Q1 ⇒ $p = 0.25$ $\quad 1-p = 0.75$

$n = 1000$ $\quad x = 240$

Central limit theorem:

$$\phi(z) = \phi\left(\left[\frac{x-\mu}{\sigma}\right]\right) = \phi\left(\left[\frac{240-\mu}{\sigma}\right]\right)$$

$$\sigma = \sqrt{np(1-p)} = 13.69$$

$$\mu = \text{mean} = n.p = E[x] = 250$$

So

$$z = \left(\frac{240-250}{13.69}\right) \approx -0.7303$$

So $\phi(z) = \phi(-0.7303) = 0.7673$

Manual Approximate calculation = 0.7673
Simulate Probability = 0.7767
Calculated using scipy = 0.7776521016

OR

codes are provides as screenshots

[28]

```python
# Question 1
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import random
import math

p = 0.25
n = 1000

BinomScipy = stats.binom(n,p)

random.seed(0)

arr = ['A','T','C','G'] # Since p = 0.25 , here also p(A) = 0.25

x = []
y = []

temp = 0
temp_arr = []
temp_1 = 0
for j in range(0,10000):
    temp = 0
    for k in range(0,1000):
        temp += 1 if (random.choice(arr) == 'A') else 0
    if(temp >= 240):
        temp_1 += 1

print("Simulated Probability of A appearing atleast 240 time  : " + str(temp_1/10000) )
print("Calculated Probability of A appearing atleast 240 time : " + str(1-BinomScipy.cdf(239.9999)))
```

```
Simulated Probability of A appearing atleast 240 time  : 0.7767
Calculated Probability of A appearing atleast 240 time : 0.7776521016977749
```

Q # 2 ⇒ $p = 0.3 \quad n = 10$

$$P[x = k] = ({}^{n}C_{k}) \cdot p^{k} \cdot (1-p)^{n-k}$$

$$P(x = 0) = {}^{10}C_{0} \times (0.3)^{0} \cdot (1 - 0.3)^{10-0}$$

$$= \underline{0.0282}$$

$$P(x = 2) = {}^{10}C_{2} \cdot (0.3)^{2} \cdot (1 - 0.3)^{8}$$

$$= \underline{0.2334}$$

$$E[X] = n \times p = 10 \times 0.3 = \underline{3}$$

$$Var[X] = n \times p \times (1-p) = 10 \times (0.3) \times (1 - 0.3)$$

$$= \underline{2.099}$$

Q #3 ⇒ a) k-mer based approach is used to check the similarities and dissimilarities in different biological sequences.

b) k-mer has extensive use in biotechnology in controlling translational efficiency.

c) They are also used to create attenuated vaccines

d) They also have application in identifying heterogenous gene expression

e) They are used in construction of De Bruijn graphs.

f) k-mer distributions are well-preserved. Because of this reason bacterial genomes can be clustered into natural groups according to k-mer distribution similarities.

g) Observed frequencies of k-words can be used to make inferences about DNA sequences.

Q # 6 =>

	G	G	C	T	G	C	A	A	C	T	A	G	C	T	C
G	x	x			x							x			
G	x	x			x							x			
G	x	x			x							x			
T				x						x				x	
A							x	x			x				
A							x	x			x				
G	x	x			x							x			
C			x			x			x				x		x
T				x						x				x	
T				x						x				x	
G	x	x			x							x			
C			x			x			x				x		x

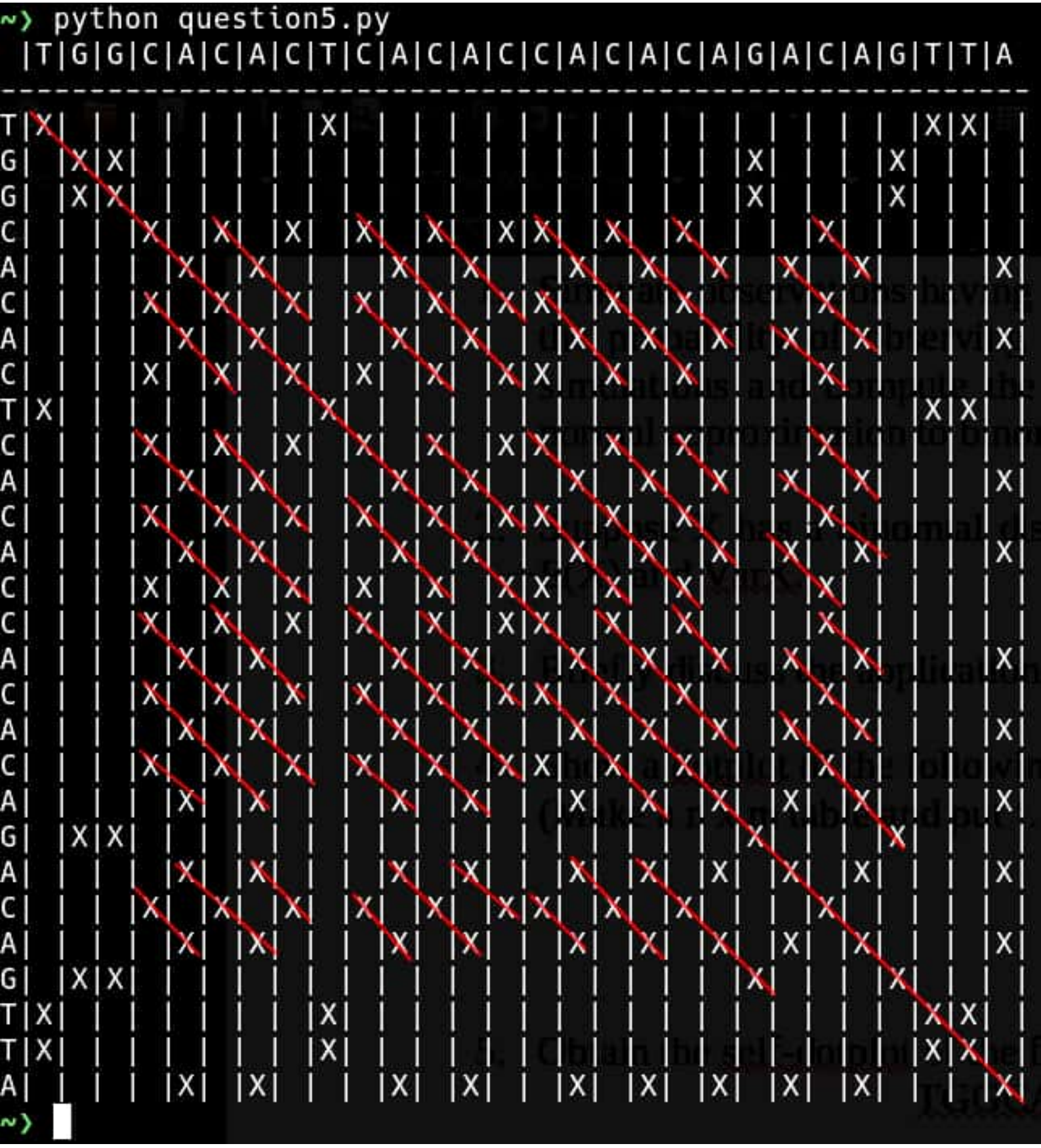
Conserved Regions: The region of similarity between 2 genomes

We use dotplot to identify conserved regions that are showing a match by x wherever a base matches or a group of Bases matches

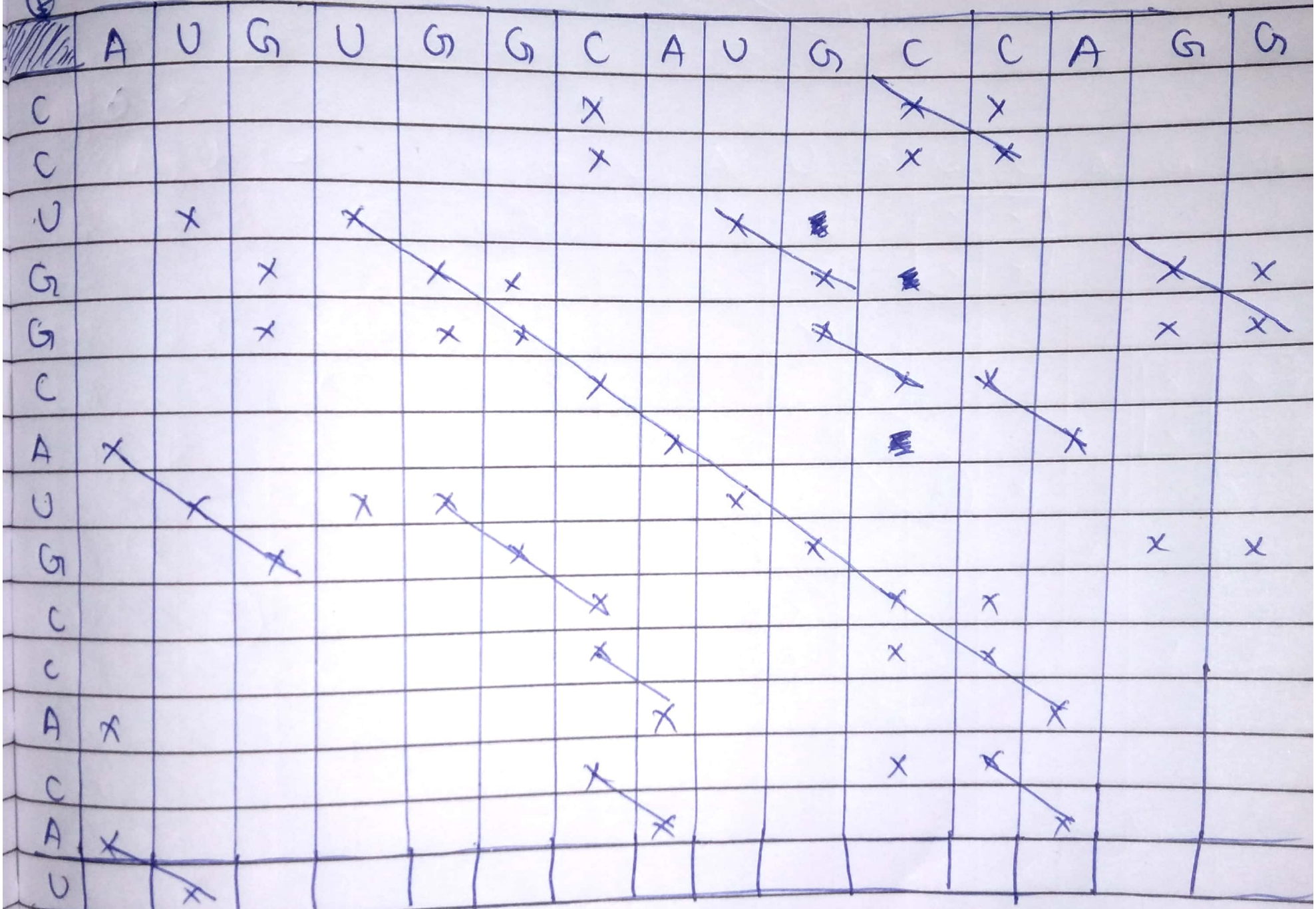
1.ipynb M question5.py ×

home > bhaeyan > question5.py > ...

```python
# Question #5 Code
import numpy
import matplotlib.pyplot as plt

def df(x,y):
    return 0 if x == y else 1

def M(seq1,seq2,i,j,k):
    return sum(df(x,y) for x,y in zip(seq1[i:i+k],seq2[j:j+k]))

def ML(seq1,seq2,k):
    n = len(seq1)
    m = len(seq2)
    return [[M(seq1,seq2,i,j,k) for j in range(m-k+1)] for i in range(n-k+1)]

def drawMatrix(M,t, seq1, seq2, nonblank = chr(0x25A0), blank = ' |'):
    print('  ',end="")
    for i in seq2:
        print("|" + i,end="")
    print("")
    print('-'*(2 + 2*len(seq2)))
    for label,row in zip(seq1,M):
        line = ''.join('X|' if s < t else blank for s in row)
        print(label + '|' + line)
        # print('-'*(2 + 2*len(seq2)))

def dotplot(seq1,seq2,k = 1,t = 1):
    M = ML(seq1,seq2,k)
    drawMatrix(M, t, seq1,seq2)

# Sequence :
seqy = seqx = 'TGGCACACTCACACCACACAGACAGTTA'

# Ploting Function
dotplot(seqx,seqy)
```


~) python question5.py
|T|G|G|C|A|C|A|C|T|C|A|C|A|C|C|A|C|A|C|A|G|A|C|A|G|T|T|A
T|X| | | | | | | |X| | | | | | | | | | | | | | | |X|X| |
G| |X|X| | | | | | | | | | | | | | | | | | | |X| | |X| | | |
G| |X|X| | | | | | | | | | | | | | | | | | | |X| | |X| | | |
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
T|X| | | | | | | |X| | | | | | | | | | | | | | | |X|X| |
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
G| |X|X| | | | | | | | | | | | | | | | | | | |X| | |X| | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
C| | | |X| |X| |X| |X| |X| |X|X| |X| |X| | | |X| | | | | |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
G| |X|X| | | | | | | | | | | | | | | | | | | |X| | |X| | | |
T|X| | | | | | | |X| | | | | | | | | | | | | | | |X|X| |
T|X| | | | | | | |X| | | | | | | | | | | | | | | |X|X| |
A| | | | |X| |X| | | |X| |X| | |X| |X| |X| |X| |X| | | |X|
~)

Q #6



Diagnols representing complementary regions are shown from top - ~~right~~ left and bottom right diagonals directions.

RNA & its reverse used

Palindrome sequence

i) AU

ii) GC

iii) UGG CAUGCCA

AUGUGGCAUGCCAGG

RNA ⇒ AUGUGGCAUGCCAGG

Reverse ⇒ CCUGGCAUGCCACAU

complement