Offset 5

$S_{CDEKD} = 1 + 1 - \infty + 0 = \infty$

= very tiny value (indeterminate)

1C

① Offset position 3 of the given sequence when aligned with probe shows highest log odds score = 10.5 dB. Indeed, this offset position shows the presence of conserved motif 'NKEDE'.

=3

---

(6) Needleman Wunsch algorithm is employed to perform global sequence alignment.

The algorithm involves three stages

a) Matrix initialization

b) Matrix fill

0.5 c) Trace back.

The matrix will be initialized with the given scoring scheme, match = 5, mismatch = -3 and INDEL = -6.

The matrix will be filled/scored as per the following condition:

$$S_{i,j} = \max \begin{cases} S_{i-1,j} + (-6) \\ S_{i,j-1} + (-6) \\ S_{i-1,j-1} + S(x_i, y_j) \end{cases}$$

0.5

(13)

Alignment 1
INDIANA
INDI-GO

Alignment 2
INDIANA
INDIG-O

Alignment 3
INDIANA
INDIGO-

① / =5

---

7) (a) In CLUSTALW, global sequence alignment of given sequences ② are performed for which the E-value statistics is not available.

(b) In order to calculate the weighted ~~sco~~ score for the given MSA, the weight for each sequences needs to be calculated from the guide tree as follows:

Let W_i be the weight for the 'i^{th}' sequence.

$$W_1 = 0.3 + \left(\frac{0.04}{3}\right) + \left(\frac{0.06}{4}\right) = 0.3283$$

$$W_2 = 0.02 + \left(\frac{0.28}{2}\right) + \left(\frac{0.04}{3}\right) + \left(\frac{0.06}{4}\right) = 0.1883$$

$$W_3 = 0.02 + \left(\frac{0.28}{2}\right) + \left(\frac{0.04}{3}\right) + \left(\frac{0.06}{4}\right) = 0.1883$$

$$W_4 = 0.34 + \left(\frac{0.06}{4}\right) = 0.355$$

$$W_5 = 0.4$$

(15) For eg, the weight of sequence 1 (W_1) is calculated by summing up the distance from sequence (1) to the root in such a way that it a particular distance

(14)

b) → BLOSUM matrices are based on sequence identity
∴ Lesser the sequence identity greater is the divergence.
(1) In BLOSUM, the smallest number will be most suitable for distantly related sequence ⇒ BLOSUM 30

→ PAM matrices are based on evolutionary distance
∴ Larger the evolutionary distance greater is the divergence
(1) In PAM, the largest number will be most suitable for aligning distantly related sequences ⇒ PAM 250

=2

4) (a) To calculate the log odds score (S), we can simply add individual alignment score of the residues from
(1) the PAM 250 matrix

$$S = S(R,R) + S(K,K) + S(R,D) + S(K,E) + S(R,R) + S(K,K) + S(R,D) + S(K,E)$$
$$= 6 + 5 + (-1) + 0 + 6 + 5 + (-1) + 0$$
$$= 20$$

(1)
=2

(b) To determine the statistical significance of an alignment of two random sequences of length equal to query & db sequences that could achieve the given score, we have to calculate the E-value (E).
From Karlin-Altschul equation,

$$E = K m n e^{-\lambda S} \longrightarrow (1)$$

(1) K & λ are constants that depend on the scoring matrix.

(5)

3) (a) $M_{a,b}$ $b \to a$ $= \dfrac{\lambda\, m_b\, A_{a,b}}{\sum_a A_{a,b}}$ ⟶ (1)

$M_{b,a}$ $a \to b$ $= \dfrac{\lambda\, m_a\, A_{a,b}}{\sum_b A_{a,b}}$ ⟶ (2)

where $M_{a,b}$ & $M_{b,a}$ → Mutation probability matrix

λ → proportionality constant that depends on the scoring matrix & accounts for the unspecified evolutionary time scale

m_a & m_b → mutability of amino acids A & B, respectively

$A_{a,b}$ → Accepted point mutation matrix

From (1) & (2), it is clear that

$$\boxed{m_a \neq m_b}$$

Therefore, $M_{a,b}$ & $M_{b,a}$ are not symmetric

(b)

Alignment

E	D	G	E	
A	N	G	E	

Amino acids	A	D	E	G	N
No of charges	1	1	1	0	1
Total composition	1	1	3	2	1
Mutability = No of charges / Total composition	1	1	1/3	0	1

$$M_{AA} = 1 - \sum_{B \neq A} M_{BA} = 1 - \frac{\lambda\, m_A \sum_B A_{BA}}{\sum_B A_{BA}} = 1 - \lambda\, m_A$$

$$= 1 - \lambda \quad [\because m_A = 1 \text{ from the above table}]$$

$f^4_{T,T} = 1 \times \frac{1}{2} = \frac{1}{2}$; $f^2_{S,T} = \frac{1}{2} \times 1 = \frac{1}{2}$

$f^2_{A,T} = \frac{1}{3} \times 1 + \frac{1}{3} \times \frac{1}{2} = \frac{1}{3} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

$f^2_{A,S} = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$; $f^2_{D,S} = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$

$f^2_{D,T} = \frac{1}{3} \times 1 + \frac{1}{3} \times \frac{1}{2} = \frac{1}{3} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

$f^2_{E,T} = \frac{1}{3} \times 1 + \frac{1}{3} \times \frac{1}{2} = \frac{1}{3} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

$f^2_{E,S} = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$

No A & T in columns 3, + & 5. Hence these columns are ignored.

Column 6 : Cluster 1 $\Rightarrow \{T\}$; Cluster 2 $\Rightarrow \{\frac{1}{2} M, \frac{1}{2} A\}$

Cluster 3 $\Rightarrow \{\frac{1}{3} G, \frac{1}{3} A, \frac{1}{3} L\}$

$f^6_{A,T} = \frac{1}{2} \times 1 + \frac{1}{3} \times 1 = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$

$f^6_{M,T} = \frac{1}{2} \times 1 = \frac{1}{2}$; $f^6_{G,T} = \frac{1}{3} \times 1 = \frac{1}{3}$; $f^6_{L,T} = \frac{1}{3} \times 1 = \frac{1}{3}$

$f^6_{A,A} = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$; $f^6_{M,A} = \frac{1}{2} \times \frac{1}{3} = 1/6$

$f^6_{A,G} = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$; $f^6_{M,G} = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$

$f^6_{A,L} = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$; $f^6_{M,L} = \frac{1}{2} \times \frac{1}{3} = 1/6$

1.5 Since there are three clusters with six columns in the given alignment, total number of possible combination is given by

$$\sum_{a,b} f_{a,b} = 3 \times 6 = 18$$

0.5 Score (A,T) is given by

$$S_{A,T} = \log_2 \left(\frac{q_{A,T}}{2 P_A P_T} \right) \longrightarrow (1)$$

From (1), In order to calculate $S_{A,T}$, $q_{A,T}$ (non-random model),

$$q_{A,T} = \frac{\sum f_{A,T}}{\sum_{\text{all columns}} f_{a,b}}$$

0.5

m & n → length of the query & db, respectively
S → log odds score

$$E = 0.09 \times 250 \times 250 \; e^{-0.229 \times 20}$$

(1)

$$= 57.6837$$

Lesser the E value (closer to 0), greater is the significance of the alignment.

(1) Based on this, the given alignment is not stat[...] significant.

=3

———— x ————

5) (a) Log odds score in bits (S) = $\log_2 \left(\frac{\text{Observed frequency of}}{\text{Background frequency of}} \right)$

(1) Background frequency of amino acid = $\frac{1}{20} = 0.05$.

(1)

Amino acid	C1	C2	C3	C4	C5
N	3.585	1	1	1	1
K	0	3.8074	1	0	1
C	1	1	3.321	2	1
D	1	1	1	3.8074	$-\infty$
E	$-\infty$	1	$-\infty$	1	4

=2

$\Rightarrow q_{T,A} = 0.0741$ [from previous calculation]

$$\Rightarrow q_{T,D} = \frac{\sum_{\text{all columns}} f_{T,D}}{\sum_{\text{all columns}} f_{alb}} = \frac{f^2_{T,D}}{18} = \frac{\frac{1}{2}}{18} = \frac{1}{36} = 0.0278$$

$$\Rightarrow q_{T,E} = \frac{\sum_{\text{all columns}} f_{T,E}}{\sum_{\text{all columns}} f_{alb}} = \frac{f^2_{T,E}}{18} = \frac{\frac{1}{2}}{18} = \frac{1}{36} = 0.0278$$

$$\Rightarrow q_{T,M} = \frac{\sum_{\text{all columns}} f_{T,M}}{\sum_{\text{all columns}} f_{alb}} = \frac{f^6_{T,M}}{18} = \frac{\frac{1}{2}}{18} = \frac{1}{36} = 0.0278$$

$$\Rightarrow q_{T,L} = \frac{\sum_{\text{all columns}} f_{T,L}}{\sum_{\text{all columns}} f_{alb}} = \frac{f^6_{T,L}}{18} = \frac{\frac{1}{3}}{18} = \frac{1}{54} = 0.0185$$

Substitute these values in (4)

$$P_T = 0.0278 + \frac{1}{2}\left[0.0185 + 0.056 + 0.0741 + 0.0278 + 0.0278 + 0.0278 + 0.0185\right]$$

$$= 0.0278 + 0.12525$$

$$= 0.15305$$

$$S_{A,T} = \log_2\left(\frac{q_{A,T}}{2 P_A P_T}\right)$$

$$= \log_2\left(\frac{0.0741}{2 \times 0.09265 \times 0.15305}\right)$$

$$= 1.3856$$

05

=5

BT-301 KEYS (SEPTEMBER 2024)

1) a) The Compositional Complexity (K) is given by

$$K = \frac{1}{L} \log_N \left(\frac{L!}{\prod_i n_i!} \right) \longrightarrow (1)$$

where $L \rightarrow$ Length of the given sequence

$N \rightarrow$ No of different kinds of residues

$N = 4$ for nucleotide sequences

0.5

$n_i \rightarrow$ No of nucleotides of 'i'th kind

(i) $K(GCGACT) = \frac{1}{6} \log_4 \left(\frac{6!}{n_A! n_C! n_G! n_T!} \right)$

$n_A = 1 \quad n_G = 2$

$n_T = 1 \quad n_C = 2$

$= \frac{1}{6} \log_4 \left(\frac{6!}{1! 2! 2! 1!} \right)$

0.5

$= \frac{1}{6} \log_4 \left(\frac{6 \times 5 \times 4 \times 3 \times 2!}{2! \times 2!} \right)$

$= 0.6243$

(ii) $K(TATATA) = \frac{1}{6} \log_4 \left(\frac{6!}{n_A! n_C! n_T! n_G!} \right)$

$n_A = 3 \quad n_G = 0$

$n_T = 3 \quad n_C = 0$

$= \frac{1}{6} \log_4 \left(\frac{6!}{3! \times 3!} \right)$

0.5

$= \frac{1}{6} \log_4 \left(\frac{6 \times 5 \times 4 \times 3!}{3! \times 3!} \right)$

$= 0.3601$

0.5 The sequence TATATA ($K = 0.3601$) is less complex than the sequence GCGACT (0.6243)

=2

From (1),

In order to calculate $S_{A,T}$, we have to calculate $q_{A,T}$ (non-random model), & P_A & P_a (random model).

$$q_{A,T} = \frac{\sum_{\text{all columns}} f_{A,T}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^2_{A,T} + f^6_{A,T}}{18} = \frac{\frac{3}{6} + \frac{5}{6}}{18}$$

0.5

$$= \frac{8}{108} = 0.0741$$

$$P_A = q_{AA} + \frac{1}{2} \sum_i q_{A,i} \longrightarrow (2)$$

$$= q_{AA} + \frac{1}{2}\left[q_{A,G} + q_{A,S} + q_{A,M} + q_{A,L} + q_{A,T}\right] \longrightarrow (3)$$

$$\Rightarrow q_{AA} = \frac{\sum_{\text{all columns}} f_{A,A}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^6_{A,A}}{18} = \frac{\frac{1}{6}}{18} = \frac{1}{108} = 0.0091$$

$$\Rightarrow q_{A,G} = \frac{\sum_{\text{all columns}} f_{A,G}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^1_{A,G} + f^6_{A,G}}{18} = \frac{\frac{1}{2} + \frac{1}{6}}{18} = \frac{\frac{4}{6}}{18} = \frac{4}{108}$$

$$= 0.037$$

$$\Rightarrow q_{A,S} = \frac{\sum_{\text{all columns}} f_{A,S}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^1_{A,S} + f^2_{A,S}}{18} = \frac{\frac{1}{2} + \frac{1}{6}}{18} = \frac{\frac{4}{6}}{18} = \frac{4}{108}$$

$$= 0.037$$

(5)

G - INDEL penalty

$S(x_i, y_j)$ - Match or mismatch score for aligning x_i vs y_j.

Make Initialization & filling

		I	N	D	I	G	O
	-	-6	-12	-18	-24	-30	-36
-	0	5	-1	-7	-13	-19	-25
I	-6	10	4	-2	-8	-14	
N	-12	15	9	3	-3		
D	-18	20	14	8			
I	-24	18	12				
A	-30	16					
N	-36						
A	-42						

Trace back

		I	N	D	I	G	O
	-	-6	-12	-18	-24	-30	-36
-	0	5	-1	-7	-13	-19	-25
I	-6	10	4	-2	-8	-14	
N	-12	15	9	3	-3		
D	-18	20	14	8			
I	-24	18	12				
A	-30	16					
N	-36						
A	-42						

of given sequences
atistics is not

ore for the given
d to be calculate

quence.

0.1883

0.1883

calculated
e (1) to the
an distance

shared by more than one sequences, then that distance should be divided by the number of shared sequences

Weighted average score (S) is calculated as follows:

$$= \frac{\sum_{i,j} W_i W_j S(x_i, y_j)}{N}$$

where W_i & W_j are the weights of sequences i & j, respectively

0.5

$S(x,y)$ → Score for aligning character x_i & y_j

N → Number of partial alignments

$$= \Big[W_1 \times W_4 \times S(A,1) + W_1 \times W_5 \times S(A,4) + W_2 \times W_4 \times S(E,1) + W_2 \times W_5 \times S(E,4) + W_3 \times W_4 \times S(D,1) + W_3 \times W_5 \times S(D,4) \Big] / 6$$

$$= \Big[0.3283 \times 0.355 \times 5 + 0.3283 \times 0.4 \times 8 + 0.1883 \times 0.355 \times (-7) + 0.1883 \times 0.4 \times (-5) + 0.1883 \times 0.355 \times (-3) + 0.1883 \times 0.4 \times (-5) \Big] / 6$$

$$= \Big[0.5827 + 1.0506 - 0.4679 - 0.3766 - 0.2005 - 0.3766 \Big] / 6$$

$$= 0.2117 / 6$$

$$= 0.0353$$

(1)

/ -3

(15)

$$\Rightarrow q_{A,M} = \frac{\sum_{\text{all columns}} f_{A,M}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^6_{A,M}}{18} = \frac{1/6}{18} = \frac{1}{108} = 0.009$$

$$\Rightarrow q_{A,L} = \frac{\sum_{\text{all columns}} f_{A,L}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^6_{A,L}}{18} = \frac{\frac{1}{6}}{18} = \frac{1}{108} = 0.0092$$

$$\Rightarrow q_{A,T} = \frac{\sum_{\text{all columns}} f_{A,T}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^2_{A,T} + f^6_{A,T}}{18} = \frac{3/6 + 5/6}{18} = \frac{8}{108}$$

$$= 0.0741$$

Substitute these values in (3)

$$P_A = q_{AA} + \frac{1}{2}\left[q_{A,G} + q_{A,S} + q_{A,M} + q_{A,L} + q_{A,T} \right] \quad (1)$$

$$= 0.0093 + \frac{1}{2}\left[0.037 + 0.037 + 0.0093 + 0.0093 + 0.0741 \right]$$

$$= 0.0093 + 0.08335 = 0.09265$$

$$P_T = q_{T,T} + \frac{1}{2}\left[q_{T,G} + q_{T,S} + q_{T,A} + q_{T,D} + q_{T,E} + q_{T,M} + q_{T,L} \right] \rightarrow (4)$$

$$\Rightarrow q_{TT} = \frac{\sum_{\text{all columns}} f_{TT}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^2_{T,T}}{18} = \frac{\frac{1}{2}}{18}$$

$$= \frac{1}{36} = 0.0278$$

$$\Rightarrow q_{T,S} = \frac{\sum_{\text{all columns}} f_{T,S}}{\sum_{\text{all columns}} f_{a,b}} = \frac{f^1_{T,S} + f^2_{T,S}}{18} = \frac{\frac{1}{2} + \frac{1}{2}}{18} = \frac{1}{18} = 0.056$$

$$\Rightarrow q_{T,G} = \frac{\sum_{\text{all}} f_{T,G}}{\sum_{\text{all}} f_{a,b}} = \frac{f^6_{T,G}}{18} = \frac{\frac{1}{3}}{18} = \frac{1}{54} = 0.0185 \quad (6)$$

query & db, respectively

$e^{-0.229 \times 20}$

greater is the

not statistically

Observed frequency of
background frequency of

25.

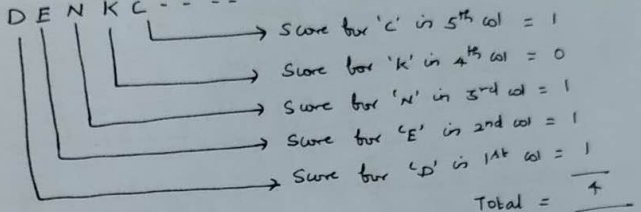
C5
1
1
1
∞
+

(b) We have to calculate the sum of log odds score after aligning the given sequence with the profile for different offset. 0.5

For example, Offset 1

DENKCDEKD (Given sequence)

DENKC----



Sum of all scores will give the log odds score for aligning 'DENKC' with the profile at offset 1.

Similarly log odds score will be calculated for different offsets by sliding the window by one nucleotide.

Offset 2

$S_{ENKCD} = -\infty + 1 + 1 + 2 - \infty$ = Some very long value (indeterminate)

Offset 3

$S_{NKCDE} = 3.585 + 3.8074 + 3.321 + 3.8074 + +$

$= 18.5208$

Offset 4

$S_{KCDEK} = 0 + 1 + 1 + 1 + 1 = 4$

(11)