DEPARTMENT OF STATISTICS

SAVITRIBAI PHULE PUNE UNIVERSITY

ST-019: STATISTICAL METHODS FOR BIOCOMPUTING

INTERNAL ASSIGNMENT

VIRUS NAME – FLAVIVIRIDAE

SUBMITTED BY,

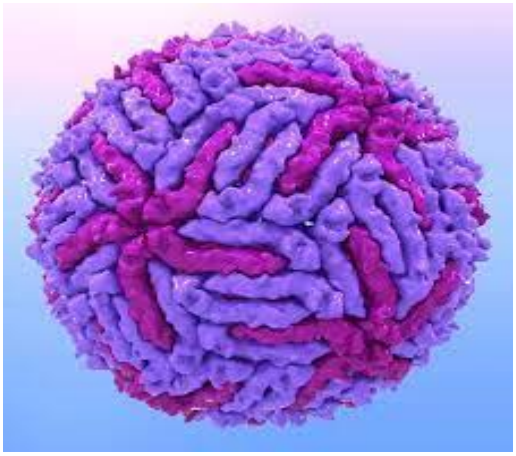
1. **JAGDISH PATIL (2141)**
2. **VIPUL DHAMALE (2112)**

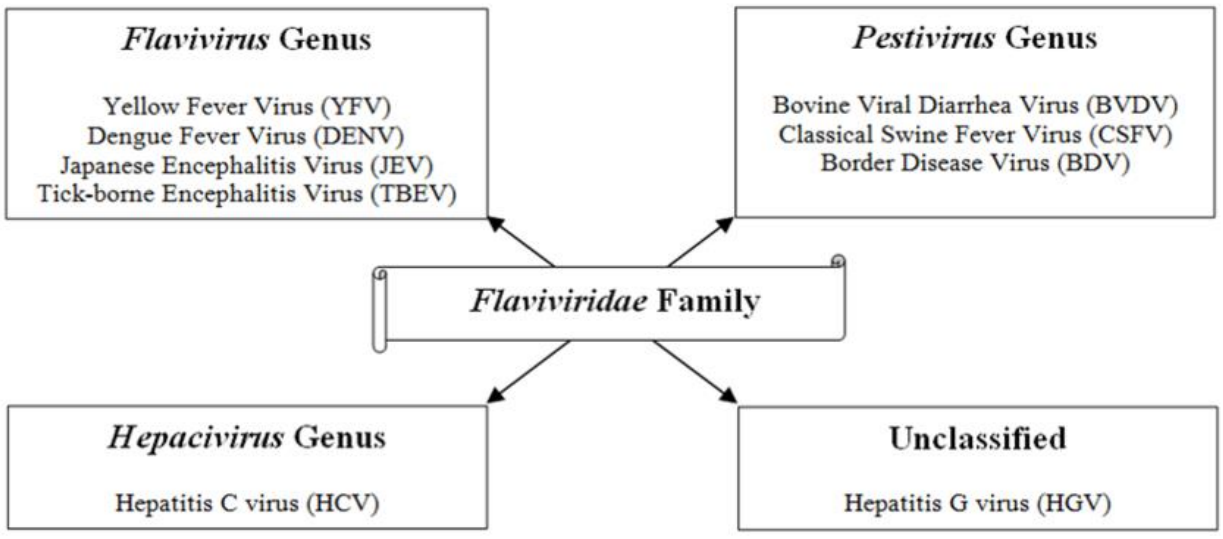
GUIDE,

DR. MOHAN KALE SIR

INTRODUCTION

The Flaviviridae family is a group of single-stranded RNA viruses that includes several important human pathogens such as Dengue virus, Zika virus, yellow fever virus, and West Nile virus. The genomic sequence of Flaviviridae viruses typically ranges from about 10,000 to 12,000 nucleotides in length and encodes a single polyprotein that is processed into several structural and non-structural proteins.





DATA DESCRIPTION AND MODIFICATIONS –

NO.	SEQUENCE NAME	LENGTH
1.	KRV	11375
2.	LGTV	10943
3.	DENV-4	10648
4.	KUNV	10664
5.	TABV	10053
6.	MMLV	10690
7.	MODV	10600
8.	JEV	10976
9.	DEN3	10943
10.	GBV-C	9250

SOFTWARE AND PACKAGES USED –

- R-STUDIO

PACKAGES USED:

library(seqinr)

library(utils)

library(infotheo)

library(phangorn)

library(ape)

library(assertthat)

library(igraph)

Q.1 Compute entropy for each sequence. Also compute mutual information content between every pair of sequences by taking

- **First 10% terms.**
- **Middle 10% terms.**
- **Last 10% terms.**
- **Complete sequence.**

Adjust this proportion to equal length by approximately adding or removing some terms. Comment on the result. Store the result in appropriate format.
Ans –

In genomics, A, C, G, and T are the four nucleotide bases that make up DNA molecules. These bases pair up to form the "rungs" of the DNA double helix ladder.

A represents Adenine, which pairs with Thymine (T).

C represents Cytosine, which pairs with Guanine (G).

The specific sequence of these four nucleotides in a DNA molecule determines the genetic information stored in that DNA.

Proportion of A,C,G,T present in each sequence is given below:

	a	c	g	t
s1	0.251077	0.238945	0.263824	0.246154
s2	0.246185	0.222425	0.320662	0.210728
s3	0.309917	0.207833	0.263054	0.219196
s4	0.273443	0.219524	0.28629	0.220743
s5	0.332438	0.169104	0.21516	0.283299
s6	0.295603	0.181385	0.260056	0.262956
s7	0.292925	0.182736	0.271887	0.252453
s8	0.277606	0.230594	0.283619	0.208182
s9	0.321335	0.207087	0.259256	0.212322
s10	0.190378	0.263351	0.304	0.24227

What is entropy?

In genomics, entropy refers to the degree of randomness or uncertainty in the distribution of nucleotides (A, C, G, and T) along a genomic sequence. The entropy of a genomic sequence is a measure of its information content and can be used to quantify its complexity or level of organization. In DNA sequences, the entropy can be calculated by measuring the diversity or frequency of nucleotides at each position in the sequence. If the distribution of nucleotides is highly biased or uneven, then the entropy will be low, indicating a low level of complexity or information content. On the other hand, if the distribution of nucleotides is more uniform or random, then the entropy will be high, indicating a higher level of complexity or information content. The entropy of genomic sequences can be useful for a variety of applications, including identifying regions of functional importance such as promoters, enhancers, and protein-coding regions, as well as for detecting signatures of natural selection, mutation, and evolutionary processes. Additionally, the entropy of genomic sequences can be used in comparative genomics to study differences and similarities between different species or populations.

Let X be a random variable that can assume values x1, x2, x3...xm with probabilities p1,p2,p3...pm. The entropy or expected information contents of X is,
H(X) = -∑ □ □□ □ □ □ 2 (□ □)
It is also called Shannon's entropy.
The range of entropy is Os Entropy s log(n) where n is the length of sequences.

	Entropy
Seq1	1.385639
Seq2	1.372266
Seq3	1.373537
Seq4	1.378992
Seq5	1.354530
Seq6	1.371418
Seq7	1.371869
Seq8	1.378174
Seq9	1.369890
Seq10	1.372617

INTERPRETATION:
We can see that entropy values of all 10 sequences are nearly equal. Hence there is same amount of uncertainty and complexity present in all sequences

What is mutual information gain?

Mutual information content (MIC) is a measure of the dependence between two random variables in information theory. In genomics, mutual information content is often used to quantify the degree of co-occurrence or co-evolution between two nucleotide positions in a DNA or RNA sequence alignment.

MIC is defined as the difference between the joint entropy of the two variables and the sum of their individual entropies. More formally, if X and Y are two discrete random variables with probability mass functions p(x) and p(y), respectively, and (X,Y) is their joint distribution, then the mutual information content I(X;Y) is given by:

I(X;Y) = H(X) + H(Y) - H(X,Y)

where H(X) and H(Y) are the entropies of X and Y, respectively, and H(X,Y) is the joint entropy of X and Y. MIC measures the amount of information that X and Y share and provides a quantitative measure of the strength of their association.

In genomics, MIC can be used to identify conserved nucleotide pairs that tend to co-occur or co-evolve across different species or to identify functionally important nucleotide positions that interact with each other in RNA or protein structures.

1]. First 10% terms.

	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9	Seq10
Seq1	1.383	0.005	0.004	0.01	0.002	0.007	0.006	0.005	0.005	0.009
Seq2	0.005	1.372	0.005	0.003	0.006	0.007	0.005	0.006	0.006	0.007
Seq3	0.004	0.005	1.377	0.002	0.005	0.004	0.008	0.011	0.011	0.005
Seq4	0.01	0.003	0.002	1.378	0.004	0.015	0.004	0.002	0.003	0.004
Seq5	0.002	0.006	0.005	0.004	1.349	0.006	0.005	0.005	0.001	0.003
Seq6	0.007	0.007	0.004	0.015	0.006	1.375	0.005	0.012	0.002	0.004
Seq7	0.006	0.005	0.008	0.004	0.005	0.005	1.372	0.003	0.006	0.005
Seq8	0.005	0.006	0.011	0.002	0.005	0.012	0.003	1.378	0.004	0.007
Seq9	0.005	0.006	0.011	0.003	0.001	0.002	0.006	0.004	1.379	0.002
Seq10	0.009	0.007	0.005	0.004	0.003	0.004	0.005	0.007	0.002	1.366

INTERPRETATION:

1. Mutual information between almost all of sequences is close to zero, it means that they are independent and there is no significant similarity or correlation between them.

2. Entropies are same because first terms of all sequences are same.

2]. Middle 10% terms.

	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9	Seq10
Seq1	1.386	0.002	0.003	0.003	0.001	0.004	0.005	0.006	0.005	0.003
Seq2	0.002	1.373	0.005	0.004	0.004	0.006	0.007	0.005	0.006	0.002
Seq3	0.003	0.005	1.371	0.003	0.003	0.004	0.003	0.005	0.006	0.007
Seq4	0.003	0.004	0.003	1.378	0.005	0.01	0.002	0.006	0.008	0.005
Seq5	0.001	0.004	0.003	0.005	1.373	0.007	0.005	0.003	0.005	0.008
Seq6	0.004	0.006	0.004	0.01	0.007	1.366	0.007	0.008	0.006	0.004
Seq7	0.005	0.007	0.003	0.002	0.005	0.007	1.373	0.002	0.005	0.004
Seq8	0.006	0.005	0.005	0.006	0.003	0.008	0.002	1.379	0.003	0.007
Seq9	0.005	0.006	0.006	0.008	0.005	0.006	0.005	0.003	1.36	0.005
Seq10	0.003	0.002	0.007	0.005	0.008	0.004	0.004	0.007	0.005	1.364

INTERPRETATION:

1. Mutual information between almost all of sequences is close to zero, it means that they are independent and there is no significant similarity or correlation between them.

2. Entropies are same because first terms of all sequences are same.

3].Last 10% terms.

	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9	Seq10
Seq1	1.38	0.004	0.002	0.007	0.007	0.006	0.005	0.003	0.005	0.005
Seq2	0.004	1.369	0.004	0.007	0.005	0.005	0.004	0.004	0.013	0.002
Seq3	0.002	0.004	1.377	0.007	0.006	0.005	0.003	0.002	0.003	0.006
Seq4	0.007	0.007	0.007	1.38	0.006	0.004	0.004	0.004	0.007	0.004
Seq5	0.007	0.005	0.006	0.006	1.352	0.002	0.006	0.008	0.001	0.002
Seq6	0.006	0.005	0.005	0.004	0.002	1.381	0.007	0.011	0.01	0.002
Seq7	0.005	0.004	0.003	0.004	0.006	0.007	1.378	0.009	0.011	0.007
Seq8	0.003	0.004	0.002	0.004	0.008	0.011	0.009	1.371	0.006	0.002
Seq9	0.005	0.013	0.003	0.007	0.001	0.01	0.011	0.006	1.369	0.006
Seq10	0.005	0.002	0.006	0.004	0.002	0.002	0.007	0.002	0.006	1.359

INTERPRETATION:

1. Mutual information between almost all of sequences is close to zero, it means that they are independent and there is no significant similarity or correlation between them.

2. Entropies are same because first terms of all sequences are same.

4] Complete

	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9	Seq10
Seq1	1.3853	0.0005	0.0005	0.0004	0.0008	0.0009	0.0007	0.0011	0.0006	0.0005
Seq2	0.0005	1.372	0.0006	0.0005	0.0004	0.0006	0.0015	0.0009	0.0007	0.0005
Seq3	0.0005	0.0006	1.3734	0.0005	0.0008	0.0011	0.0005	0.0016	0.001	0.0007
Seq4	0.0004	0.0005	0.0005	1.3791	0.0007	0.0011	0.0012	0.0005	0.0004	0.0005
Seq5	0.0008	0.0004	0.0008	0.0007	1.3544	0.0015	0.0004	0.0006	0.0003	0.0007
Seq6	0.0009	0.0006	0.0011	0.0011	0.0015	1.3707	0.0013	0.0005	0.0007	0.0005
Seq7	0.0007	0.0015	0.0005	0.0012	0.0004	0.0013	1.3712	0.0005	0.0006	0.0007
Seq8	0.0011	0.0009	0.0016	0.0005	0.0006	0.0005	0.0005	1.3792	0.0006	0.0009
Seq9	0.0006	0.0007	0.001	0.0004	0.0003	0.0007	0.0006	0.0006	1.3697	0.0009
Seq10	0.0005	0.0005	0.0007	0.0005	0.0007	0.0005	0.0007	0.0009	0.0009	1.3726

INTERPRETATION:

1. Mutual information between almost all of sequences is close to zero, it means that they are independent and there is no significant similarity or correlation between them.

2. Entropies are same because first terms of all sequences are same.
3.To check alignment of sequences

Q.2 Using UPGMA algorithm reconstruct a phylogenetic tree topology for this group of sequences with distance function as

1. **Difference between entropy of two sequences.**
2. **Frequency of A, G, C and T based distance function of your choice.**
3. **Any distance function you have chosen. Comment on results**.

ANS –
Terminologies used:

- **Operational taxonomic units (OTUS):** It is method used in microbiology and ecology to classify microorganisms into clusters based on their genetic similarity
- **Phylogenetic tree**: A two dimensional graph depicting nodes and branches that illustrates evolutionary relationships between molecules and organisms
- **Nodes:** The points that connect branches and usually represent the taxonomic units
- **Branches**: A branch connects any two nodes

UPGMA (Un-weighted Pair Group Method with Arithmetic mean):

● Unweighted Pair Group Method with Arithmetic Mean (UPGMA) is a hierarchical clustering algorithm used to construct phylogenetic trees based on genetic distance. UPGMA assumes that the rate of evolution is constant over time and that the distance between two sequences is proportional to the time since they last shared a common ancestor.
●The UPGMA algorithm begins by creating a distance matrix that specifies the genetic distance between all pairs of sequences. The algorithm then proceeds as follows:
1. Find the two sequences that are closest together (i.e.,have the smallest distance) and group them into a cluster.
2. Compute the average distance between the new cluster and all other clusters.
3. Repeat steps 1 and 2 until all sequences are grouped into a single cluster.
●At each step of the algorithm, the distance between clusters is computed as the arithmetic mean of the distances between all pairs of sequences in the two clusters. This is why the algorithm is called "Unweighted Pair Group Method

with Arithmetic Mean."

- The output of the UPGMA algorithm is a rooted tree that represents the evolutionary relationships between the sequences. The height of each node in the tree corresponds o the average genetic distance between the sequences in the cluster represented by that node. The branch lengths represent the evolutionary distance between nodes and are proportional to the genetic distance between the sequences they connect.
- UPGMA is a relatively simple and fast algorithm that is widely used in bioinformatics and molecular biology. However, it has several limitations, including its assumption of a constant rate of evolution and its sensitivity to errors in the distance matrix.

Principle of working: Principle of decreasing similarity. The most similar sequences will be clustered first then next best similar and so on.....

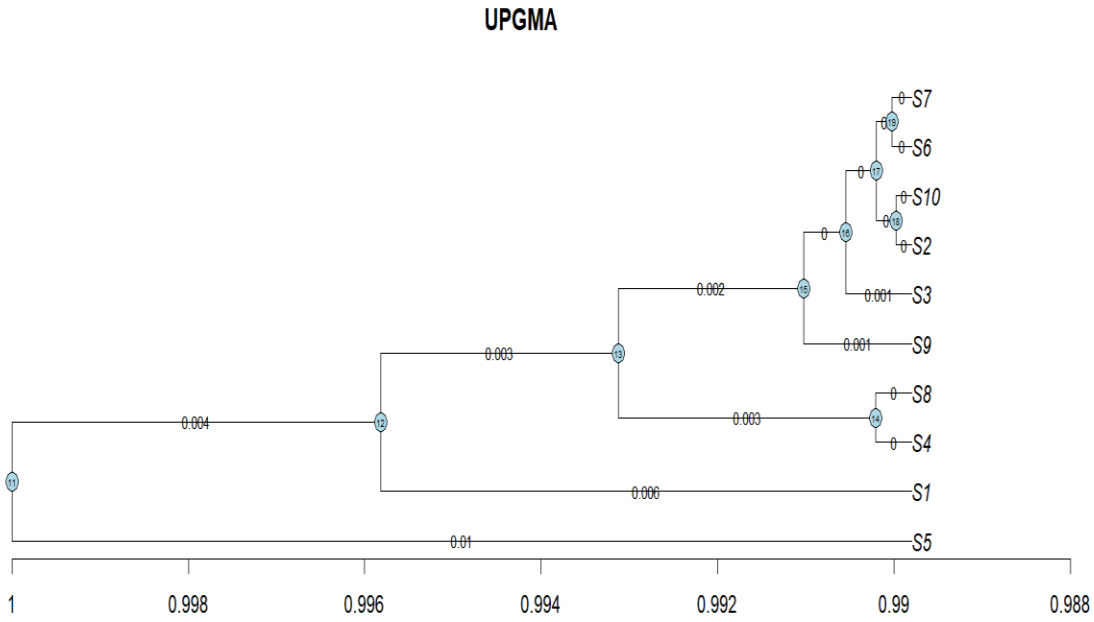
Assumption: The rate of evolution is approximately constant among different lineages so that an approximate linear relationship exists between evolutionary distance and divergence time.

Algorithm:

Step 1) Decide the distance function in an optimal way.
Step 2) Initialization: dij=0 for all i,j
Step 3) Calculate the pairwise distance using a distance function chosen. Step 4) Arrange the pairwise distance function into a matrix having diagonal entries as '0'.
Step 5) Choose a pair of distance from a collection of distance values (distance matrix) di such that min dij = di where i !=j
Step 6) Connect the sequence i" and j" by a branch having length $d_{(i*j*)}/2$ to an ancestor.
Step 7) Recalculate distance matrix D1-((dij*)).
Step 8) If there are sequences left in the database go to step 5 with modified distance matrix
Step 9) Print the tree structure and stop.

✠Difference between entropy of two sequences.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0	0.013	0.012	0.007	0.031	0.014	0.014	0.007	0.016	0.013
S2	0.013	0	0.001	0.007	0.018	0.001	0	0.006	0.002	0
S3	0.012	0.001	0	0.005	0.019	0.002	0.002	0.005	0.004	0.001
S4	0.007	0.007	0.005	0	0.024	0.008	0.007	0.001	0.009	0.006
S5	0.031	0.018	0.019	0.024	0	0.017	0.017	0.024	0.015	0.018
S6	0.014	0.001	0.002	0.008	0.017	0	0	0.007	0.002	0.001
S7	0.014	0	0.002	0.007	0.017	0	0	0.006	0.002	0.001
S8	0.007	0.006	0.005	0.001	0.024	0.007	0.006	0	0.008	0.006
S9	0.016	0.002	0.004	0.009	0.015	0.002	0.002	0.008	0	0.003
S10	0.013	0	0.001	0.006	0.018	0.001	0.001	0.006	0.003	0



INTERPRETATION:

1. Sequence 7 and sequence 6 have evolved from a common ancestor (node 19) with the evolutionary time equal to 0.0000 units.

2. Sequence 10 and sequence 2 have evolved from a common ancestor (node18) with the evolutionary time equal to 0.0000 units.

3. Node 18 and node 19 have evolved from node 17 with evolutionary time equal to 0.00000 units.

4. Sequence 3 and node 17 have evolved from a common ancestor (node 16) with the evolutionary time equal to 0.001 units.

5. Sequence 9 and node 16 have evolved from node 15 with evolutionary time equal to 0.001 units.

6. Sequence 8 and sequence 4 have evolved from node 14 with evolutionary time equal to 0.000 units.

7. Node 15 and node 14 have evolved from node 13 with evolutionary time equal to 0.003 units.

8. Sequence 1 and node 13 have evolved from node 12 with evolutionary time equal to 0.006 units.

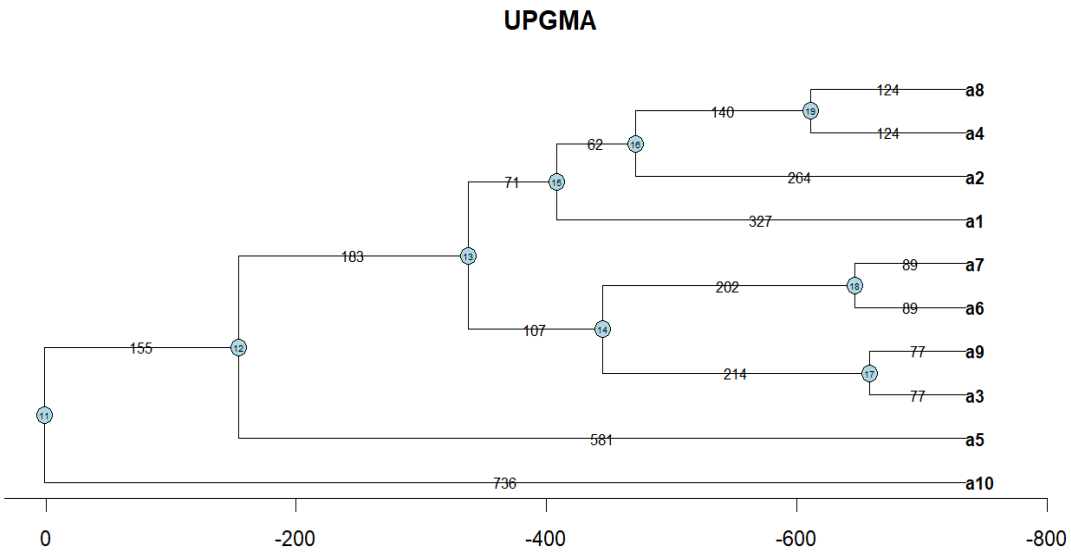
9. Node 12 and sequence 5 have evolved from node 11 with evolutionary time equal to 0.01 units.

10. The evolutionary time for all the 10 sequences from node 11 is 0.01 units.

11. Node 11 is the origin of evolution (first ancestor) for all the sequences 1,2,3,4,5, 6,7,8,9, and 10

✠ Frequency of A, G, C and T based distance function of your choice.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	0	780	842	589	1406	865	838	591	960	1275
a2	780	0	958	518	1750	1117	973	540	1069	1166
a3	842	958	0	477	968	568	488	515	153	1558
a4	589	518	477	0	1276	710	576	248	610	1189
a5	1406	1750	968	1276	0	687	812	1413	990	1957
a6	865	1117	568	710	687	0	178	866	667	1591
a7	838	973	488	576	812	178	0	750	603	1500
a8	591	540	515	248	1413	866	750	0	606	1325
a9	960	1069	153	610	990	667	603	606	0	1691
a10	1275	1166	1558	1189	1957	1591	1500	1325	1691	0



INTERPRETATION:

1. Sequence 8 and sequence 4 have evolved from a common ancestor node 19 .

2. Node 19 and sequence 2 have evolved from a common ancestor node16 .

3. Node 16 and sequence 1 have evolved from a common ancestor node 15 .

4. Sequence 7 and sequence 6 have evolved from a common ancestor node 18 .

5. Sequence 9 and sequence 3 have evolved from a common ancestor node 17.

6. Node 18 and node 17 have evolved from node 14.

7. Node 15 and node 14 have evolved from node 13 .

8. Sequence 5 and node 13 have evolved from node 12 .

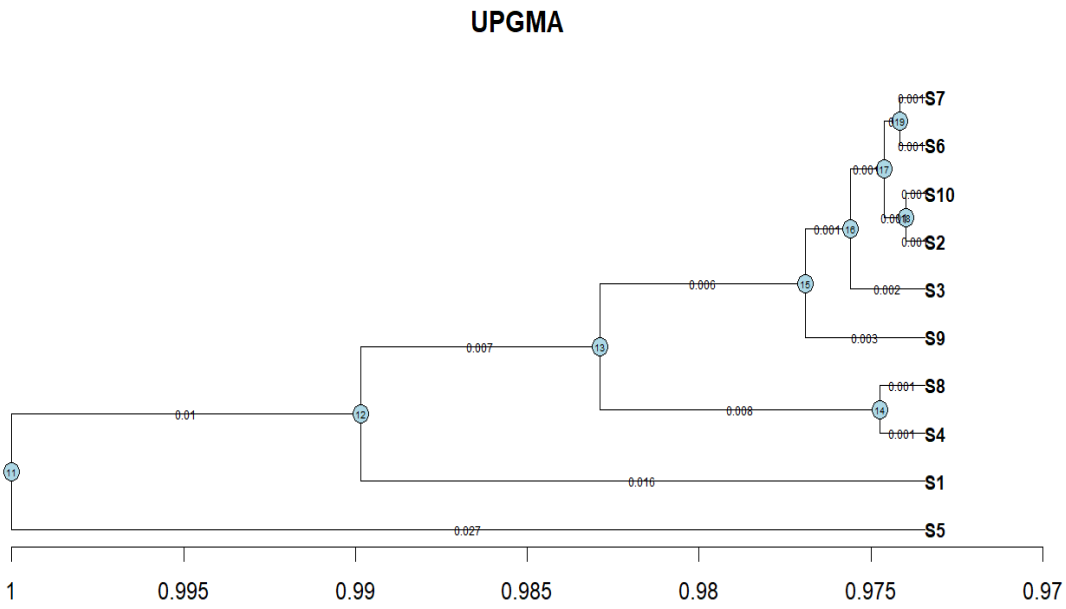
9. Node 12 and sequence 10 have evolved from node 11.

10. The evolutionary time for all the 10 sequences from node 11 is 736 units.

11. Node 11 is the origin of evolution (first ancestor) for all the sequences 1,2,3,4,5, 6,7,8,9 and 10

✠Using Minkowski distance function

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0	0.037	0.034	0.021	0.051	0.037	0.037	0.023	0.037	0.036
S2	0.037	0	0.004	0.02	0.054	0.003	0.001	0.018	0.007	0.001
S3	0.034	0.004	0	0.017	0.055	0.006	0.005	0.015	0.009	0.003
S4	0.021	0.02	0.017	0	0.055	0.021	0.021	0.003	0.022	0.019
S5	0.051	0.054	0.055	0.055	0	0.053	0.053	0.055	0.049	0.054
S6	0.037	0.003	0.006	0.021	0.053	0	0.001	0.019	0.005	0.003
S7	0.037	0.001	0.005	0.021	0.053	0.001	0	0.018	0.006	0.002
S8	0.023	0.018	0.015	0.003	0.055	0.019	0.018	0	0.02	0.017
S9	0.037	0.007	0.009	0.022	0.049	0.005	0.006	0.02	0	0.008
S10	0.036	0.001	0.003	0.019	0.054	0.003	0.002	0.017	0.008	0



INTERPRETATION:

1. Sequence 7 and sequence 6 have evolved from a common ancestor node 19.

2. Sequence 10 and sequence 2 have evolved from a common ancestor node18.

3. Node 18 and node 19 have evolved from node 17.

4. Sequence 3 and node 17 have evolved from a common ancestor node 16.

5. Sequence 9 and node 16 have evolved from node 15.

6. Sequence 8 and sequence 4 have evolved from node 14 .

7. Node 15 and node 14 have evolved from node 13 .

8. Sequence 1 and node 13 have evolved from node 12.

9. Node 12 and sequence 5 have evolved from node 1.

10. The evolutionary time for all the 10 sequences from node 11.

12. Node 11 is the origin of evolution (first ancestor) for all the sequences 1,2,3,4,5, 6,7,8,9 and 10

*Tree caculated using difference between entropy of two sequences and using minkowski distance function are same.

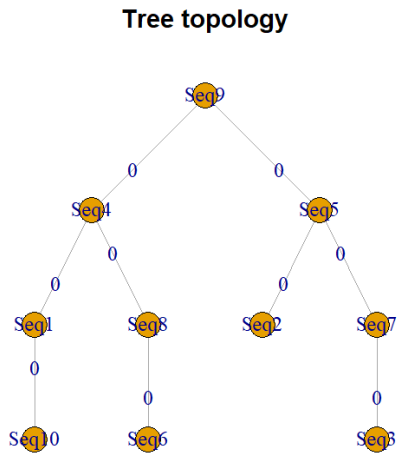
Q.3 Explain how do you use mutual information content to obtain tree topology. Using your suggested algorithm obtain the tree topology for your data.

ANS

Tree Topology- Tree topology refers to the arrangement of branches in a phylogenetic tree, which depicts the evolutionary relationships among a set of biological entities, such as species or genes. In a phylogenetic tree, the topology describes the pattern of branching, including which nodes are connected to which other nodes and the order of these connections. The tree topology is a crucial aspect of phylogenetic analysis and represents the inferred evolutionary relationships among biological entities. It is determined by the method of analysis and can be represented in various forms depending on the analysis and intended use.

Mutual information content (MIC) can be used to obtain a tree topology by identifying the pairwise relationships between sequences based on the amount of shared information. Here is one possible algorithm to obtain a tree topology using MIC:

- Calculate the mutual information content between all pairs of sequences.
- Identify the pair of sequences with the highest mutual information content.
- Connect the two sequences with a branch in the tree.
- Remove the two sequences from the set of sequences and repeat steps 1-3 until only one sequence is left.



Q.4 Select three distance functions of your choice. Obtain the distance matrix for the each one of them. Verify which distance function satisfies ultrametric condition as well as four point condition. Using N-J method obtain tree topology corresponding to each distance function. Comment on the result.

ANS – In the context of phylogenetics, a distance function is a mathematical function that measures the evolutionary distance between two sequences or other biological entities based on their molecular data. The distance function is used to infer the evolutionary relationships between the entities and to construct a phylogenetic tree.

Condition for distance function

1. **Non-negativity**: The distance between any two sequences must be a non-negative value.
2. **Identity**: The distance between a sequence and itself must be zero.
3. **Symmetry**: The distance between sequence A and sequence B must be the same as the distance between sequence B and sequence A.
4. **Triangle inequality**: The distance between sequence A and sequence C must be less than or equal to the sum of the distances between sequence A and sequence B, and between sequence B and sequence C. Formally, this can be written as $d(A,C) \leq d(A,B) + d(B,C)$.
5. **Additivity:** The distance between two sequences that have diverged from a common ancestor must be the sum of the distances between each sequence and the common ancestor. This condition is only applicable to rooted trees.

Ultra-Metric condition- An ultra-metric tree is a tree in which all branches have equal or greater lengths than any branch leading to a more recent node.

Conditions

1. Rootedness: The tree must have a single root node from which all branches emerge.
2. Equal branch lengths: All branches in the tree must have equal lengths.
3. Terminal nodes: Each leaf node of the tree must represent a distinct entity (e.g., a sequence or a species).
4. Additivity: The distance between any two leaf nodes must be the sum of the branch lengths connecting them to their most recent common ancestor.
5. Ultrametricity: The distance from the root node to any leaf node must be the same for all leaf nodes.

Rooted and Unrooted Dendrogram- A rooted dendrogram has a single designated root node, which represents the most recent common ancestor of all the sequences or organisms included in the tree. The root node divides the tree into two halves, with each half representing a distinct lineage that diverged from the common ancestor. Rooted dendrograms are commonly used in evolutionary biology to infer the direction and timing of evolutionary events, such as speciation or gene duplication.

In contrast, an unrooted dendrogram does not have a designated root node or outgroup. Instead, it shows only the relative distances or similarities between the sequences or organisms, without specifying their evolutionary history or directionality. Unrooted dendrograms are useful for visualizing the overall patterns of relationships and for comparing the similarities or differences between different groups of sequences or organisms.

N-J Method

The Neighbor-Joining (NJ) method is a popular algorithm for constructing phylogenetic trees from distance matrices. Here is a general overview of the steps involved in the NJ method:

1. Calculate the pairwise distances between all sequences and construct a distance matrix.
2. Create a phylogenetic tree with all sequences represented by leaf nodes.
3. For each internal node of the tree, calculate the branch lengths that minimize the total branch length of the subtree rooted at that node.
4. Join the two sequences that have the shortest distance, creating a new internal node.
5. Recalculate the distance matrix by replacing the two joined sequences with the new internal node.
6. Repeat steps 3-5 until the tree is fully resolved.

The NJ method is designed to minimize the total branch length of the tree, while maintaining the consistency of the distance matrix. This allows the method to construct accurate trees even in the presence of noise or errors in the distance matrix.

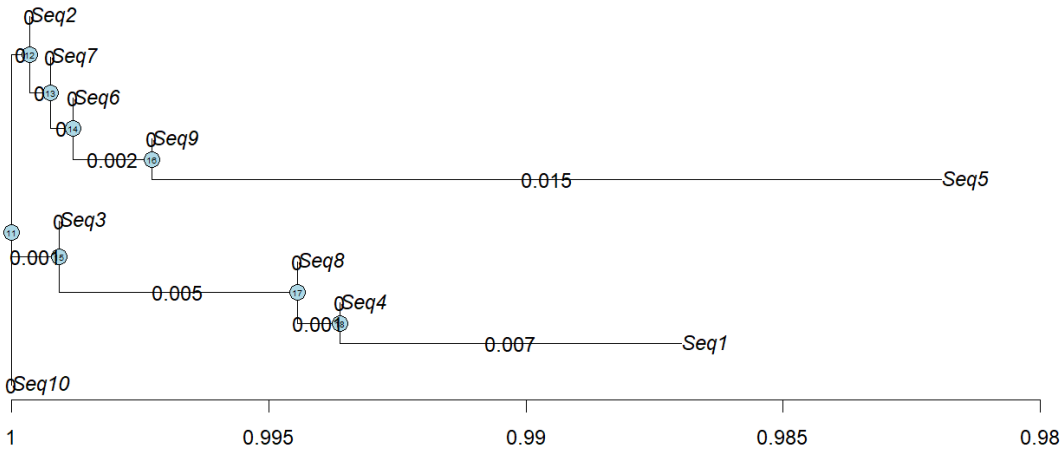
1.Using Euclidean distance function

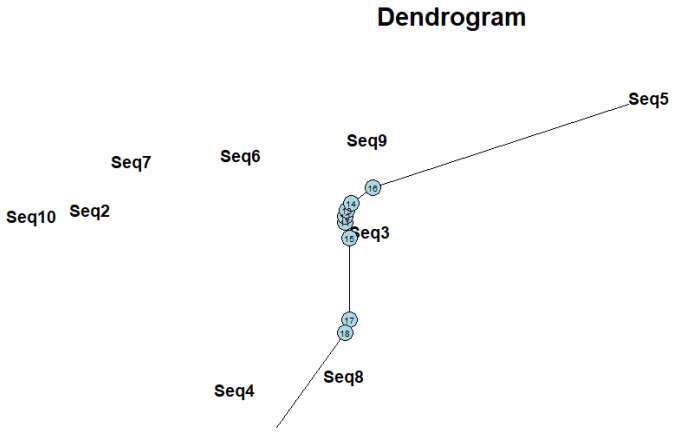
```
distances <- as.matrix(dist1)
# Test the four-point condition
assertthat::assert_that(all(distances >= 0)) # Non-negativity property
[1] TRUE
assertthat::assert_that(all(distances == t(distances))) # Symmetry property
[1] TRUE
 assertthat::assert_that(all(diag(distances) == 0)) # Identity property
[1] TRUE
# Check the triangle inequality property
for (i in 1:nrow(distances))
{
for (j in 1:nrow(distances))
{
for (k in 1:nrow(distances))
{
 if (i != j && j != k && i != k)
{
```

```
 assertthat::assert_that(distances[i, j] <= distances[i, k] + distances[k, j])
}
}
}
}
is.ultrametric(tree1)
[1] FALSE
```

INTERPRETATION:

Distance function satisfies four point condition.But not the ultrametric condition.





Interpretation:

1.Sequence 1 is evolved from the node 18 with the evolutionary time equal to 0.07 units.
2.Sequence 4 is evolved from the node 18 with the evolutionary time equal to 0 units.
3.Node 18 is evolved from the node 17 with the evolutionary time equal to 0 units.
4.Sequence 8 is evolved from the node 17 with the evolutionary time equal to 0 units.
5.Node 17 is evolved from the node 15 with the evolutionary time equal to 0.005 units.
6.Sequence 3 is evolved from the node 15 with the evolutionary time equal to 0 units.
7.Sequence 10 is evolved from the node 11 with the evolutionary time equal to 0 units.
8.Node 15 is evolved from the node 11 with the evolutionary time equal to 0.05 units.
9. Sequence 5 is evolved from the node 16 with the evolutionary time equal to 0.015 units.
10. Sequence 9 is evolved from the node 16 with the evolutionary time equal to 0 units.
11. Node 16 is evolved from the node 14 with the evolutionary time equal to 0.002 units
12. Sequence 6 is evolved from the node 14 with the evolutionary time equal to 0 units
13.Node 14 and sequence 7 are evolve from node 13 with the evolutionary time equal to 0 units
14. Node 13and sequence 2 are evolve from node 12 with the evolutionary time equal to 0 units
15 . Node 12 evolve from node 11

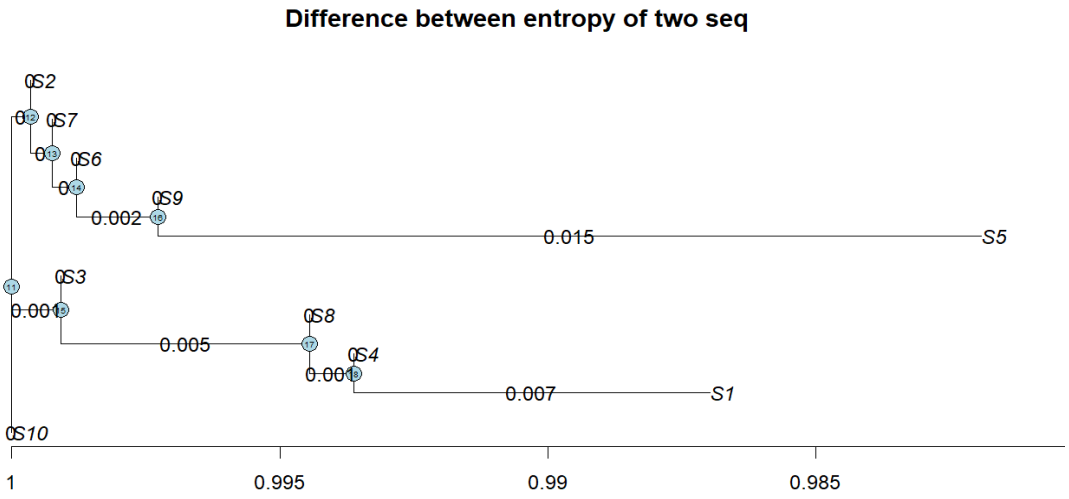
2.Difference between entropy of two sequences.

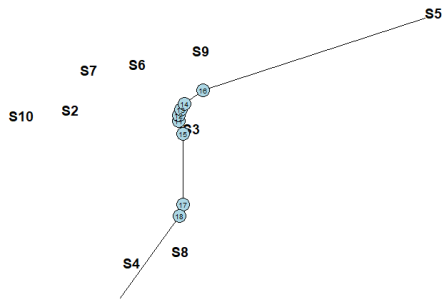
```
# Test the four-point condition

assertthat::assert_that(all(distances >= 0)) # Non-negativity property

[1] TRUE

assertthat::assert_that(all(distances == t(distances))) # Symmetry property

[1] TRUE

assertthat::assert_that(all(diag(distances) == 0)) # Identity property

[1] TRUE

# Check the triangle inequality property

for (i in 1:nrow(distances))

{

for (j in 1:nrow(distances))

 {

 for (k in 1:nrow(distances))

{

if (i != j && j != k && i != k)

{
```

```
assertthat::assert_that(distances[i, j] <= distances[i, k] + distances[k, j])
  }
  }
  }
  }
  is.ultrametric(tree2)
  [1] FALSE
```

INTERPRETATION:

Distance function satisfies four point condition.But not the ultrametric condition.





Interpretation:

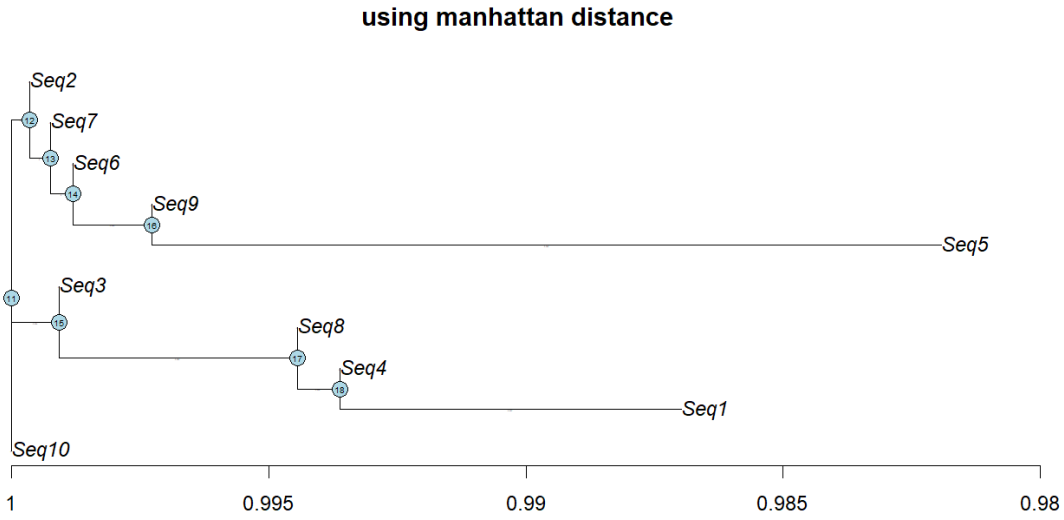
1.Sequence 1 is evolved from the node 18 with the evolutionary time equal to 0.07 units.
2.Sequence 4 is evolved from the node 18 with the evolutionary time equal to 0 units.
3.Node 18 is evolved from the node 17 with the evolutionary time equal to 0 units.
4.Sequence 8 is evolved from the node 17 with the evolutionary time equal to 0 units.
5.Node 17 is evolved from the node 15 with the evolutionary time equal to 0.005 units.
6.Sequence 3 is evolved from the node 15 with the evolutionary time equal to 0 units.
7.Sequence 10 is evolved from the node 11 with the evolutionary time equal to 0 units.
8.Node 15 is evolved from the node 11 with the evolutionary time equal to 0.05 units.
9. Sequence 5 is evolved from the node 16 with the evolutionary time equal to 0.015 units.
10. Sequence 9 is evolved from the node 16 with the evolutionary time equal to 0 units.
11. Node 16 is evolved from the node 14 with the evolutionary time equal to 0.002 units
12. Sequence 6 is evolved from the node 14 with the evolutionary time equal to 0 units
13.Node 14 and sequence 7 are evolve from node 13 with the evolutionary time equal tounits
14. Node 13and sequence 2 are evolve from node 12 with the evolutionary time equal to 0 units
15 . Node 12 evolve from node 11

3.Using Manhatten distance function

Test the four-point condition

```
assertthat::assert_that(all(distances >= 0)) # Non-negativity property
[1] TRUE
assertthat::assert_that(all(distances == t(distances))) # Symmetry property
[1] TRUE
 assertthat::assert_that(all(diag(distances) == 0)) # Identity property
[1] TRUE
# Check the triangle inequality property
for (i in 1:nrow(distances))
{
for (j in 1:nrow(distances))
{
for (k in 1:nrow(distances))
{
if (i != j && j != k && i != k)
{
assertthat::assert_that(distances[i, j] <= distances[i, k] + distances[k, j])
}
}
}
}
 is.ultrametric(tree3)

[1] FALSE
```

INTERPRETATION:
Distance function satisfies four point condition.But not the ultrametric condition.



Interpretation:

1.Sequence 1 is evolved from the node 18 with the evolutionary time equal to 0.07 units.
2.Sequence 4 is evolved from the node 18 with the evolutionary time equal to 0 units.
3.Node 18 is evolved from the node 17 with the evolutionary time equal to 0 units.
4.Sequence 8 is evolved from the node 17 with the evolutionary time equal to 0 units.
5.Node 17 is evolved from the node 15 with the evolutionary time equal to 0.005 units.
6.Sequence 3 is evolved from the node 15 with the evolutionary time equal to 0 units.
7.Sequence 10 is evolved from the node 11 with the evolutionary time equal to 0 units.
8.Node 15 is evolved from the node 11 with the evolutionary time equal to 0.05 units.
9. Sequence 5 is evolved from the node 16 with the evolutionary time equal to 0.015 units.
10. Sequence 9 is evolved from the node 16 with the evolutionary time equal to 0 units.
11. Node 16 is evolved from the node 14 with the evolutionary time equal to 0.002 units
12. Sequence 6 is evolved from the node 14 with the evolutionary time equal to 0 units
13.Node 14 and sequence 7 are evolve from node 13 with the evolutionary time equal to units
14. Node 13and sequence 2 are evolve from node 12 with the evolutionary time equal to 0 units
15 . Node 12 evolve from node 11

Conclusion:
We can see that for all three distance function we had chosen the tree topology is similar.

Q.5 Choosing anyone distance function obtain distance matrix for your data and obtain topological tree for this distance matrix manually using NJ method. Also show important steps.

Q.6 By assuming every sequence is a Markov chain with state space (A, C, G., T) and initial probability distribution P(X = a) = 1/4, a E (A, C,G,T). Obtain the estimates of one step transition probability matrix. Are these Markov chains ergodic? Justify your answer.

ANS –
Definition: A Markov chain is a time-homogeneous Markovian random process which takes values in a state space S
Here S = {A, G, C, T}**.**
'Markovian' means: At each time, only the current state is important for the future: $P(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = P(X_{n+1} = i_{n+1} | X_n = i_n)$.

The states visited are not necessarily numerical. They are A, G, C and T.

Irreducible Markov chain –
A Markov chain in which every state can be reached from every other state is called an irreducible Markov chain.
Aperiodic Markov chain -
The period of a state i is the greatest common divisor of the set {n ∈ : pn (i,i) > 0}. If every state has period 1 then the Markov chain (or its transition probability matrix) is called aperiodic.
Ergodic Markov Chains-
A Markov chain is called an ergodic or irreducible Markov chain if it is possible to eventually get from every state to every other state with positive probability. Ex: The wandering mathematician in previous example is an ergodic Markov chain.

❒**The estimate of one step transition probability matrix for sequence 1**

	a	c	g	t
a	0.2899160	0.2321429	0.2254902	0.2524510
c	0.2928624	0.2538631	0.1994113	0.2538631
g	0.2849050	0.1899367	0.2819060	0.2432522
t	0.1343337	0.2840300	0.3461951	0.2354412

❒**The estimate of one step transition probability matrix for sequence 2**

	a	c	g	t
a	0.2349161	0.2392003	0.3320243	0.1938593
c	0.3312253	0.2422925	0.1608696	0.2656126
g	0.2877915	0.1961591	0.3278464	0.1882030
t	0.1063386	0.2218515	0.4645538	0.2072560

❒**The estimate of one step transition probability matrix for sequence 3**

	a	c	g	t
a	0.3139205	0.2062500	0.2744318	0.2053977
c	0.3962822	0.2543304	0.1035065	0.2458809
g	0.3553248	0.1667783	0.2933691	0.1845278
t	0.1659336	0.2167133	0.3594562	0.2578968

❒**The estimate of one step transition probability matrix for sequence 4**

	a	c	g	t
a	0.2769923	0.2255784	0.2869537	0.2104756
c	0.3408270	0.2340426	0.1573665	0.2677640
g	0.3157248	0.2051597	0.2896192	0.1894963
t	0.1483101	0.2143141	0.4083499	0.2290258

❒**The estimate of one step transition probability matrix for sequence 5**

	a	c	g	t
a	0.3685880	0.1723956	0.20571126	0.2533051
c	0.4161004	0.2195504	0.04966022	0.3146890
g	0.3602434	0.1314402	0.26977688	0.2385396
t	0.2187306	0.1611699	0.28842564	0.3316739

❒**The estimate of one step transition probability matrix for sequence 6**

	a	c	g	t
a	0.3430440	0.1795482	0.2434602	0.2339477
c	0.3775411	0.2115198	0.0856728	0.3252662
g	0.3265789	0.1594056	0.2863897	0.2276258
t	0.1552129	0.1853838	0.3744552	0.2849480

❒**The estimate of one step transition probability matrix for sequence 7**

	a	c	g	t
a	0.3323326	0.1768238	0.25607926	0.2347643
c	0.3778205	0.2189150	0.08929429	0.3139702
g	0.3209076	0.1688817	0.29141005	0.2188006
t	0.1551304	0.1801739	0.39895652	0.2657391

❒**The estimate of one step transition probability matrix for sequence 8**

	a	c	g	t
a	0.2883765	0.2233102	0.2962729	0.1920404
c	0.3372940	0.2414718	0.1682637	0.2529705
g	0.3090062	0.2204969	0.2903727	0.1801242
t	0.1584699	0.2366541	0.3816730	0.2232030

❒**The estimate of one step transition probability matrix for sequence 9**

	a	c	g	t
a	0.3189679	0.2141093	0.2654406	0.2014823
c	0.4186837	0.2250531	0.1116773	0.2445860
g	0.3512411	0.1880313	0.2896974	0.1710303
t	0.1893224	0.2020534	0.3527721	0.2558522

❒**The estimate of one step transition probability matrix for sequence 10**

	a	c	g	t
a	0.2011358	0.2858495	0.2683389	0.2446758
c	0.2187915	0.2991368	0.2081673	0.2739044
g	0.1870524	0.2337439	0.3563449	0.2228588
t	0.1359681	0.2509065	0.3817984	0.2313270

INTERPRETATION:

We can see in each tpm every state can reached to every other state
Hence all tpms are irreducible and aperiodic
Therefore all sequences are ergodic.