Indian Statistical Institute

M.Tech. (CS), Second Year, Mid-Sem of First Semester Examination, 2024-25

Computational Molecular Biology and Bioinformatics

Answer Keys

1. (a) Out of the 64 codons, 1 corresponds to start, 2 correspond to stop, and the rest correspond to 20 amino acids. Given there is a many-to-one mapping, a codon and its mutated version might correspond to the same amino acid. Hence, the translated protein will remain the same.

   (b) A Poly-A tail is ideally shorter than 200nt in length and consists of successive ‘A’s. Therefore, it cannot include CpG islands.

   (c) Its length is more than 200bp, GC content is more than 50%, and the observed Obs/Exp CpG is 900*(450/(180*270)) = 8.33, which is more than 0.6. Hence, it is a CpG island.

2.

	–	**T**	**A**	**G**	**A**	**T**
–	0	-3	-6	-9	-12	-15
T	-3	(+2) 2	(-2) -1	(-2) -4	(-2) -7	(+2) -10
A	-6	(-2) -1	(+2) 4	(-2) 1	(+2) -2	(-2) -5
T	-9	(+2) -4	(-2) 1	(-2) 2	(-2) -1	(+2) 0
A	-12	(-2) -7	(+2) -2	(-2) -1	(+2) 4	(-2) 1
G	-15	(-2) -10	(-2) -5	(+2) 0	(-2) 1	(-2) 2

The best alignment is as follows:

T A T A G
T A G A T

The best alignment score is (+2) + (+2) + (-2) + (+2) + (-2) = 2.

3. (a)

^HNCNHH$	^HNCNHH$ $^HNCNHH H$^HNCNH HH$^HNCN NHH$^HNC CNHH$^HN NCNHH$^H HNCNHH$^	CNHH$^HN HH$^HNCN HNCNHH$^ H$^HNCNH NCNHH$^H NHH$^HNC ^HNCNHH$ $^HNCNHH	NN^HHC$H

The transformed sequence is: NN^HHC$H.

   (b) Trie can be used for better implementing lexicographic sorting.

4. (a) $p(l,d)=\sum_{i=0}^{d}\binom{l}{i}\left(\frac{19}{20}\right)^{i}\left(\frac{1}{20}\right)^{l-i}$.

(b) This can be done by thresholding (based on the frequency threshold F, based on the uniqueness threshold U) and by statistical significance analysis (using z-score, using abundance score, using significance profile).