Bio-informatics Assignment Submission 01

Student:
Tajkia Nuri Ananna
Reg. No: 2016331092
Email: tazkiaaltaaf@gmail.com

Instructor:
Moqsadur Rahman
Assistant Professor
Email: moqsadsust@gmail.com

Date of Submission: November 23, 2020

Department of Computer Science and Engineering

Contents

1 Biological Section

1.1 Questions for One Marks

1.1.1 Q1: Write down the chemical composition of a cell .

The chemical composition (by weight) of a cell is divided into the following three parts:

- 70% water
- 7% small molecules: salts, lipids, amino acids, nucleotides
- 23% macro-molecules: proteins, polysaccharides, lipids .

1.1.2 Q2: By which nucleotide DNA differs from RNA?

Thymine.

1.1.3 Q3: What are the termination codons?

UAA, UAG UGA.

1.1.4 Q5: What is a genome?

A genome is an organism's complete set of DNA, including all of its genes.

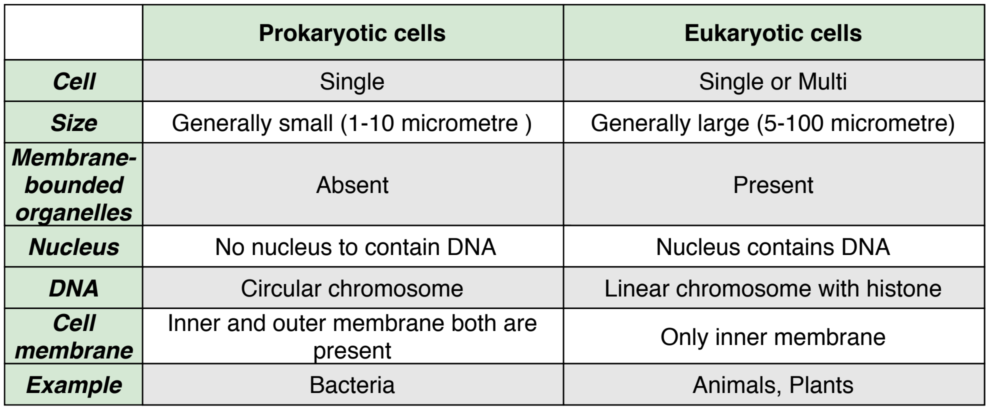
1.1.5 Q5: What is punnet square?

Punnet square is a tool to do genetic crosses.

1.2 Questions for Five Marks

1.2.1 Q1:What are the differences between Prokaryotic and Eukaryotic cells?

The differences between Prokaryotic and Eukaryotic cells are shown below (Fig. 1):



	Prokaryotic cells	Eukaryotic cells
Cell	Single	Single or Multi
Size	Generally small (1-10 micrometre)	Generally large (5-100 micrometre)
Membrane-bounded organelles	Absent	Present
Nucleus	No nucleus to contain DNA	Nucleus contains DNA
DNA	Circular chromosome	Linear chromosome with histone
Cell membrane	Inner and outer membrane both are present	Only inner membrane
Example	Bacteria	Animals, Plants

Figure 1: Prokaryotic vs Eukaryotic cells

1.2.2 Q2: What is a cell cycle? Explain the cell cycle process.

Cell cycle: Every living cell go through a series of phases such as they grow, copy their chromosomes and then divide to form new cells). This whole process is known as a cell cycle.
Cell cycle process: The cell cycle process has 4 phases. These are:

1. **G1 phase:** This is known as the growth phase. In this phase the cell grows and cellular contents excluding chromosome duplicates.

2. **S phase:** This is known as the DNA synthesis phase. Each of the 46 chromosomes are duplicated in this phase. After completion of this phase each cell contains two sister chromatids connected by centromere.

3. **G2 phase:** This is a preparation phase for the cell division. The cell double checks the chromosomes for any kind of error and repaires them if needed.

4. **M phase:** The cell conducts two important tasks in this phase : 1) The copied chromosomes are separated to form two new sets (Mitosis) 2) The cell is then divided into two cells (cytokinesis)

G0 phase: Cells who are not dividing leave the cell cycle (exit in G1 phase) and stays in G0 (Resting state).

The G1, S and G2 phase are known as interphase and the M phase is known as the mitosis phase. The cell spends 10% of it's time in mitosis and 90% in interphase. The cell cycle process is shown in Fig. 2.

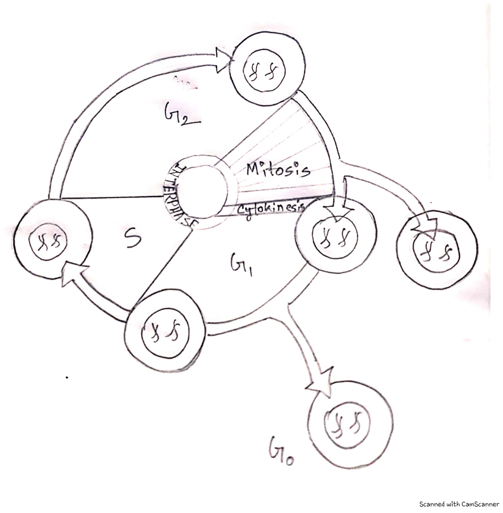


Figure 2: Cross of heterozygous purple and homozygous white flower

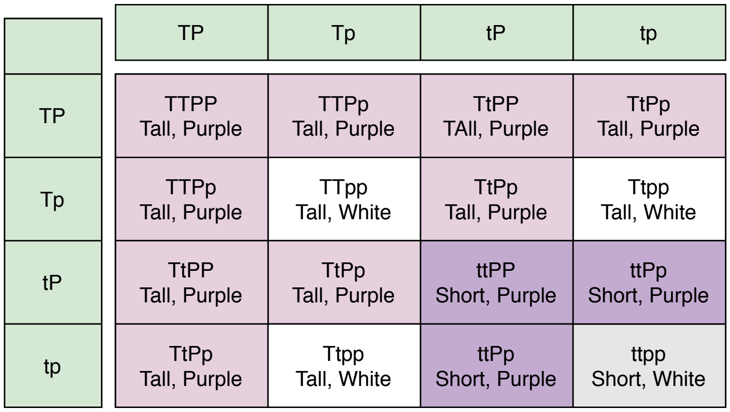
1.2.3 Q3: Explain Mendel's second law with proper example.

Mendel's second law is known as *The Law of Independent Assortment*. This law states that, transmission of one trait (characteristic) does not affect the transmission of other traits.

Let,
T = Stem length tall
t = Stem length short
P = Allele for purple
p = Allele for white

In parent generation, a tall and purple flower is crossed with white and short flower. The F1 generation cross was between two heterozygous tall and purple flowers. The cross between two F1 generation generates F2 generation. The punnett chart of this dihybrid cross (F2 generation) between two heterozygous tall, purple flower is shown in Figure 3. When we examine the punnet chart, we can see the results as following:

1. 9 tall and purple flower
2. 3 short and purple flower
3. 3 tall and white flower



	TP	Tp	tP	tp
TP	TTPP Tall, Purple	TTPp Tall, Purple	TtPP TAll, Purple	TtPp Tall, Purple
Tp	TTPp Tall, Purple	TTpp Tall, White	TtPp Tall, Purple	Ttpp Tall, White
tP	TtPP Tall, Purple	TtPp Tall, Purple	ttPP Short, Purple	ttPp Short, Purple
tp	TtPp Tall, Purple	Ttpp Tall, White	ttPp Short, Purple	ttpp Short, White

Figure 3: Example of Mendel's second law

4. 1 short and white flower

From the output we can observe that all the trait combinations are present in the F2 generation whereas they were lost in F1 generation. This proves that, the genes for each trait are not dependant on any other trait.

1.3 Questions for Ten Marks

1.3.1 Q: What is DNA replication? Write down the DNA replication process step by step.

DNA replication is the process by which DNA makes a copy of itself during cell division process.
DNA replication process:

1. The first step is to unzip the double helix structure of DNA. This process is carried out by an enzyme named *helicase* which breaks the hydrogen bonds between the complementary base pairs (A with T, C with G).

2. The separation of the strands creates a 'Y' shaped structure called the *replication fork*.These two separated strand will act as template for making the new strands of DNA. This process is shown in Fig 4.

3. One of the strand oriented in 3' to 5' direction toward the *replication fork* is known as *leading strand*. The other strand oriented in 5' to 3' direction

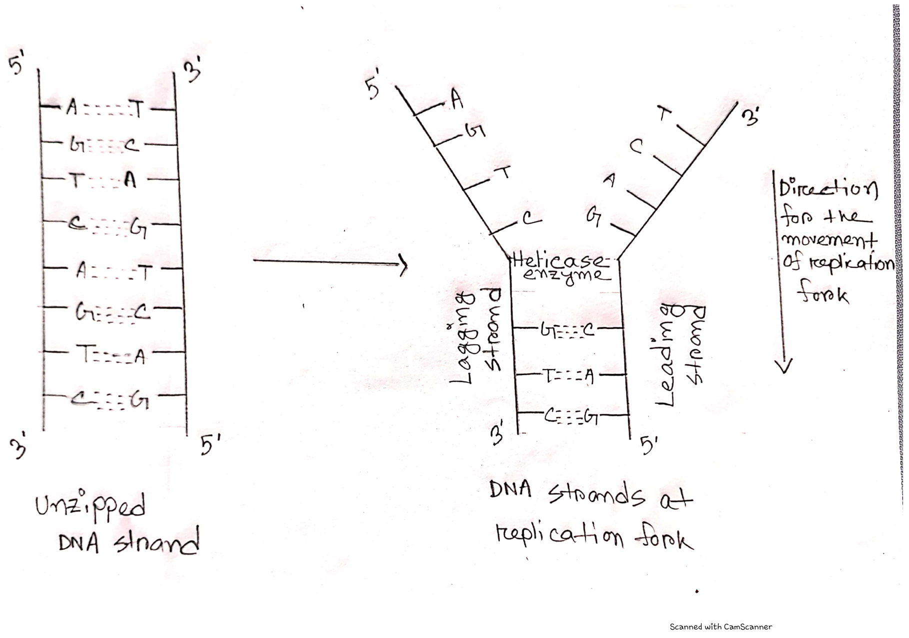


Figure 4: DNA strand at the replication fork

away from the replication fork is known as *lagging strand*. These two strands are replicated differently due to their different orientation.

4. Leading strand:

   (a) A short price of RNA named *primer* (produced from *primase* enzyme) binds itself at the end of the leading strand. The primer acts as starting point for DNA synthesis.

   (b) Another enxyme named *DNA polymerase* binds itself with the leading strand and adds complementary nucleotide bases. *DNA polymerase* can only add nucleotides in 5' to 3' direction. Therefore, adding nucleotide bases in the leading strand is a continuous process (Shown in Fig. 5).

5. Lagging strand:

   (a) Number of RNA primers are produced from *primase* enzyme and are binded in different positions of a lagging strand.

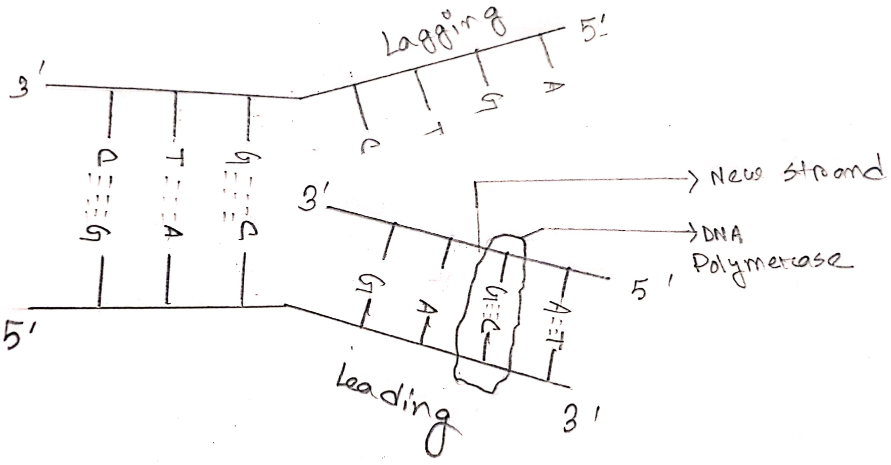


Figure 5: Leading strand operation

(b) *DNA polymerase* then creates chunk of DNA and adds them to the lagging strand in t 5' to 3' direction. These chunks are known as "Okazaki fragments".

(c) These types of replication is known as discontinuous replication as the *Okazaki fragments* will need to be joined up later.

6. After all the bases are matched up, an enzyme named *exonuclease* strips out all the primers from both the strands. The gaps created by primers are filled up by more complementary nucleotides.

7. The new strands are proofreaded so that there does not exist any mistakes in the DNA sequence.

8. Finally an enzyme named *DNA ligase* seals up the sequence of DNA (in both strand) to form a continuous double strands.

9. the result of DNA replication is two DNA molecules. DNA replication is a semi-conservative process where the replicated DNA(s) consist of half new and half old strand.

10. Upon completion of the DNA replication the new DNA automatically winds up into a double helix.

2 Chapter 01

2.1 Questions for One Marks

2.1.1 Q1: What is viral vector?

Viral vectors are tools used by biologists which can penetrate cell walls and deliver genetic materials to the cell.

2.1.2 Q2: What is an oriC?

oriC (also called replication origin) is the genomic region from where DNA replication begins.

2.1.3 Q3: What is DnaA box?

DnaA protein binds to a short segment within the OriC, known as DnaA box.

2.1.4 Q4: What is the complexity of FREQUENTWORDS algorithm?

$\mathcal{O}\left(|Text|^2 \cdot k\right)$

2.1.5 Q5: What is Forward half strand?

If traversing Dna from oriC to terC is in $5\prime$ to $3\prime$ direction, then it is forward half strand.

2.2 Questions for Five Marks

2.2.1 Q1: Draw a skew diagram for the genome sequence: CTGCAATGCATGACAC

Skew diagram for the sequence is shown in Fig. 6

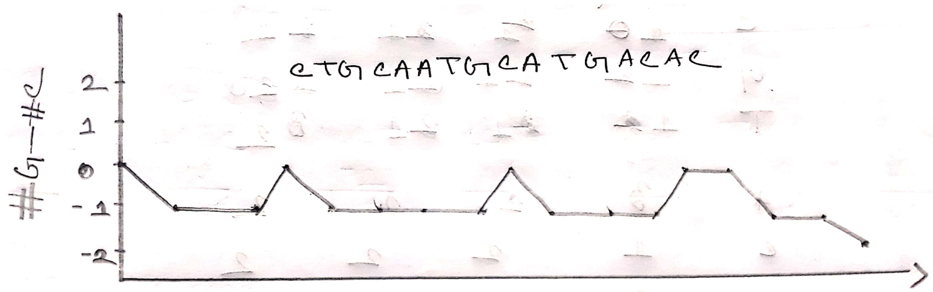


Figure 6: Skew diagram for the given sequence

2.2.2 Find the hamming distance:

(a) AATCGCGGG, AGTCTCGAG
(b) CTTGATCAT, CCTGATTAT
Find the reverse compliment for the following string:
(c) TCTTGATCA
(d) CTCTTGATC

Hamming distance:
a) Hamming distance between AATCGCGGG and AGTCTCGAG is 3.
b) Hamming distance between CTTGATCAT, CCTGATTAT is 2.
Reverse complement:
c) TCTTGATCA

1. Complement of TCTTGATCA: AGAACTAGT
2. Reverse of AGAACTAGT: TGATCAAGA

Reverse complement of TCTTGATCA is TGATCAAGA
d) CTCTTGATC

1. Complement of CTCTTGATC: GAGAACTAG
2. Reverse of GAGAACTAG: GATCAAGAG

Reverse complement of CTCTTGATC is GATCAAGAG

2.2.3 Q3: Draw skew graphs for the following situations.

The answer is shown in Fig. 7

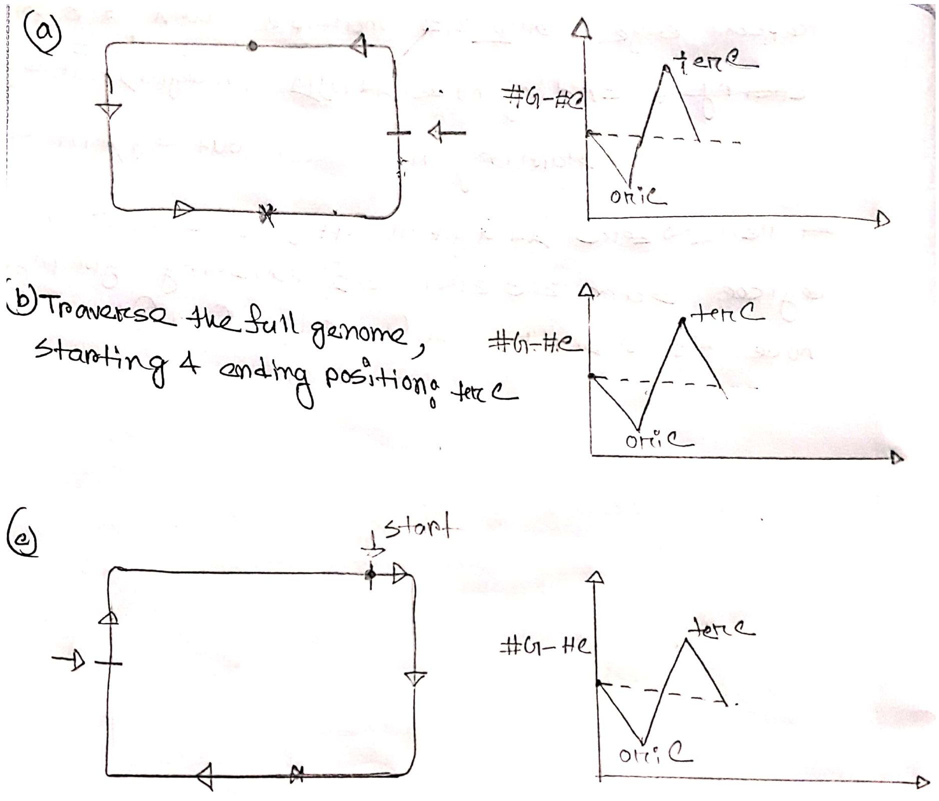

(a)
#G-#C
terC
oriC
b) Traverse the full genome,
Starting & ending position: terC
#G-#C
terC
oriC
(c)
start
#G-#C
terC
oriC

Figure 7: Skew graph

2.3 Questions for Ten Marks

2.3.1 Q: Form (12, 3)-clump for this given genome: ATGATGTGCATGGGGAAAGGG. (k = 3)

All possible 12 length strings: ATGATGTGCATG, TGATGTGCATGG, GATGTGCATGGG, ATGTGCATGGGG, TGTGCATGGGGA, GTGCATGGGGAA, TGCATGGGGAAA, GCATGGGGAAAG, CATGGGGAAAGG, ATGGGGAAAGGG.

From the table 1 we can say that, (12, 3)-clumps are: ATG and GGG.

Table 1: All occurrences

Pattern	0-11	1-12	2-13	3-14	4-15	5-16	6-17	7-18	8-19	9-20
ATG	3	2	2	2	1	1	1	1	1	1
TGA	1	1								
GAT	1	1	1							
TGT	1	1	1	1	1					
GTG	1	1	1	1	1	1				
TGC	1	1	1	1	1	1	1			
GCA	1	1	1	1	1	1	1	1		
CAT	1	1	1	1	1	1	1	1	1	
TGG		1	1	1	1	1	1	1	1	1
GGG			1	2	2	2	2	2	2	3
GGA					1	1	1	1	1	1
GAA						1	1	1	1	1
AAA							1	1	1	1
AAG								1	1	1
AGG									1	1

3 Chapter 02

3.1 Questions for One Marks

3.1.1 Q1: What is the complexity of Median String solution of Motif finding problem?

$\mathcal{O}\left(4^k \cdot n \cdot k \cdot t\right)$

3.1.2 Q2: What is Circadian clock?

Circadian clock is the internal timekeeper which controls daily schedule of animal, plants, bacteria etc.

3.1.3 Q3: What is gene expression profile?

The on and off states of all genes is known as gene expression profile.

3.1.4 Q4: What is the main contribution of clock genes?

Clock genes mainly control genes when sun-rises, including genes relate to photosynthesis, photo-reception, flowering.

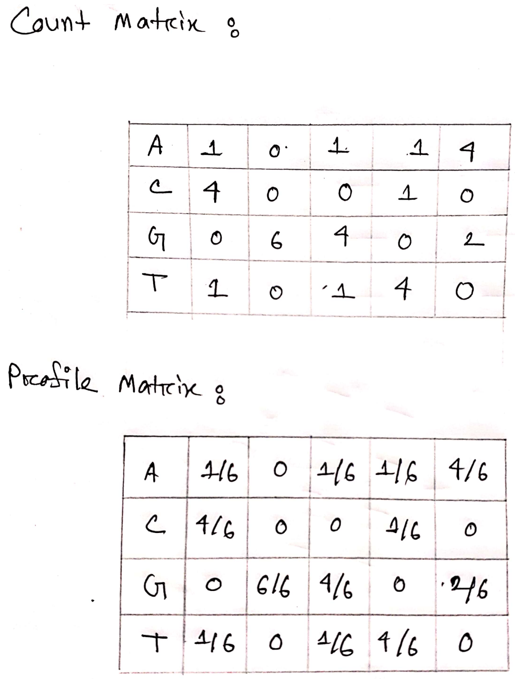
3.1.5 Q5: What is regulatory motifs?

The sequence where regulatory protein binds is known as regulatory motifs.

3.2 Questions for Five Marks

3.2.1 Q1: Find the count and profile matrix from the below Motifs: CGGCg, CGtTA, CGGTA, aGaTA, CGGag, tGGTA

The count matrix and profile matrix are shown in Fig. 8:



Count Matrix :

A	1	0	1	1	4
C	4	0	0	1	0
G	0	6	4	0	2
T	1	0	1	4	0

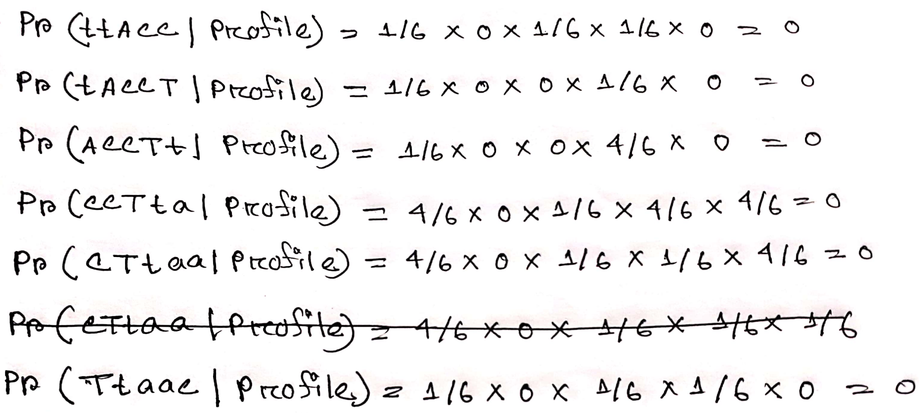
Profile Matrix :

A	1/6	0	1/6	1/6	4/6
C	4/6	0	0	1/6	0
G	0	6/6	4/6	0	2/6
T	1/6	0	1/6	4/6	0

Figure 8: Count matrix and profile matrix

3.2.2 Q2: Find profile most probable 5-mer for the given sequence: ttACCTtaac. Use the profile matrix from previous question

Here, Given string is ttACCTtaac.
All possible k-mers (k=5) are: ttACC, tACCT, ACCTt, CCTta, CTtaa, Ttaac.
Probability of the 4-mers for the given profile matrix are shown below (Fig 9):



Pr (ttAcc | Profile) = 1/6 × 0 × 1/6 × 1/6 × 0 = 0

Pr (tAccT | Profile) = 1/6 × 0 × 0 × 1/6 × 0 = 0

Pr (AccTt | Profile) = 1/6 × 0 × 0 × 4/6 × 0 = 0

Pr (ccTta | Profile) = 4/6 × 0 × 1/6 × 4/6 × 4/6 = 0

Pr (cTtaa | Profile) = 4/6 × 0 × 1/6 × 1/6 × 4/6 = 0

~~Pr (cTtaa | Profile) = 4/6 × 0 × 1/6 × 1/6 × 1/6~~

Pr (Ttaac | Profile) = 1/6 × 0 × 1/6 × 1/6 × 0 = 0

Figure 9: Probabilities of k-mers

From the figure above we can say that all the probabilities of occurring 5-mers are zero. Therefore, there is no profile most probable k-mer present in this sequence.

3.2.3 Q3: What are the drawbacks of RandomizedMotifSearch algorithm? Calculate profile matrix for ttAC and calculate the probability of occurring TGTC using this profile matrix. You may use pseudocounts if necessary.

Problems in RandomizedMotifSearch: RandomizedMotifSearch takes a set of motifs based on Profile then checks whether it has better score or not. If the score doesn't improve, it may removes the entire motifs set based on profile matrix and again choose a new one in the next iteration. But there may exists few motifs which are in our actual motifs set.

2nd part:
Profile matrix for ttAC:

As there are 0's in matrix profile , we will use pseudocounts and make our new matrix profile.

Table 2: Profile Matrix for ttAC

	1st	2nd	3rd	4th
A	0	0	1	0
C	0	0	0	1
G	0	0	0	0
T	1	1	0	0

Table 3: Count Matrix for ttAC

	1st	2nd	3rd	4th
A	1	1	1	1
C	1	1	1	1
G	1	1	1	1
T	1	1	1	1

Table 4: Profile Matrix for ttAC

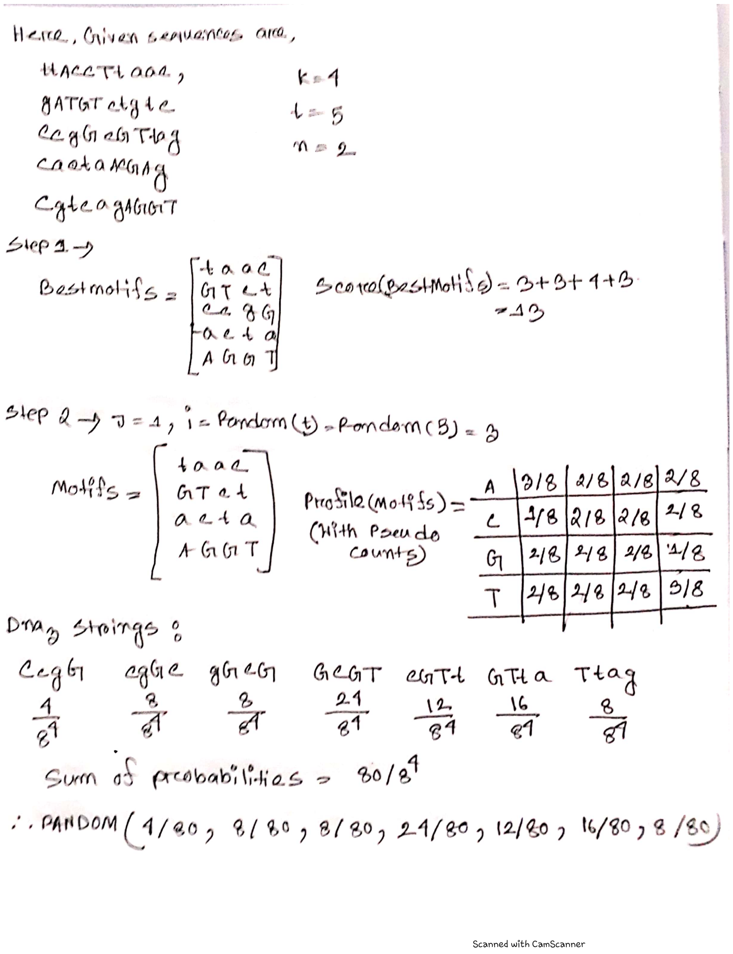
	1st	2nd	3rd	4th
A	1/5	1/5	2/5	1/5
C	1/5	1/5	1/5	2/5
G	1/5	1/5	1/5	1/5
T	2/5	2/5	1/5	1/5

Pr(TGTC | Profile) $= (2/5) * (1/5) * (1/5) * (2/5) = 4/5^4$

3.3 Questions for Ten Marks

3.3.1 Q: Traverse Gibbs sampling algorithm for N = 2, k = 4 and t = 5. Given sequences are: ttACCTtaac, gATGTctgtc, CcgGcGTtag, CactaACGAg, CgtcagAGGT

The answer is shown in Fig. 10 and 11.



Here, Given sequences are,

ttACCTtaaa, $k=4$
gATGTctgtc $t=5$
ccgGcGTtag $n=2$
caataACGAg
cgtcagAGGT

Step 1 →

$$\text{Bestmotifs} = \begin{bmatrix} t & a & a & c \\ G & T & c & t \\ c & a & g & G \\ a & c & t & a \\ A & G & G & T \end{bmatrix}$$

Score(BestMotifs) = 3+3+4+3
= 13

Step 2 → J = 1, i = Random(t) = Random(5) = 3

$$\text{Motifs} = \begin{bmatrix} t & a & a & c \\ G & T & c & t \\ a & c & t & a \\ A & G & G & T \end{bmatrix}$$

Profile(Motifs) =
(With Pseudo counts)

A	3/8	2/8	2/8	2/8
C	1/8	2/8	2/8	2/8
G	2/8	2/8	2/8	1/8
T	2/8	2/8	2/8	3/8

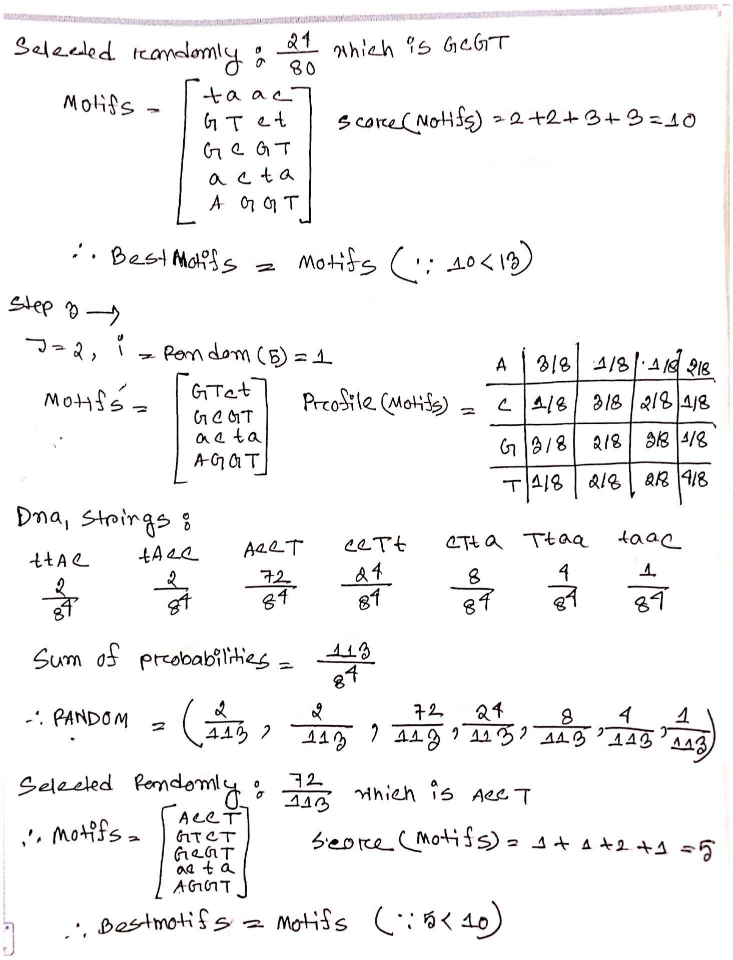
Draw strings :

ccgG	cgGc	gGcG	GcGT	cGTt	GTta	Ttag
$\frac{4}{8^4}$	$\frac{8}{8^4}$	$\frac{8}{8^4}$	$\frac{24}{8^4}$	$\frac{12}{8^4}$	$\frac{16}{8^4}$	$\frac{8}{8^4}$

Sum of probabilities = $80/8^4$

∴ PANDOM(4/80, 8/80, 8/80, 24/80, 12/80, 16/80, 8/80)

Figure 10: Question 3.3.1 (Gibbs sampling part 1)



Selected randomly : $\frac{24}{80}$ which is GcGT

$$\text{Motifs} = \begin{bmatrix} t\,a\,a\,c \\ G\,T\,c\,t \\ G\,c\,G\,T \\ a\,c\,t\,a \\ A\,G\,G\,T \end{bmatrix} \qquad \text{score(Motifs)} = 2+2+3+3=10$$

$\therefore$ BestMotifs = Motifs ($\because$ 10<13)

Step 3 →

J = 2, i = Random(5) = 1

$$\text{Motifs}' = \begin{bmatrix} G\,T\,c\,t \\ G\,c\,G\,T \\ a\,c\,t\,a \\ A\,G\,G\,T \end{bmatrix} \qquad \text{Profile(Motifs)} =$$

A	3/8	1/8	1/8	2/8
c	1/8	3/8	2/8	1/8
G	3/8	2/8	3/8	1/8
T	1/8	2/8	2/8	4/8

Dna, Strings :

ttAc	tAcc	AccT	ccTt	cTta	Ttaa	taac
$\frac{2}{8^4}$	$\frac{2}{8^4}$	$\frac{72}{8^4}$	$\frac{24}{8^4}$	$\frac{8}{8^4}$	$\frac{4}{8^4}$	$\frac{1}{8^4}$

Sum of probabilities = $\frac{113}{8^4}$

$\therefore$ RANDOM $= \left(\frac{2}{113}, \frac{2}{113}, \frac{72}{113}, \frac{24}{113}, \frac{8}{113}, \frac{4}{113}, \frac{1}{113}\right)$

Selected Randomly : $\frac{72}{113}$ which is AccT

$$\therefore \text{Motifs} = \begin{bmatrix} A\,c\,c\,T \\ G\,T\,c\,T \\ G\,c\,G\,T \\ a\,c\,t\,a \\ A\,G\,G\,T \end{bmatrix} \qquad \text{Score(Motifs)} = 1+1+2+1=5$$

$\therefore$ Bestmotifs = Motifs ($\because$ 5< 10)

Scanned with CamScanner

Figure 11: Question 3.3.1 (Gibbs sampling part 2)