1. We have thousands of SARS-CoV-2 genome sequences available in the database. What is the advantage gained by re-sequencing the viral genome? **[1]**

Soln: understand how the virus is evolving, understand its migration pattern, identify strain-specific mutations for drug/vaccine development.

2. What information can be obtained by comparing sequences? Give two applications of sequence comparison. **[1+1]**

Soln: extent of shared similarity along the length of the two sequences
Function extrapolation - if two proteins share significant seq similarity
Identifying species – as in the case of DNA barcoding
Phylogenetic analysis – to find evolutionary relatedness between species
Genome comparison between individuals in a population – for structural variation analysis
Genome comparison between diseased (e.g., cancer) and normal cells – for identifying variations responsible for the disease
Genome comparison between species – for understanding genome evolution

3. Why does DNA have a helical form? **[1]**

Soln: if the bases are stacked not directly on top of each other but offset slightly and rotated, base pairs can fit snugly on top of each other and minimize the destabilizing effect of water molecules.

4. DNA and RNA polymerases both carry out template-dependent nucleotide polymerization. **[1]**
   True
   False

Soln: True

5. In a sample of double-stranded DNA, 10% of the nitrogenous bases are guanine (G). What percentage of the nitrogenous bases in the sample are adenine (A)? **[1]**

A. 90%
B. 80%
C. 10%
D. 40%

Soln: D: 40%

6. Which of the following statements correctly describes each new molecule of DNA produced after DNA replicates? **[1]**
   A. Each strand of DNA is combined with a strand of RNA.
   B. Each new molecule retains the A, C, and G bases in the DNA sequence, but replaces the T base with U.
   C. Each new molecule is half the length of the original molecule.
   D. Each new molecule contains one strand from the parent molecule and one newly synthesized strand.

Soln: D

7. Many antibiotics interfere with the transfer of genetic information from RNA to protein, preventing bacteria from growing. Which of the following processes is affected by antibiotics? **[1]**
   A. Transmission

B. Translation
C. Replication
D. Transcription

Soln. B

8. What is the probability of finding a long homopolymer region (AAAAAAAAAAAA) in a random DNA sequence? Would you expect its probability to be lower or higher in a protein coding gene sequence? Give reasons to support your answer. **[1+1]**

Soln: Probability of a homopolymer of A, 12 bases long in a random DNA sequence with equiprobability of base composition is $(1/4)^{12}$.
In a coding sequence, this corresponds to a contiguous stretch of 4 Lysine residues in the protein sequence which is highly unlikely since long homopolymer would not encode proteins having biochemical function. Hence its probability will be much lower in coding region. Further, most genes are GC-rich, and the homopolymer A freq will be further lower compared to a random sequence.

9. For a sequence with 60% GC content, find the probability of observing the codons TTT and TTC coding for the amino acid Phe. Compute the relative frequency of observing the codon TTT. **[1+1]**

Soln: P(T) = 0.2, P(C)=0.3, p(TTT) = 0.2x0.2x0.2 = .008, p(TTC) = 0.2x0.2x0.3 = 0.012
Rel freq of observing TTT = p(TTT)/(p(TTT)+p(TTC) = 0.008/(0.012+0.008) = 0.008/0.02 = 8/20 = 4/10 = 0.4

10. Which of the following statements are true: **[1]**
    A. Codon usage differ from organism to organism in genic region.
    B. For eukaryotes gene regions have a different base composition than non-genic regions
    C. Highly expressed genes have different codon usage frequencies, compared to other genes in a genome.
    D. Only statements A and C are true.
    E. All the statements A, B, C

Soln: E

11. If you were to digest the DNA with a restriction enzyme such as Hind III (AAGCTT), approx. how many fragments would be obtained in a genome that is 50% G+C and 5Mb in size? Will the number of fragments be the same if the genome is circular or linear? **[1+1]**

Soln: Prob of observing this pattern AAGCTT is L* p(A)p(A) p(G) p(C)p(T)p(T), where L is the length of the sequence.
$5x10^6x(0.25)^6$ = = 5 x 10^6 x 0.000244140 = 5 x 244.14 = 1220.7 ~1221.
No. of fragments (linear) = 1222 (n is the obsd freq, n+1 – no. of fragments)
No. of fragments (circular) = 1221 (n is the obsd freq, n – no. of fragments)

12. What is the minimum asymptotic amount of space needed to compute the score of a global alignment between sequences of lengths N and M, ignoring traceback? Describe briefly how this is done. **[1+1]**

Soln: If only the score of the alignment is required, we need to keep traceback of the pointers. In this case we need not store (N x M x 3) matrix in memory. Only keeping two rows is sufficient, i.e. (2 x M).

13. Genes are stable entities and inherited from generation to generation. Are mutated genes also stable entities and inherited in the same way as normal genes? **[1]**

    YES

    NO

Soln: Yes

14. A template strand of DNA is 3′ TACCGATTGCA 5′. Which of the following DNA strand is created from this template during replication? **[1]**
    A. 5′ TGCAATGCCTA 3′
    B. 5′ ATGGCTAACGT 3′
    C. 5′ TAGGCATTGCA 3′
    D. 5′ AUGGCUAACGU 3′

Soln: B

15. Give a simple approach for finding restriction recognition sites in a DNA sequence. **[1+1]**
Soln: using the property the Res are palindromic sequences, consider a window of length k, and check if the first and last character complementary to each other, next two, and so on as you move inwards.

16. To construct a physical map of a genome would you consider digesting the genome with a single restriction enzyme or with multiple restriction enzymes. Give reasons to support your answer. **[1+1]**
Soln: single digests will only determine which fragments are present in the unknown DNA. To order and orient the fragments correctly, multiple digests is required.

17. Why two primers are used in PCR amplification, but only one in sequencing by PCR? **[1]**
Soln: For geometric amplification of a DNA fragment. In sequencing, we need to sequence only one strand of DNA, the other can be obtained by complementarity rule.

18. For sequencing a novel bacterial species, would you consider clone-by-clone hierarchical shotgun sequencing method or whole genome shotgun sequencing method? Gives reasons. **[1+1]**
Soln: whole genome shotgun sequencing method. As there are fewer repeat regions in a bacterial genome, and bacterial genomes are also smaller, there is no need to go through the laborious procedure of going through clone-by-clone hierarchical shotgun sequencing method.

19. Using third generation sequencer which generates reads of length 10K bases long, 10MB of reads were generated for a human genome (3 Billion bases). Give the coverage of the genome obtained. What is the time complexity of aligning 1MB reads to the human genome? **[1+1]**
Soln: Reqd reads = Genome Length x Desired coverage/Read Length
Coverage = Number of Reads x Read Length/Genome Length
$C = 10 \times 10^6 \times 10000/3 \times 10^9 = 3.33 \times 10 = 33.33$
Time complexity $O(NxM) = 10 \times 10^6 \times 10000 \times 3 \times 10^9 = 3 \times 10^{20}$.

20. Give affine scoring scheme for penalizing gaps in sequence alignment. What is the advantage of using it instead of linear scoring scheme for penalizing gaps? **[1+1]**

Soln: To treat long insertions/deletions as a single mutational event instead of separate g mutational events. Since, if the occurrence of a mutation is a random event then the probability of continuous residues being inserted or deleted by g random mutational events is highly improbable.

21. Give one example when (i) memory space and (ii) time can be an issue when using dynamic programming approach for sequence alignment. **[1+1]**

Soln: (i) memory space when comparing large genomic sequences
(ii) during database search

22. What information can be obtained by the following k-mer analysis: (i) k=1 and (ii) k=3? **[2+2]**

Soln: (i) k=1: identifying the replication of origin by computing GC skew: (#G - #C)/ (#G+#C), melting temperature of DNA, GC content of the sequence – which can help in identifying protein-coding regions, identifying horizontally transferred regions
(ii) k=3: identify protein-coding regions as triplet frequencies differ in coding regions compared to non-coding regions, identify highly expressed genes by computing codon adaptation index (CAI) since different gene classes have different codon usage frequencies, codon usage differs from organism to organism and can be useful in identifying horizontally transferred regions

23. Give a simple strategy to identify self-complementarity regions in a RNA sequence using Dotplots. **[1]**

Sol. Diagonal lines

24. Give the boundary conditions and recursive relation for obtaining all possible conserved regions between a pair of sequences. Why would you be interested in identifying all sub-optimal matches? **[1+1+1]**

Soln: Boundary Conditions: F(0,j) = 0

$$F(i,0) = max \begin{cases} F(i-1,0) \\ F(i-1,j) - T \end{cases} \quad j = 1,...m$$

Recurrence Relation:

$$F(i,j) = max \begin{cases} F(i,0) \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

If one or both the sequences are long, it is possible to have many different local alignments with a significant score, and one may be interested in all of these.
Example - many copies of a repeated domain or motif in a protein, in comparisons of multi-domain proteins, distantly related sequences that have more than one conserved regions.