGCGCCTTAAATAACGTACTCATCTCTATACATTCTTGACAATCTATCGAGCGACTTGATTATTAACAGAT
GTCTTGCAGTTCTAATCTTTTACCAACATCGTAATAGCCTCCAAGAGATTGATGATAGTTATAGGCACAAAACTGAGACGACGCCGATGGATAGCGGACGAGCCGGGATTTGGGTTTTGGTCAACCACAATTCCCT
ATGAGACAGATCCTGCCGTGTACATCATTTTGAATATACAAGCAACCCAAGAGAGCTGAGCCTAGACTCAGCTGGTTCCTGAGTAAGCTCGAGACTTGAGATAACAGCTCTTTATACATAGAATGGGGGCGTCGAA
CGGTCGTGAAAGTCATAGTACCTCGGATACCAACTTACTGAGGATATTGTTTGAAGCTGTACTATTTTAGGGGGGGAACGCTGAAGATCTCTTCTTCTCATGACTGAACTCGCAAGGGTCGTGATGTCGATTCCTT
CAAAGGTTAAAAAACAAAGACTTACTGTGCACAGAGGAACGTCTATTTAACGGTTGGTATCTTGAATCCTCGGTCCCTTTTGTCTTTCCAGATTAATTCATTTCCCTCATTCACAAGCTTACCAAGTCAATATTG
```



```
>Sequence5(start at position 407)
ATATATGAATGCAATCTTGAAGAGGCCACTTAAAAATGGCAGTAGTTAATACTTTAAACTCCATTTGGTTAATTCGTGTATCACCGCGATAGGCTGATAAAGGTTTAATATTGTATAACAAGATACTTCCGGTCTC
AATGAATGGCCGGGAAAGGTACACGCGTGGTATGGGAGGATTAAGAAACCAATAGAAAGGCTTCTTTCTCACTTGCTAGAAGGCAATTGTATAATAATGCTTACTATATCGATACATAAAACATATCCATTGGTTG
TCCAAACTGTAAAGTGTCTATCACCCCTAGGCCCGTTTTCTGCATATAAACGCCAGGTTGTATCCGTATTTGATGCTACCATGGATGAGTCAGCGTCGAACATGCAACATTTATTGCATGAGTAGGGTTGACTAGG
GCAAGGATTAGGGAGAACCGTTAGATGCCTCGCTGTACTAATAATTGTCAACAAATCATCAAGATTAGAAAATGGTACCAGCATTTTTAAAGGTTCTCTAACTAGTATGGATAACTGTGTTTTCACTATGTTGCGG
TTACTCATTATCTGAAATCCAGTTGATGTCAAGCCATTCCCTGTCTAAGACGCCGTATGTAATAAAATATATACATTGCTCGGGTTCACTCCGATCCGTTCTGAGTCGACCAAGGACACAATCGAATTCCGATTTG
TATTATCAAGAAACTTGTATCCAACCCCCGTAGTTTACTAGCTCTTCAGATATCATGGAGCCTATGGTTGAACGTGTCCGATAACAAACTTCGACATGATAAAGTTCCCCCTCGCGACTACCAGAGAAGAAGAC


>Sequence6(start at position 299)
TACTGAATTGAGCATTCCCAGCACTTTAACTAAGGAAGCTACCAATTTTTAGTTTTTAAGTGTTACGTCTGACCTCGTAGATAGATTGCCAAACATAGAGCTTATGAGTCAGCGAAAACAATAAGGCCTTTTTAAG
TATGGGGAGTAAGTGATCAAACGCTTCAGATATGACTATATACTTAGGTTAGATCTCGTCCCGTGAATTTTAATCCTCATCAATTATAAAATATAAGGTAAGCCAAAAAAGCACGTGGTGGCGTTCACCGACTGTT
CCCAAACTGTAACTCATTGTTCTGTCGGGCCGGGCATAGGGAAGGTCTAACTTATTTCCCGGCCCTTTCTATGTGCGGACCATATTGTCCTAATTCTTTGGTTATGTTTCCGATGTAGGAGTGAATCTACTTTCGT
TTGCGTCTTATTACCAATGAAAAAGCTATGCACTTTGTATAGGGTACCATCAGGTTTCTGAACTCTCAGATAGTGGAGATCCCGGGAAAAGACCTATATTTGCGGTTCAACTTAGGCATAAACCTCGATGCTACCT
ACTCAGACCTACTCTGCACGAAGTAAATATGGCATTCATCCCAGCTGGTTCTTGGCGTTCTACGCAGCCACATGTTCATTAACAGTTGTTTGGTAGCACAAAAGTATTATCATAGTCCTAGAAATTCAGCAGAGTT
AATTCGAACCTAATGTCACAAATGAGATAGAACGCCAATGAGTATTAGACATTAGGTCGAGTTCAGTTCGGTAACGGAGAAACTCTGCGGCATACTTAATTATACATATGAAACGCGCCCAAGTGATGCTAAACA


>Sequence7(start at position 11)
TACTGAATTGGGGCCGGGACCAGGGAGCATTCCCAGCACTTTAACTAAGGAAGCTACCAATTTTTAGTTTTTAAGTGTTACGTCTGACCTCGTAGATAGATTGCCAAACATAGAGCTTATGAGTCAGCGAAAACAA
TAAGGCCTTTTTAAGTATGGGGAGTAAGTGATCAAACGCTTCAGATATGACTATATACTTAGGTTAGATCTCGTCCCGTGAATTTTAATCCTCATCAATTATAAAATATAAGGTAAGCCAAAAAAGCACGTGGTGG
CGTTCACCGACTGTTCCCAAACTGTAACTCATTGTTCTGTCAAGGTCTAACTTATTTCCCGGCCCTTTCTATGTGCGGACCATATTGTCCTAATTCTTTGGTTATGTTTCCGATGTAGGAGTGAATCTACTTTCGT
TTGCGTCTTATTACCAATGAAAAAGCTATGCACTTTGTATAGGGTACCATCAGGTTTCTGAACTCTCAGATAGTGGAGATCCCGGGAAAAGACCTATATTTGCGGTTCAACTTAGGCATAAACCTCGATGCTACCT
ACTCAGACCTACTCTGCACGAAGTAAATATGGCATTCATCCCAGCTGGTTCTTGGCGTTCTACGCAGCCACATGTTCATTAACAGTTGTTTGGTAGCACAAAAGTATTATCATAGTCCTAGAAATTCAGCAGAGTT
AATTCGAACCTAATGTCACAAATGAGATAGAACGCCAATGAGTATTAGACATTAGGTCGAGTTCAGTTCGGTAACGGAGAAACTCTGCGGCATACTTAATTATACATATGAAACGCGCCCAAGTGATGCTAAACA


>Sequence8(start at position 201)
TACTGAATTGAGCATTCCCAGCACTTTAACTAAGGAAGCTACCAATTTTTAGTTTTTAAGTGTTACGTCTGACCTCGTAGATAGATTGCCAAACATAGAGCTTATGAGTCAGCGAAAACAATAAGGCCTTTTTAAG
TATGGGGAGTAAGTGATCAAACGCTTCAGATATGACTATATACTTAGGTTAGATCTCGTCCCGTGATCCGGGATTAGGGGAATTTTAATCCTCATCAATTATAAAATATAAGGTAAGCCAAAAAAGCACGTGGTGG
CGTTCACCGACTGTTCCCAAACTGTAACTCATTGTTCTGTCAAGGTCTAACTTATTTCCCGGCCCTTTCTATGTGCGGACCATATTGTCCTAATTCTTTGGTTATGTTTCCGATGTAGGAGTGAATCTACTTTCGT
TTGCGTCTTATTACCAATGAAAAAGCTATGCACTTTGTATAGGGTACCATCAGGTTTCTGAACTCTCAGATAGTGGAGATCCCGGGAAAAGACCTATATTTGCGGTTCAACTTAGGCATAAACCTCGATGCTACCT
ACTCAGACCTACTCTGCACGAAGTAAATATGGCATTCATCCCAGCTGGTTCTTGGCGTTCTACGCAGCCACATGTTCATTAACAGTTGTTTGGTAGCACAAAAGTATTATCATAGTCCTAGAAATTCAGCAGAGTT
AATTCGAACCTAATGTCACAAATGAGATAGAACGCCAATGAGTATTAGACATTAGGTCGAGTTCAGTTCGGTAACGGAGAAACTCTGCGGCATACTTAATTATACATATGAAACGCGCCCAAGTGATGCTAAACA
```

Other result is included in the following file (output)

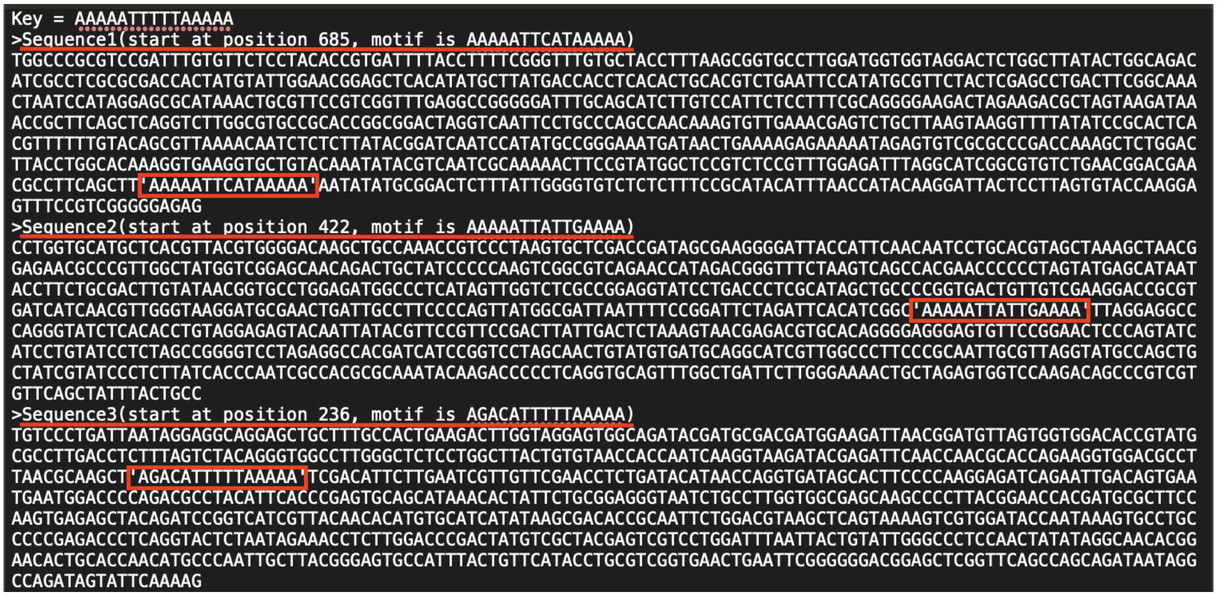
The testing data TA provide is all in the **testing_data** folder, the result is in the **output** folder.

The following is the sample output of testing data 1

- Key1

  The first line is the final motif.

  >SequenceN(start at position x, motif is m) and the motif of every sequence is quote with ‘’.



```
Key = AAAAATTTTTAAAAA
>Sequence1(start at position 685, motif is AAAAATTCATAAAAA)
TGGCCCGCGTCCGATTTGTGTTCTCCTACACCGTGATTTTACCTTTTCGGGTTTGTGCTACCTTTAAGCGGTGCCTTGGATGGTGGTAGGACTCTGGCTTATACTGGCAGAC
ATCGCCTCGCGCGACCACTATGTATTGGAACGGAGCTCACATATGCTTATGACCACCTCACACTGCACGTCTGAATTCCATATGCGTTCTACTCGAGCCTGACTTCGGCAAA
CTAATCCATAGGAGCGCATAAACTGCGTTCCGTCGGTTTGAGGCCGGGGGATTTGCAGCATCTTGTCCATTCTCCTTTCGCAGGGGAAGACTAGAAGACGCTAGTAAGATAA
ACCGCTTCAGCTCAGGTCTTGGCGTGCCGCACCGGCGGACTAGGTCAATTCCTGCCCAGCCAACAAAGTGTTGAAACGAGTCTGCTTAAGTAAGGTTTTATATCCGCACTCA
CGTTTTTTGTACAGCGTTAAAACAATCTCTCTTATACGGATCAATCCATATGCCGGGAAATGATAACTGAAAAGAGAAAAATAGAGTGTCGCGCCCGACCAAAGCTCTGGAC
TTACCTGGCACAAAGGTGAAGGTGCTGTACAAATATACGTCAATCGCAAAAACTTCCGTATGGCTCCGTCTCCGTTTGGAGATTTAGGCATCGGCGTGTCTGAACGGACGAA
CGCCTTCAGCTT'AAAAATTCATAAAAA'AATATATGCGGACTCTTTATTGGGGTGTCTCTCTTTCCGCATACATTTAACCATACAAGGATTACTCCTTAGTGTACCAAGGA
GTTTCCGTCGGGGGAGAG
>Sequence2(start at position 422, motif is AAAAATTATTGAAAA)
CCTGGTGCATGCTCACGTTACGTGGGGACAAGCTGCCAAACCGTCCCTAAGTGCTCGACCGATAGCGAAGGGGATTACCATTCAACAATCCTGCACGTAGCTAAAGCTAACG
GAGAACGCCCGTTGGCTATGGTCGGAGCAACAGACTGCTATCCCCCAAGTCGGCGTCAGAACCATAGACGGGTTTCTAAGTCAGCCACGAACCCCCCTAGTATGAGCATAAT
ACCTTCTGCGACTTGTATAACGGTGCCTGGAGATGGCCCTCATAGTTGGTCTCGCCGGAGGTATCCTGACCCTCGCATAGCTGCCCCGGTGACTGTTGTCGAAGGACCGCGT
GATCATCAACGTTGGGTAAGGATGCGAACTGATTGCCTTCCCCAGTTATGGCGATTAATTTTCCGGATTCTAGATTCACATCGGC'AAAAATTATTGAAAA'TTAGGAGGCC
CAGGGTATCTCACACCTGTAGGAGAGTACAATTATACGTTCCGTTCCGACTTATTGACTCTAAAGTAACGAGACGTGCACAGGGGAGGGAGTGTTCCGGAACTCCCAGTATC
ATCCTGTATCCTCTAGCCGGGGTCCTAGAGGCCACGATCATCCGGTCCTAGCAACTGTATGTGATGCAGGCATCGTTGGCCCTTCCCGCAATTGCGTTAGGTATGCCAGCTG
CTATCGTATCCCTCTTATCACCCAATCGCCACGCGCAAATACAAGACCCCCTCAGGTGCAGTTTGGCTGATTCTTGGGAAAACTGCTAGAGTGGTCCAAGACAGCCCGTCGT
GTTCAGCTATTTACTGCC
>Sequence3(start at position 236, motif is AGACATTTTTAAAAA)
TGTCCCTGATTAATAGGAGGCAGGAGCTGCTTTGCCACTGAAGACTTGGTAGGAGTGGCAGATACGATGCGACGATGGAAGATTAACGGATGTTAGTGGTGGACACCGTATG
CGCCTTGACCTCTTTAGTCTACAGGGTGGCCTTGGGCTCTCCTGGCTTACTGTGTAACCACCAATCAAGGTAAGATACGAGATTCAACCAACGCACCAGAAGGTGGACGCCT
TAACGCAAGCT'AGACATTTTTAAAAA'TCGACATTCTTGAATCGTTGTTCGAACCTCTGATACATAACCAGGTGATAGCACTTCCCCAAGGAGATCAGAATTGACAGTGAA
TGAATGGACCCCAGACGCCTACATTCACCCGAGTGCAGCATAAACACTATTCTGCGGAGGGTAATCTGCCTTGGTGGCGAGCAAGCCCCTTACGGAACCACGATGCGCTTCC
AAGTGAGAGCTACAGATCCGGTCATCGTTACAACACATGTGCATCATATAAGCGACACCGCAATTCTGGACGTAAGCTCAGTAAAAGTCGTGGATACCAATAAAGTGCCTGC
CCCCGAGACCCTCAGGTACTCTAATAGAAACCTCTTGGACCCGACTATGTCGCTACGAGTCGTCCTGGATTTAATTACTGTATTGGGCCCTCCAACTATATAGGCAACACGG
AACACTGCACCAACATGCCCAATTGCTTACGGGAGTGCCATTTACTGTTCATACCTGCGTCGGTGAACTGAATTCGGGGGGACGGAGCTCGGTTCAGCCAGCAGATAATAGG
CCAGATAGTATTCAAAAG
```