1. **Multiple sequence alignment**

   Given the following sequences:

   CTATTAATAC

   TATTAATAC

   CTATTAATC

   CATTAATAC

   Assume a match score of +1, a gap penalty of -1 and a mismatch score of -1. Suppose we choose the first sequence as the center, use the Star algorithm to construct a multiple sequence alignment.

   You may find optimal pair-wise alignment manually or by using your program from last assignment. Also calculate the SP score (sum of pairs) for the multiple sequence alignment.

Answer:

- Sequence 1 → CTATTAATAC
- Sequence 2 → TATTAATAC
- Sequence 3 → CTATTAATC
- Sequence 4 → CATTAATAC

Let's choose Sequence 1 as our center. Since Sequence 1 is chosen as center we will now find optimal pair-wise alignment.

Sequence 1 → CTATTAATAC

Sequence 2 → -TATTAATAC

Score = 8

Sequence 1 → CTATTAATAC

Sequence 2 → CTATTAAT-C

Score = 8

Sequence 1 → CTATTAATAC

Sequence 4 → C-ATTAATAC

Score = 8

So the multiple sequence alignment is:

Sequence 2 → -TATTAATAC

Sequence 1 → CTATTAATAC

Sequence 3 → CTATTAAT-C

Sequence 4 → C-ATTAATAC

SP Scores for Multiple alignment:

Score = score {S1, S2} + score {S1, S3} + score {S1, S4} + score {S2, S3} + score {S2, S4} + score {S3, S4}

Score = 8 + 8 + 8 + 6 + 6 + 6

Score = 42

2.

a. Construct an alignment of all instances of motifs, and compute the profile and consensus sequence of the motif. (In the event of a "tie", select one nucleotide as a representative for that position)

GAACTCATGGTG
AAAAGCACGGTC
TCAAAGCAAGGC
CCTAATCAGGGC
AAGTATGGACTC
ACTAAGCAGGGT
TCTCACGGCCCA
CCTCGTGGTGGG
TACCGTATGGTT
ACCACTCGTCGA

Answer:

Sequence	Motifs
GAACTCATGGTG	ACTCATGG
AAAAGCACGGTC	AAGCACGG
TCAAAGCAAGGC	AAGCAAGG
CCTAATCAGGGC	AATCAGGG
AAGTATGGACTC	AAGTATGG
ACTAAGCAGGGT	AAGCAGGG
TCTCACGGCCCA	TCTCACGG
CCTCGTGGTGGG	CCTCGTGG
TACCGTATGGTT	CCGTATGG
ACCACTCGTCGA	ACTCGTCG

Profile Matrix:

A	0.7	0.5	0	0	0.8	0.1	0	0
C	0.2	0.5	0	0.8	0	0.2	0.1	0
G	0	0	0.5	0	0.2	0.2	0.9	1
T	0.1	0	0.5	0.2	0	0.5	0	0

Consensus sequence: ACGCATGG

b. Compute the entropy score of the motif profile.

Answer:

Entropy score= $-\sum_{i=1}^{N} p_i \cdot \log p_i$

Col 1 → $(0.7\log_2 0.7) + (0.2\log_2 0.2) + (0\log_2 0) + (0.1\log_2 0.1)$
Col 2 → $(0.5\log_2 0.5) + (0.5\log_2 0.5) + (0\log_2 0) + (0\log_2 0)$
Col 3 → $(0\log_2 0) + (0\log_2 0) + (0.5\log_2 0.5) + (0.5\log_2 0.5)$
Col 4 → $(0\log_2 0) + (0.8\log_2 0.8) + (0\log_2 0) + (0.2\log_2 0.2)$
Col 5 → $(0.8\log_2 0.8) + (0\log_2 0) + (0.2\log_2 0.2) + (0\log_2 0)$
Col 6 → $(0.1\log_2 0.1) + (0.2\log_2 0.2) + (0.2\log_2 0.2) + (0.5\log_2 0.5)$
Col 7 → $(0\log_2 0) + (0.1\log_2 0.1) + (0.9\log_2 0.9) + (0\log_2 0)$
Col 8 → $(0\log_2 0) + (0\log_2 0) + (1\log_2 1) + (0\log_2 0)$

Col 1 → ((-0.359) + (-0.464) + (0)+(-0.332)) = -1.155
Col 2→ (-0.5) + (-0.5) +(0) +(0) = -1
Col 3 → (0)+(0)+(-0.5)+(-0.5) = -1
Col 4 → (0) + (-0.257) +(0)+(-0.464) = -0.721
Col5 → (-0.257) + (0) + (-0.464) + (0)= -0.721
Col6 → (-0.332) + (-0.464) + (-0.464) + (-0.5) = -1.759
Col7 → (0) + (-0.332) + (-0.1368) + (0) = -0.4688
Col8 → (0) + (0) + (0) + (0) =-0

=6.8248

c. Compute the likelihood ratio of getting "ACTCATGG" according to motif profile vs. a background model of each base having equal probability of 25%.

Answer:

Likelihood ratio= P1/ P2

P1= 0.7 * 0.5 * 0.5 * 0.8 * 0.8 * 0.5 * 0.9 * 1

= 0.0504 (chances of getting **ACTCATGG** for a motif profile)

P2 = 0.25 * 0.25 * 0.25 * 0.25 * 0.25 * 0.25* 0.25 *0.25

= $(0.25)^8$ (background model of each base having equal probability of 25%)

Hence, Likelihood ratio = P1/ P2

= 0.0504 / $(0.25)^8$

= 3303.0144

3. Let S= {AAT, ATC, ATG, CAT, GAA, TCA, TGG} be a 3-mer spectrum.

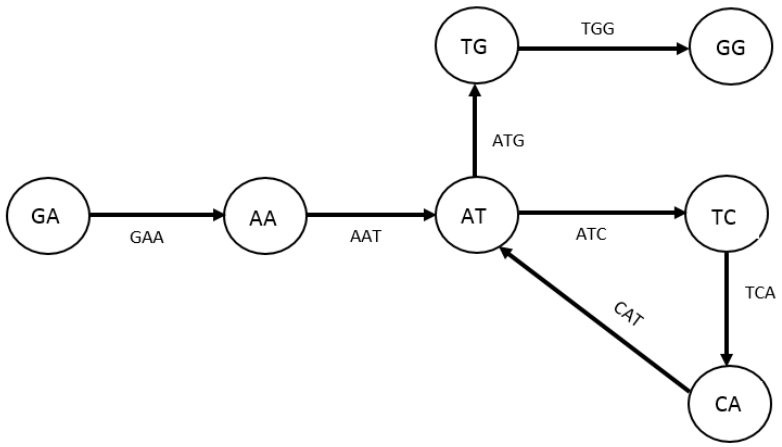
a) Show the de Bruijn graph that represents this spectrum.

Answer:

De Bruij graph:

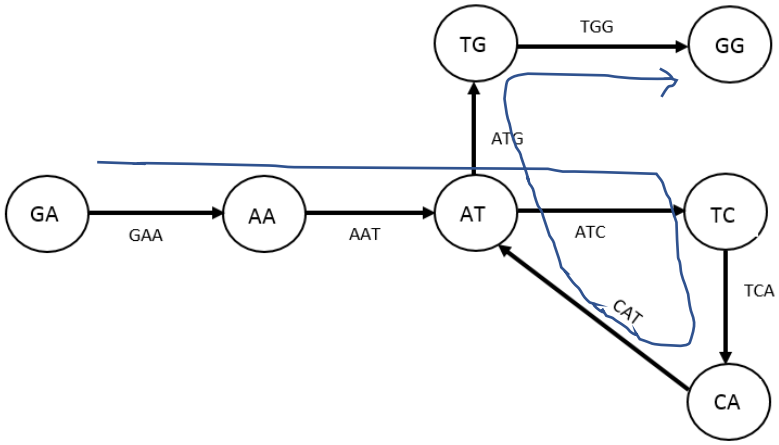
S= {AAT, ATC, ATG, CAT, GAA, TCA, TGG}

We get **GAATCATGG** as the shortest sequence from above 3-mers.



b) Show all Eulerian paths for this graph, and the assembled sequence each one represents

Answer:



4. Programming assignment

OUTPUT :

Motif Length --> 8

DNA[0][20:28] ggagtcag

DNA[1][7:15] tgtgtcat

DNA[2][22:30] tgacacag

DNA[3][26:34] tgagtcag

DNA[4][33:41] taagtcac

DNA[5][32:40] tgactcat

DNA[6][18:26] tgattcag

DNA[7][20:28]	tgcgtcag
DNA[8][36:44]	tgagtcag
DNA[9][25:33]	tgagtcag
DNA[10][24:32]	ggagtcat
DNA[11][40:48]	tgcgtcag
DNA[12][42:50]	tgattaag
DNA[13][28:36]	tgagtcac
DNA[14][40:48]	tgactcag
a	0.0,0.07,0.8,0.0,0.07,0.07,1.0,0.0
c	0.0,0.0,0.13,0.2,0.0,0.93,0.0,0.13
t	0.87,0.0,0.07,0.13,0.93,0.0,0.0,0.2
g	0.13,0.93,0.0,0.67,0.0,0.0,0.0,0.67
Consensus	tgagtcag

Motif Length --> 9

DNA[0][19:28]	tggagtcag

DNA[1][34:43]	cccagtcag
DNA[2][21:30]	atgacacag
DNA[3][25:34]	gtgagtcag
DNA[4][32:41]	ctaagtcac
DNA[5][31:40]	ctgactcat
DNA[6][17:26]	atgattcag
DNA[7][19:28]	ctgcgtcag
DNA[8][35:44]	ctgagtcag
DNA[9][24:33]	atgagtcag
DNA[10][23:32]	gggagtcat
DNA[11][39:48]	ctgcgtcag
DNA[12][9:18]	gtgactaat
DNA[13][27:36]	gtgagtcac
DNA[14][39:48]	ctgactcag

a	0.2,0.0,0.07,0.87,0.0,0.07,0.07,1.0,0.0
c	0.47,0.07,0.07,0.13,0.27,0.0,0.93,0.0,0.13
t	0.07,0.8,0.0,0.0,0.07,0.93,0.0,0.0,0.2
g	0.27,0.13,0.87,0.0,0.67,0.0,0.0,0.0,0.67
Consensus	ctgagtcag

Motif Length --> 10

DNA[0][18:28]	ttggagtcag
DNA[1][5:15]	cctgtgtcat
DNA[2][20:30]	gatgacacag
DNA[3][24:34]	tgtgagtcag
DNA[4][31:41]	gctaagtcac
DNA[5][30:40]	tctgactcat
DNA[6][16:26]	aatgattcag
DNA[7][18:28]	tctgcgtcag
DNA[8][34:44]	cctgagtcag

DNA[9][23:33]	catgagtcag
DNA[10][22:32]	tgggagtcat
DNA[11][38:48]	cctgcgtcag
DNA[12][8:18]	agtgactaat
DNA[13][26:36]	tgtgagtcac
DNA[14][38:48]	tctgactcag
a	0.13,0.2,0.0,0.07,0.8,0.0,0.07,0.07,1.0,0.0
c	0.27,0.47,0.0,0.0,0.13,0.27,0.0,0.93,0.0,0.13
t	0.47,0.07,0.87,0.0,0.07,0.07,0.93,0.0,0.0,0.27
g	0.13,0.27,0.13,0.93,0.0,0.67,0.0,0.0,0.0,0.6
Consensus	tctgagtcag

Motif Length --> 11

DNA[0][17:28]	tttggagtcag
DNA[1][4:15]	tcctgtgtcat

DNA[2][19:30]	ggatgacacag
DNA[3][23:34]	ttgtgagtcag
DNA[4][30:41]	ggctaagtcac
DNA[5][29:40]	ctctgactcat
DNA[6][15:26]	aaatgattcag
DNA[7][17:28]	atctgcgtcag
DNA[8][33:44]	ccctgagtcag
DNA[9][22:33]	gcatgagtcag
DNA[10][21:32]	ttgggagtcat
DNA[11][37:48]	gcctgcgtcag
DNA[12][7:18]	gagtgactaat
DNA[13][25:36]	gtgtgagtcac
DNA[14][37:48]	gtctgactcag
a	0.13,0.13,0.2,0.0,0.07,0.8,0.0,0.07,0.07,1.0,0.0
c	0.13,0.27,0.47,0.0,0.0,0.13,0.27,0.0,0.93,0.0,0.13

t	0.27,0.47,0.07,0.87,0.0,0.07,0.07,0.93,0.0,0.0,0.27
g	0.47,0.13,0.27,0.13,0.93,0.0,0.67,0.0,0.0,0.0,0.6

5.

Consensus gtctgagtcag