BT3040 BIOINFORMATICS ENDSEM

G. Shiva Ram Raj
BE18B023

3) I GTGCTGCACG
II AGCTGCAAGC
III GTGCTGCACT
IV GTATCACACG
V GTATCACATA

a)

	I	II	III	IV	V
I	–				
II	9	–			
III	① 9	–			
IV	4	8	5	–	
V	6	8	6	2	–

→

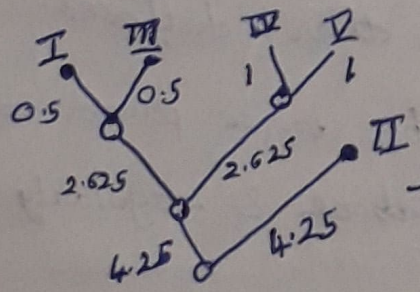
	I,III	II	IV	V
I,III	–			
II	9	–		
IV	4.5	8	–	
V	6	8	②	–

↓

	I,III	II	IV,V
I,III	–		
II	9	–	
IV,V	⑤ 5.25	8	–

←

	(I,III),(IV,V)	II
(I,III),(IV,V)	–	
II	8.5	–



→ Scaled Tree

b)

	1	2	3	4	5	6	7	8	9	10
A	1	0	2	0	0	2	1	5	0	1
G	4	1	2	0	1	2	0	0	1	2
C	0	0	1	2	2	1	4	0	3	1
T	0	4	0	3	2	0	0	0	1	1

→ Freq matrix

Use the formula $\ln\left[\frac{(n_{ij}+P_i)/(N+1)}{P_i}\right]$ N → no. of sequences

For Ex:- If $n_{ij}=1$ then $\ln\left(\frac{1.25}{(5+1)\times 0.25}\right)$

$= \ln\left(\frac{1.25}{1.5}\right) = -0.18$

Similary, calculate the elements of weight matrix using other values

	1	2	3	4	5	6	7	8	9	10
A	0.182	-1.79	0.405	-1.79	-1.79	0.405	-0.182	1.252	-1.79	-0.182
G	1.04	-0.18	0.405	-1.79	-0.182	0.405	-1.79	-1.79	-0.182	0.405
C	-1.79	-1.79	-0.182	0.405	0.405	-0.182	1.04	-1.79	0.773	-0.182
T	-1.79	1.04	-1.79	0.773	0.405	-1.79	-1.79	-1.79	-0.182	-0.182

3c) Position 3 ÷ G freq : 0.4
A freq : 0.4
C freq : 0.2

Using entropy measure
$C.S = 0.4\ln 0.4 + 0.4\ln 0.4 + 0.2\ln 0.2$
$= -1.05492$

Position 8 ÷ A freq : 1
$CS = 1\ln 1 = 0$

3d) score: 2 mismatch: -1 gap penalty: -3

		A	G	C	T	G	C	A	A	G	C
	0	0	0	0	0	0	0	0	0	0	0
G	0	0	2	0	0	2	0	0	0	2	0
T	0	0	0	1	2	0	1	0	0	0	1
G	0	0	0	0	0	4	0	0	0	2	0
C	0	0	0	0	1	2	3	1	1	6	3
T	0	0	0	1	0	1	3	3	5	2	0
G	0	0	2	0	0	0	2	5	2	4	4
C	0	0	0	4	1	0	1	3	4	3	6
A	0	2	0	1	3	2	0	2	6	5	3
C	0	0	4	2	0	2	4	1	5	7	6
G	0	0	2	3	1	2	1	6	4	8	8

Local alignment

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

$2+2+2+2+2+2-1+2 = 13$

1) a) Life sciences is used to get information of biological entities

Computing is used for storing it in a computer so that the read information is easily accessible

Maths, Statistics is used to propose different methods for calculating different parament.

Physics is used to define the forces that are involved in biological molecules.

IT (Information Tech) is used to collect large sets of data

b) i) To predict the protein structures & the amino acids involved in it, so that we can find the function of that respective protein.

ii) To align large no. of sequences & find the conserved subset of sequences

iii) Phylogeny Analysis

iv) To understand the freq. of mutations, insertions & deletions

v) To construct diff. databases.

vi) Also for structure based drug design

vii) To understand protein folding & its stability

2) a) Database contains collection of information in a computer readable form

d) DDBJ, EMBL

c b) i) The data which is in a proper order & format

ii) Also organize the database with proper definitions (to maintain)

iii) Different routes for retrieval of data & also to use this in other programs

iv) Proper presentation of results.

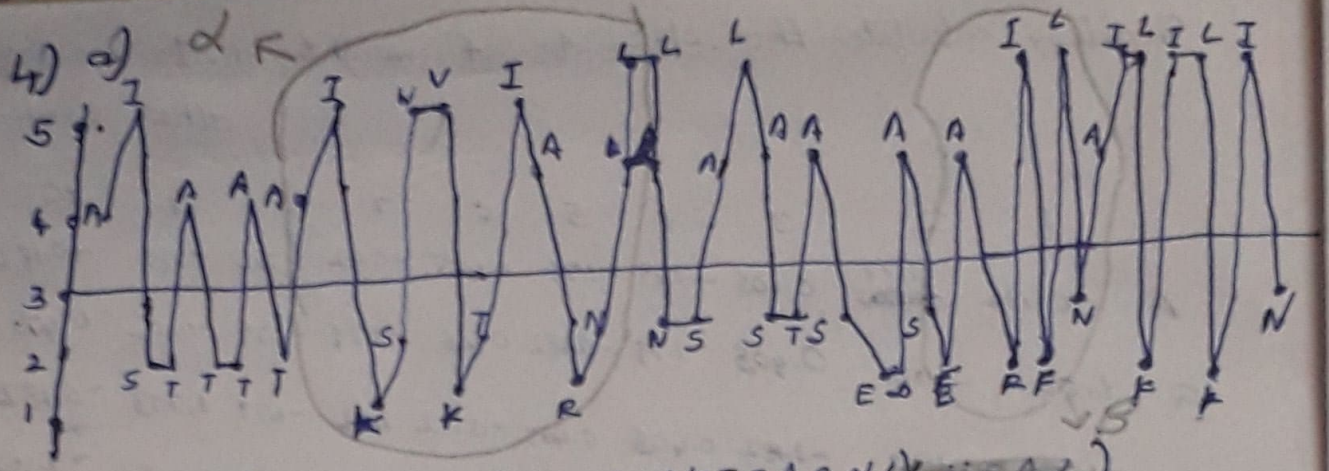
v) Also the links for original publications

vi) Interlinkage with other databases.

b) i) The data should be in a proper order

ii) Maintaing the database & regularly updating it.

4) a)



alpha segment (AIKSWVKTIARNLLNSAL)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

$a_1 = \frac{4+5+5+5}{4} = 4.75 \rightarrow \frac{H_{A1} + H_{N5} + H_{I9} + H_{L13}}{4}$

Similarly $a_2 = \frac{5+5+4}{3} = 4.67$ $\quad a_3 = \frac{1+1+1+2}{3} = 1.25$ ↘ $\left(\frac{H_{K3} + H_{K7} + H_{A11} + H_{N15}}{3}\right)$

Similarly $a_4 = \frac{2+2+2}{3} = 2.2$

Power of amphipathicity = $|(a_1 + a_2) - (a_3 + a_4)|$ (R0.1)

$= |(4.75 + 4.67) - (1+2)|$

$= |1.75 + 4.67|$

$= 6.42$

beta segment EARIKLNA
1 2 3 4 5 6 7 8

$b_1 = \frac{1+1+1+2}{4} = 1.25$

$b_2 = \frac{4+5+5+4}{4} = 4.5$

P.0.1 $= |b_1 - b_2|$

$= 3.25$

b) alpha AIKSWVKTIARNLLNSAL

First 6 : AIKSWV = 1 + 0.5 + 0.5 + 0 + 1 + 1 = 4 = 4

Next 4 : KTIA = 1.07 + 0.82 + 1.00 + 1.45 > 4

RNLL = 0.79 + 0.73 + 1.34 + 1.34 < 4 ● Terminate

TIAR = 0.82 + 1.00 + 1.45 + 0.79 > 4

IARN = 1.00 + 1.45 + 0.79 + 0.73 > 4

ARNL = 1.45 + 0.79 + 0.73 + 1.34 > 4

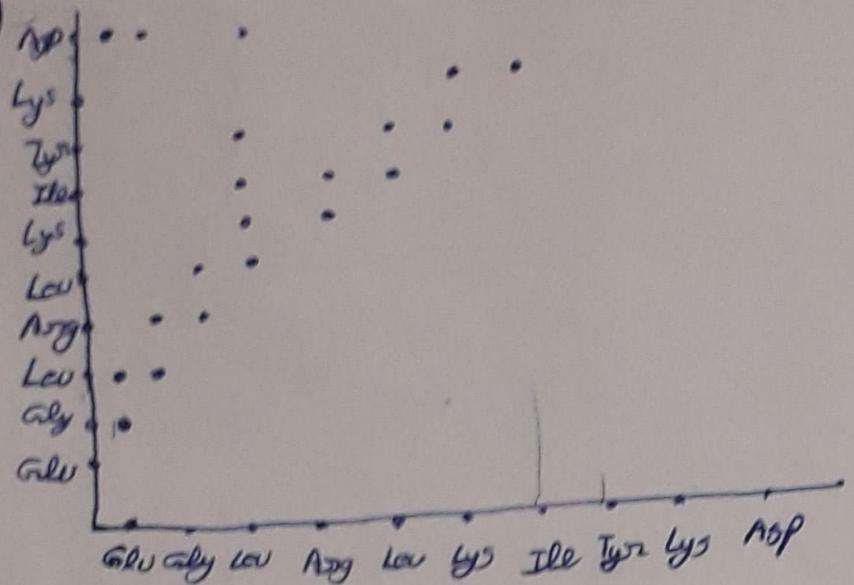
5) a) Contact maps represents the distance b/w all possible residue pair.

Short range contacts have sequence distance $< \pm 3$

medium range contacts have S.D ± 3 or ± 4

long-range contact have S.D $> \pm 4$

b)



c) i) S.R contacts (Glu^{11}, Gly^{12}) & (Arg^{14}, Lys^{16})

ii) M.R contacts (Arg^{14}, Tyr^{18}) & (Arg^{14}, Ile^{17})

iii) L.R contacts (Gly^{12}, Asp^{20}) & (Leu^{13}, Asp^{20})

d) Surrounding hydrophobicity of the residue. Lys^{16}

$H_i = \sum n_j h_i$ $\quad n_L \to 2 \quad n_R \to 1 \quad n_D \to 1 \quad n_G \to 1 \quad n_L \to 1 \quad n_E \to 1 \quad n_I \to 1 \quad n_K \to 1$

$= (2\times5) + (0\times1) + (0\times1) + (0\times4) + (0\times4) + (0\times1) + (0\times5) + (0\times1)$

$= 27$

6) b)

		Predicted +	Predicted −
Expt	+	7	3
	−	6	44

$TP = 7 \quad FN = 3$

$FP = 6 \quad TN = 44$

$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{7}{7+3} = 0.7$

$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} = \frac{13}{60} = 0.21667$

$\text{Specificity} = \frac{TN}{TN+FP} = \frac{6}{50} = \frac{3}{25} = 0.12$