Department of Statistics

Savitribai Phule Pune University

ST O19: Statistical Methods for Bio-computing,

Internal Assignment Work II

Jan-May 2023 **Maximum Marks:90**

Note:

- You have to complete following task on or before 15th April, 2023.
- Choose 10 sequences from the data base of sequences provided to you.
- Give the details of these sequences such as accession number, name of the organism, type of organism and type of sequences.
- State formulas and results you have used clearly.
- Describe/present algorithm and flow chart for the procedure you have used.
- Diagrams should be properly labeled whenever necessary.

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. (15)

Q.2 Using UPGMA algorithm reconstruct a phylogenetic tree topology for this group of sequences with distance function as (15)

- Difference between entropy of two sequences.

- Frequency of A, G, C and T based distance function of your choice.
- Any distance function you have chosen. Comment on results.

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. (10)

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. (15)

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. (10)

Q.6 By assuming every sequence is a Markov chain with state space $\{A, C, G, T\}$ and initial probability distribution $P(X = a) = 1/4, a \in \{A, C, G, T\}$. Obtain the estimates of one step transition probability matrix. Are these Markov chains ergodic? Justify your answer. (15)

Q.7 Check whether in each of your sequence there is a CP_G island. (10)

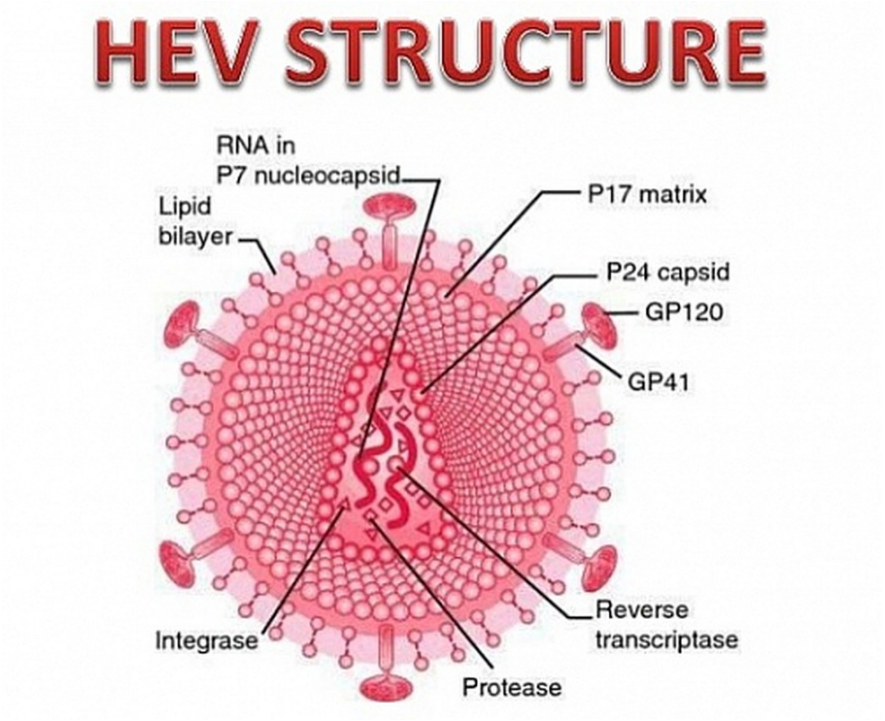
****

ST-019 Statistical Methods for Bio-computing
Internal Assignment Work II

Abhijit Chavan (2108)
Saurav Jadhav (2120)

15.04.2023

Hepatitis E Virus (HEV)



INTRODUCTION TO HEPATITIS E VIRUS (HEV):

Hepatitis E Virus (HEV) is a single-stranded RNA virus that causes Hepatitis E infection. It is a significant cause of acute viral hepatitis in developing countries.

HEV has 4 main genotypes:

- Genotypes 1 and 2 primarily infect humans and are found mainly in developing countries with poor sanitation. They are spread through the fecal-oral route, usually through contaminated water.

- Genotypes 3 and 4 infect both humans and animals like pigs, deer, etc. They cause sporadic cases in both developing and developed countries. Spread is usually through undercooked infected meat.

- HEV infection typically causes an acute illness with symptoms like fatigue, loss of appetite, nausea, vomiting, joint pain, and jaundice. Most people recover within 4-6 weeks, but sometimes it can lead to acute liver failure in pregnant women, the elderly, and those with underlying liver disease.

The HEV Virion is non-enveloped, spherical, and 27-34 nm in diameter. Its genome is a single-stranded, positive-sense RNA molecule of approximately 7.2 kilobases in length. It has 3 open reading frames (ORF1, ORF2, and ORF3) that encode functional proteins.

ORF1 encodes nonstructural proteins like RNA-dependent RNA polymerase, methyltransferase, protease, etc. required for replication and protein processing.

ORF2 encodes viral capsid protein. ORF3 encodes a small protein that may be involved in virus-host interactions.

HEV diagnosis is based on detecting anti-HEV antibodies (IgG, IgM) and viral RNA in blood or stool samples using enzyme immunoassays and reverses transcription-PCR respectively.

There is no specific treatment for HEV. Supportive care and rest are recommended. Ribavirin may be used for some chronic cases. A recombinant vaccine based on the HEV ORF2 antigen is available in China and shown to be effective.

Good sanitation, hygiene, proper cooking of meat, and avoiding contaminated food/water can help prevent HEV infection.

DATA DESCRIPTION:

From the given database, we choose ten random sequences of HEV.

Initially, the sequences were of unequal length. They are all made of equal length.

We considered the sequence which was of the minimum length and all the other sequences were made of that size.

Now, the length of all the sequences is the same and is equal to 2104.

The accession number of those sequences are:

Accession Number
4AB097812
4AB291968
3FJ906896
3bAB291953
3fEU723514
4AB220973
4AB074915
3bAB291955
1AF051351
4GU206559

Table 1.1

These 10 sequences are RNA sequences consisting of 4 nucleotides:
Adenine (A), Guanine (G), Cytosine (C), Thymine (T)

SOFTWARES AND PACKAGES USED:

Softwares used: R-Studio, Google Docs

Packages used:
phangorn
seqinr
ape
Biostrings
markovchain
transport
emdist

Question 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 an appropriate format.

Solution:

Entropy is a measure of the randomness or unpredictability of a system.

Entropy is calculated by:
Let X be a random variable having values $x_1, x_2, x_3, \ldots\ldots, x_m$ with probabilities $p_1, p_2, p_3, \ldots\ldots p_m$. The entropy also known as Shannon's entropy of X is given by,

$$H(X) = \sum_{\forall X_i \in \{A,G,C,T\}} - P_i \log P_i$$

The range of entropy is 0 ≤ Entropy ≤ log(n) where n is the length of sequences. In biological sequence analysis, the higher the entropy value, the higher the uncertainty in the appearance of nucleotides in a particular way which means that the sequence/organism is highly evolving. On the other hand, a low entropy value is an indicator of the sequence/organism is somewhat conserved. Generally, entropy is measured in “bits”. The maximum entropy of a DNA sequence can have been “2 bits”.

We rank the entropy values from high to low in the below table.

Accession Number	Entropy	Entropy-based Rank
4GU206559	1.981666099	1
3bAB291955	1.981648022	2
1AF051351	1.979906087	3
4AB291968	1.979782436	4
4AB097812	1.978882672	5
3fEU723514	1.978417802	6
3FJ906896	1.977849392	7
4AB074915	1.977592751	8
4AB220973	1.976367021	9
3bAB291953	1.963407378	10

Table 1.2

Based on the information provided, the entropy values for the sequences appear to be very similar, indicating that there is a comparable degree of uncertainty and complexity across all of the sequences. This suggests that each sequence contains a roughly equivalent amount of randomness and variation.

Mutual Information Content:

$$I(X,Y) = H(X) - H(X|Y)$$

Mutual information is a measure of the statistical dependence between two random variables. In the context of DNA sequences, mutual information can be used to quantify the degree of similarity between two sequences.

It is calculated by comparing the frequencies of occurrence of different pairs of nucleotides in the two sequences. If the two sequences are independent, then the mutual information between them will be zero. However, if the two sequences are highly similar, then the mutual information will be higher.

Mutual information can be useful in various applications such as predicting gene regulatory networks, identifying conserved regions in multiple sequence alignments, and detecting functional relationships between genes. It provides a quantitative measure of the degree of similarity or correlation between two DNA sequences, which can help in understanding the underlying biological processes that they represent.

Now, as already stated the length of the sequence is 2104. Hence, 10% of that will be 210.

So, for the first 10%, we have the mutual information content between all pairs of sequence as

Mutual Information based on the First 10% Sequence									
1.3740	1.2471	0.0178	0.0091	0.9334	1.2471	0.0170	0.0079	0.0225	1.1249
1.2471	1.3753	0.0203	0.0132	0.9818	1.3275	0.0139	0.0121	0.0282	1.1211
0.0178	0.0203	1.3602	0.0135	0.0239	0.0187	0.0081	0.0136	0.0164	0.0165
0.0091	0.0132	0.0135	1.3732	0.0304	0.0142	0.0376	1.3244	0.0469	0.0115
0.9334	0.9818	0.0239	0.0304	1.3677	0.9568	0.0204	0.0279	0.0268	0.9124
1.2471	1.3275	0.0187	0.0142	0.9568	1.3753	0.0136	0.0130	0.0263	1.1211
0.0170	0.0139	0.0081	0.0376	0.0204	0.0136	1.3783	0.0368	0.0304	0.0184
0.0079	0.0121	0.0136	1.3244	0.0279	0.0130	0.0368	1.3720	0.0456	0.0102
0.0225	0.0282	0.0164	0.0469	0.0268	0.0263	0.0304	0.0456	1.3559	0.0236
1.1249	1.1211	0.0165	0.0115	0.9124	1.1211	0.0184	0.0102	0.0236	1.3698

Table 1.3

Now, from the above table we can observe that:

1. Sequence 1&4 have a low value of mutual information content. Indicating, they are close to independent.

2. Also, sequence 2&6 share high value mutual information, indicating they are very similar to one another.

3. Similarly, sequence 4&8 share high value mutual information, indicating they are very similar to one another.
4. Sequences 10&8 have a low value of mutual information content. Indicating, they are close to independent.

Now we consider Mutual information for the middle 10% of the sequences.

Mutual Information based on Middle 10% Sequence									
1.3604	1.0833	0.0242	0.0445	0.6731	1.1256	0.0196	0.0445	0.0236	0.8705
1.0833	1.3537	0.0331	0.0309	0.6974	1.3097	0.0121	0.0309	0.0213	0.9160
0.0242	0.0331	1.3610	0.0204	0.0220	0.0349	0.0313	0.0204	0.0131	0.0298
0.0445	0.0309	0.0204	1.3490	0.0391	0.0349	0.0080	1.3490	0.0126	0.0331
0.6731	0.6974	0.0220	0.0391	1.3445	0.7166	0.0155	0.0391	0.0155	0.7130
1.1256	1.3097	0.0349	0.0349	0.7166	1.3588	0.0137	0.0349	0.0213	0.9544
0.0196	0.0121	0.0313	0.0080	0.0155	0.0137	1.3505	0.0080	0.0152	0.0091
0.0445	0.0309	0.0204	1.3490	0.0391	0.0349	0.0080	1.3490	0.0126	0.0331
0.0236	0.0213	0.0131	0.0126	0.0155	0.0213	0.0152	0.0126	1.3711	0.0161
0.8705	0.9160	0.0298	0.0331	0.7130	0.9544	0.0091	0.0331	0.0161	1.3537

Table 1.4

Now, from the above table we can observe that:

1. Sequence 7&10 share a low value of mutual information content. Indicating, they are close to independent.

2. Sequence 7&5 share a low value of mutual information content. Indicating, they are close to independent.

3. Sequences 2&6 share high value mutual information, indicating they are very similar to one another.

4. Sequences 4&8 share high value mutual information, indicating they are very similar to one another.

5. It is important to note that for sequences 4&8, for both, first 10% and middle 10% of the part, the Mutual information content is high.

Now we consider Mutual information for the last 10% of the sequences.

Mutual Information based on Last 10% Sequence									
1.3497	0.8571	0.0294	0.0310	0.3483	0.8591	0.0287	0.0310	0.0299	0.6563
0.8571	1.3353	0.0186	0.0149	0.3779	1.2198	0.0199	0.0149	0.0307	0.6750
0.0294	0.0186	1.3593	0.0124	0.0159	0.0185	0.0399	0.0122	0.0085	0.0200
0.0310	0.0149	0.0124	1.3649	0.0265	0.0163	0.0405	1.3169	0.0220	0.0137
0.3483	0.3779	0.0159	0.0265	1.3531	0.3873	0.0251	0.0265	0.0187	0.3730
0.8591	1.2198	0.0185	0.0163	0.3873	1.3394	0.0210	0.0163	0.0338	0.6690
0.0287	0.0199	0.0399	0.0405	0.0251	0.0210	1.3618	0.0467	0.0068	0.0244
0.0310	0.0149	0.0122	1.3169	0.0265	0.0163	0.0467	1.3649	0.0220	0.0137
0.0299	0.0307	0.0085	0.0220	0.0187	0.0338	0.0068	0.0220	1.3512	0.0292
0.6563	0.6750	0.0200	0.0137	0.3730	0.6690	0.0244	0.0137	0.0292	1.3290

Table 1.5

Now, from the above table we can observe that:

1. Sequence 4&8 have a large value of mutual information content, indicating they are very similar to one another.

2. For the sequence 9&7 we have a low value of mutual information content, indicating they are very dissimilar to one another.

3. Sequence 5&1 have a large value of mutual information content, indicating they are very similar to one another.

4. Sequence 5&2 have a large value of mutual information content, indicating they are very similar to one another.

Now we consider Mutual information for the complete sequence:

Mutual Information based on the Complete Sequence									
1.3713	0.9895	0.0077	0.0043	0.5264	0.9992	0.0023	0.0045	0.0040	0.8619
0.9895	1.3709	0.0072	0.0032	0.5395	1.2625	0.0015	0.0035	0.0031	0.8565
0.0077	0.0072	1.3699	0.0018	0.0071	0.0074	0.0028	0.0018	0.0019	0.0086
0.0043	0.0032	0.0018	1.3736	0.0032	0.0033	0.0021	1.3426	0.0015	0.0045
0.5264	0.5395	0.0071	0.0032	1.3723	0.5364	0.0025	0.0035	0.0040	0.5226
0.9992	1.2625	0.0074	0.0033	0.5364	1.3717	0.0014	0.0036	0.0028	0.8647
0.0023	0.0015	0.0028	0.0021	0.0025	0.0014	1.3708	0.0017	0.0015	0.0022
0.0045	0.0035	0.0018	1.3426	0.0035	0.0036	0.0017	1.3736	0.0017	0.0048
0.0040	0.0031	0.0019	0.0015	0.0040	0.0028	0.0015	0.0017	1.3609	0.0034
0.8619	0.8565	0.0086	0.0045	0.5226	0.8647	0.0022	0.0048	0.0034	1.3724

Now, from the above table we can observe that:

1. Sequence 3&8 we have a low value of mutual information content, indicating they are very dissimilar to one another.

2. Similarly, for the sequence 4&9 we have a low value of mutual information content, indicating they are very dissimilar to one another.

3. For the sequence 8&4 we have a high value of mutual information content, indicating they are very similar to one another.

4. For the sequence, 8&9 we have a low value of mutual information content, indicating they are very dissimilar to one another.

5. For the sequence, 2&6 we have a high value of mutual information content, indicating they are very similar to one another.

Question 2

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

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

Solution:

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a hierarchical clustering algorithm commonly used in phylogenetics to construct a phylogenetic tree from a set of molecular sequences. The algorithm is based on the assumption that the rate of evolution is constant over time and across all lineages.

UPGMA works by iteratively merging the two closest clusters into a new cluster, until all sequences are included in the final cluster. The distance between two clusters is calculated as the average distance between all pairs of sequences in the two clusters.

At each iteration, a new node is added to the tree to represent the newly merged cluster. The branch length between the new node and each of its child nodes is set to half the distance between the two clusters being merged. This assumes that the rate of evolution is constant over time and that the distance between two sequences is proportional to the time since they diverged from a common ancestor.

UPGMA is a fast and efficient algorithm that is relatively easy to implement. However, it has several limitations. One of the main limitations is that it assumes a constant rate of evolution over time, which may not be true for all lineages. Additionally, UPGMA does not account for homoplasy (convergent evolution or parallel evolution), which can lead to incorrect tree topologies.

Despite its limitations, UPGMA is still widely used in phylogenetics, particularly for constructing preliminary trees or for clustering sequences based on their overall

similarity. It can also be used as a starting point for more sophisticated algorithms that account for more complex evolutionary models.

- To reconstruct a phylogenetic tree topology using the UPGMA algorithm with the distance function as the difference between the entropy of two sequences, we can follow these steps:

    1. Calculate the pairwise distances between all sequences based on their entropy values.
    2. We consider the absolute differences of the entropy values of the pair of sequences as the pairwise distance.
    3. Create a distance matrix using the pairwise distances.
    4. Identify the two sequences with the smallest distance in the distance matrix, and group them into a new cluster.
    5. Calculate the average distance between the new cluster and all other clusters.
    6. Update the distance matrix with the average distance between the new cluster and all other clusters.
    7. Repeat steps 4-6 until all sequences are included in the final cluster.

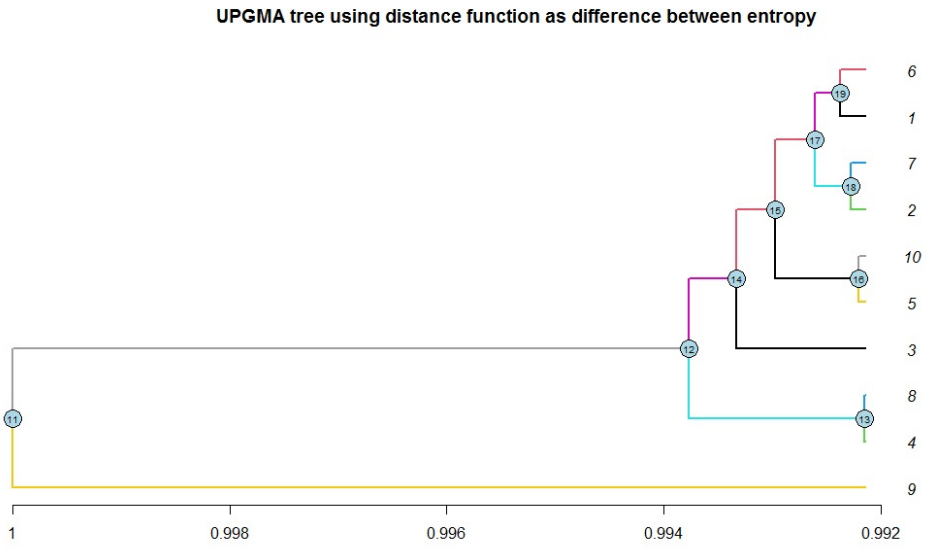


Figure 2.1

Interpretation of the phylogenic tree topology:

1. Initially, sequence 1 and sequence 6 evolved from a common ancestor (node 19). This means that sequences 1 and 6 have the highest

degree of kinship among all the other sequences.

2. Sequences 2 and 7 have evolved from a common ancestor (node 18).

3. Sequences 1,6 and 2,7 evolved from a common ancestor (node 17).

4. Sequences 10 and sequence 5 have evolved from a common ancestor(node 16).
5. Ancestral Node 15 evolved with node 16 to form the common ancestral node with the sequence number 3 (node 14).

6. Sequences, 8 & 4 evolved together to form a common ancestral node of the number 13.

- To reconstruct a phylogenetic tree topology using the UPGMA algorithm with distance function as the frequency of {A, G, C, T}

  For this purpose, we consider the distance function as the Manhattan Distance.

  To reconstruct a phylogenetic tree topology using the UPGMA algorithm with the distance function as the Manhattan distance, we can follow the steps Below:

  1. Calculate the pairwise Manhattan distances between all sequences based on their nucleotide frequencies. Manhattan distance is the sum of the absolute differences between the corresponding nucleotide frequencies in the two sequences.

  2. Create a distance matrix using the pairwise Manhattan distances.

  3. Identify the two sequences with the smallest distance in the distance matrix, and group them into a new cluster.

  4. Calculate the average distance between the new cluster and all other clusters.

  5. Update the distance matrix with the average distance between the new cluster and all other clusters.

  6. Repeat steps 3-5 until all sequences are included in the final cluster.

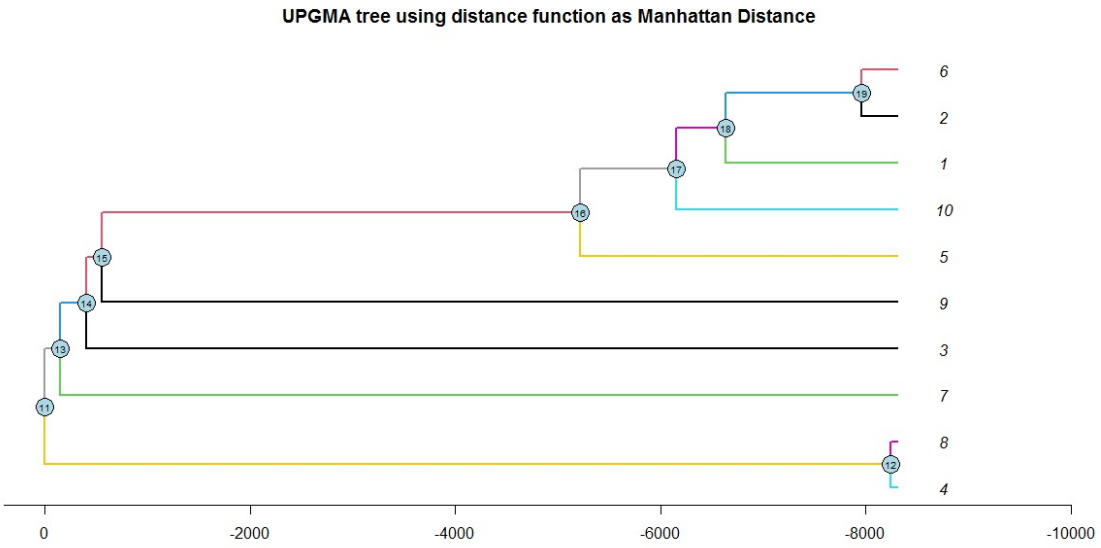


Figure 2.2

Interpretation of the phylogenic tree topology:

1. Initially, sequence 2 and sequence 6 evolved from a common ancestor (node 19). This means that sequences 2 and 6 have the highest degree of kinship among all the other sequences.

2. Sequences 1 and node 19 have evolved from a common ancestor (node 18).

3. Sequences 10 and node 18 evolved from a common ancestor (node 17).

4. Sequences 5 and node 17 have evolved from a common ancestor(node 16).

5. Ancestral Node 16 evolved with sequence 9 to form the common ancestral node 15.

6. Sequences, 3 & node 15 evolved together to form a common ancestral node of the number 14.

7. Sequences 8 & 4 evolved together to form a common ancestral node of the number 12.

- To reconstruct a phylogenetic tree topology using the UPGMA algorithm with a chosen distance function.

    1. Calculate the pairwise cosine distances between all sequences based on their nucleotide frequencies. Cosine distance is a measure of the angle between two vectors in a high-dimensional space. In this case, the nucleotide frequencies for each sequence can be represented as a vector, and the cosine distance between two sequences is the cosine of the angle between their corresponding vectors.

    2. Create a distance matrix using the pairwise cosine distances.

    3. Identify the two sequences with the smallest distance in the distance matrix, and group them into a new cluster.

    4. Calculate the average distance between the new cluster and all other clusters.

    5. Update the distance matrix with the average distance between the new cluster and all other clusters.

    6. Repeat steps 3-5 until all sequences are included in the final cluster.

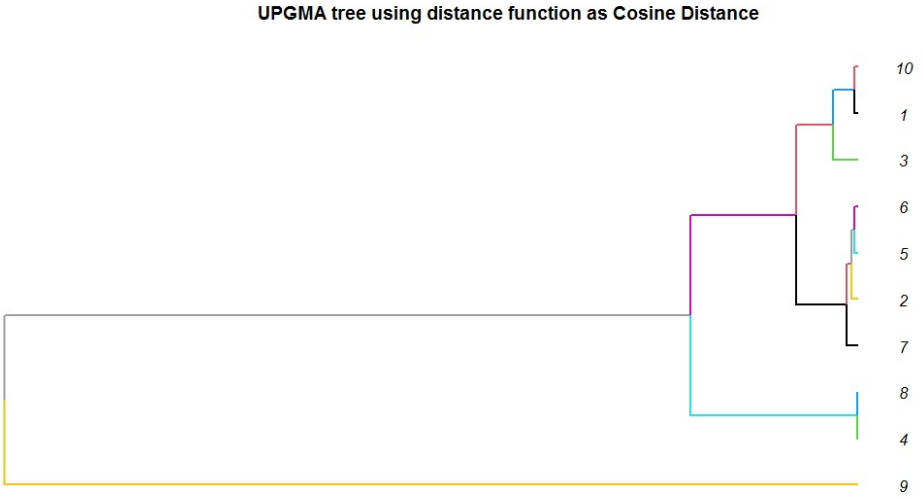


Figure 2.3

Interpretation of the phylogenic tree topology:

1. Initially, sequence 10 and sequence 1 evolved from a common ancestor (node 19). This means that sequences 10 and 1 have the highest degree of kinship among all the other sequences.

2. Sequences 3 and node 19 have evolved from a common ancestor (node 18).

3. Sequences 6 and 5 evolved from a common ancestor.

4. Sequences 8&4 are consistently evolved from a common ancestor for all three distance functions.

5. Even though the cosine distance function is a bit sketchy, it still brings a different perspective to analyze the sequence data.

Question 3

Explain how you use mutual information content to obtain tree topology. Using your suggested algorithm, obtain tree topology for your data.

Solution:

In the case of sequences, mutual information can be used to determine the similarity between different sequences, which can in turn be used to construct a tree topology that reflects the evolutionary relationships between those sequences.

To obtain the tree topology for a set of sequences, one would first need to calculate the pairwise mutual information between all pairs of sequences. This can be done using a variety of methods, such as calculating the mutual information between individual nucleotides or between whole sequences.

Once the mutual information values have been calculated, **the neighbor-joining algorithm** can be used to construct the tree topology. In this case, the distance

between two sequences can be defined as the negative of their mutual information. The algorithm then proceeds as follows:

1. We initialize a distance matrix based on the mutual information values.
2. For this case the distance matrix is $D_{XY} = -\ log(I(X;Y))$
3. Find the pair of sequences with the smallest distance and group them together.
4. Compute the distances between the new group and all the remaining sequences, and update the distance matrix accordingly.
5. Repeat steps 3 and 4 until all sequences have been grouped together into a single cluster.

Use the resulting clustering to construct the tree topology.

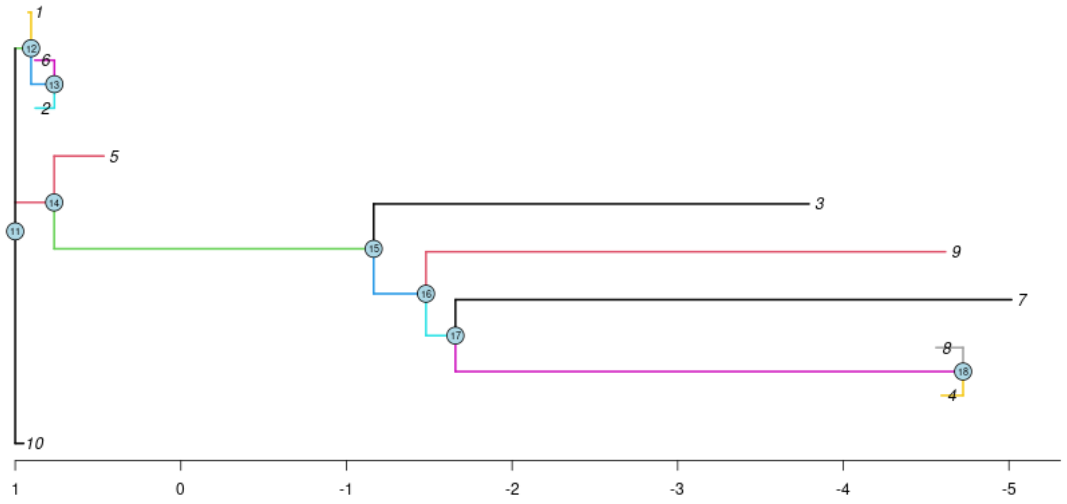


Figure 3.1: **Tree Topology based on Mutual Information Content using NJ Method**

Interpretation of the phylogenic tree topology:

1. The sequences 8 and 4 evolve from a common ancestor node 18. This means that sequences 8 and 4 have the highest degree of kinship among all other sequences.

2. Node 18 and sequence 7 evolve from a common ancestor node 17.

3. Sequence 9 and node 17 evolve from a common ancestor node 16.
4. Sequence 3 and node 16 evolve from a common ancestor node 15.
5. Sequence 5 and node 15 evolve from a common ancestor node 14.
6. Sequences 2 and 6 evolve from a common ancestor node 13.
7. Sequence 1 and node 13 evolve from a common ancestor node 12.
8. Node 12 and node 14 evolve from a common ancestor node 11.
9. Sequence 10 and node 11 evolve from a common ancestor node.

Question 4

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

Solution:

Neighbor-Joining Method:

Neighbor-joining is a popular hierarchical clustering method used to construct a phylogenetic tree from a distance matrix representing the pairwise distances between a set of taxa or sequences. The method was first introduced by Saitou and Nei in 1987.

The neighbor-joining method works by iteratively joining pairs of neighboring taxa or sequences based on their pairwise distances until a complete tree is constructed. At each iteration, the two taxa or sequences with the shortest distance are joined to form a new internal node, and the distance between this new node and each of the remaining taxa or sequences is calculated based on their distances to the two joined taxa. This process is repeated until all taxa or sequences are included in the tree.

The neighbor-joining method has several advantages over other tree reconstruction methods. It is relatively fast and efficient, and it is less sensitive to errors in the distance matrix than other methods such as a maximum likelihood or Bayesian

inference. It also allows for the inclusion of multiple sequences from the same taxon, which is useful for reconstructing phylogenetic trees from molecular data.

However, like all methods, neighbor-joining has its limitations and assumptions. It assumes that the evolution of the sequences or taxa follows a tree-like pattern, and it can be sensitive to the choice of distance metric and the presence of outliers or long-branch attraction.

Overall, neighbor-joining is a powerful and widely used method for constructing phylogenetic trees from distance data, and it has contributed significantly to our understanding of evolutionary relationships among organisms.

Molecular clock property

The molecular clock hypothesis is the assumption that the rate of evolution of a given gene or sequence is constant over time and among different lineages. The molecular clock property is a key assumption in many phylogenetic methods, including the construction of molecular trees using distance-based methods.

Two important conditions that are often used to test whether a molecular clock is operating are the ultrametric condition and the four-point condition.

Ultrametric Condition

The ultrametric condition states that the distances between any three taxa in a molecular tree should satisfy the triangle inequality and that the distance between the root of the tree and any leaf should be the same for all leaves. In other words, the tree should have a "clock-like" structure where all branches are of equal length and the time from the root to any leaf is the same.

Four-Point Condition

The four-point condition is a test for whether a molecular clock is operating by comparing the distances between four taxa in a tree. Specifically, the four-point condition states that if the distances between each of the four taxa are known, then the distance between any two of the four taxa should be equal to the sum of the distances from each of those taxa to their most recent common ancestor in the tree.

If a molecular tree satisfies both the ultrametric condition and the four-point condition, this is evidence in support of the molecular clock hypothesis. However, violations of these conditions can indicate that the molecular clock is not operating or that the tree has been incorrectly inferred.

Overall, the ultrametric condition and the four-point condition are important tools for testing the molecular clock hypothesis and ensuring the accuracy of phylogenetic inference using molecular data.

Three Distance Functions

1. **Hamming distance**
2. **Minkowski distance**
3. **Earth Mover's Distance**

1. Hamming Distance

- Hamming distance is a measure of the difference between two sequences of the same length, where the distance is defined as the number of positions at which the corresponding symbols differ. In other words, the Hamming distance is the minimum number of substitutions required to transform one sequence into the other.

- For example, consider the two DNA sequences "ACGTAGT" and "ACATAGT". The Hamming distance between these sequences is 1 because only the third position differs between the two sequences.

- Hamming distance is commonly used in molecular biology and genetics to compare sequences of DNA, RNA, or protein. It is often used to identify mutations or polymorphisms in genetic sequences or to compare the similarity of sequences between different organisms.

- To calculate the Hamming distance between two sequences, you simply compare the symbols at each position in the two sequences and count the number of differences. The sequences must be of the same length for this calculation to be meaningful.

- It is worth noting that Hamming distance is a simple and intuitive measure of sequence similarity, but it does not take into account the possibility of insertions or deletions (indels) in the sequences, which can complicate sequence comparisons.

Hamming's distance matrix for genome sequence data

0	238	1592	1604	558	234	1580	1602	1518	315
238	0	1583	1587	546	47	1582	1585	1521	330
1592	1583	0	1539	1578	1590	1581	1536	1598	1585
1604	1587	1539	0	1588	1587	1569	10	1554	1594
558	546	1578	1588	0	548	1600	1588	1503	561
234	47	1590	1587	548	0	1581	1585	1522	324
1580	1582	1581	1569	1600	1581	0	1570	1579	1595
1602	1585	1536	10	1588	1585	1570	0	1551	1592
1518	1521	1598	1554	1503	1522	1579	1551	0	1517
315	330	1585	1594	561	324	1595	1592	1517	0

Table 4.1

Hamming's distance neither satisfies ultrametric condition nor it satisfies the four-point condition.

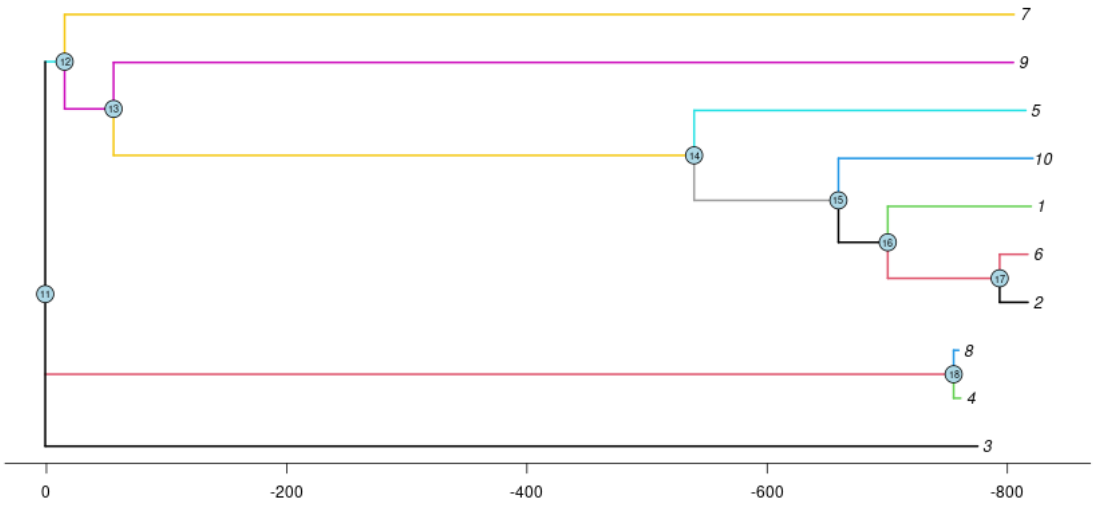


Figure 4.1: Tree Topology using Hamming's distance

Interpretation of the phylogenic tree topology:

1. The sequences 8 and 4 evolve from a common ancestor node 18. This means that sequences 8 and 4 have the highest degree of kinship among all other sequences.

2. The sequences 6 and 2 evolve from a common ancestor node 17.

3. Sequence 1 and node 17 evolve from a common ancestor node 16.

4. Sequence 10 and node 16 evolve from a common ancestor node 15.

5. Sequence 5 and node 15 evolve from a common ancestor node 14.

6. Sequence 9 and node 14 evolve from a common ancestor node 13.

7. Sequence 7 and node 13 evolve from a common ancestor node 12.

8. Node 12 and node 18 evolve from a common ancestor node 11.

9. Node 11 and sequence 3 evolve from a common ancestor node.

2. Minkowski distance

- Minkowski distance is a generalization of the Euclidean distance and Manhattan distance, which are commonly used distance metrics for numerical data. Minkowski distance can also be used to compare sequences of DNA, RNA, or protein, but it requires a numerical representation of the sequences.

- The Minkowski distance here is considered with p=4.

Minkowski distance matrix for genome sequence data

0.00	2,881.80	11,029.02	9,717.95	5,898.23	2,715.73	10,569.25	9,715.73	14,561.70	3,654.09
2,881.80	0.00	11,392.89	10,043.62	6,176.23	1,074.70	10,948.58	10,043.85	15,084.86	3,491.32
11,029.02	11,392.89	0.00	9,995.98	10,790.50	11,112.96	10,937.18	9,989.04	15,666.94	10,601.29
9,717.95	10,043.62	9,995.98	0.00	9,431.94	9,830.70	9,828.88	616.03	14,135.38	9,361.76
5,898.23	6,176.23	10,790.50	9,431.94	0.00	6,033.40	10,468.09	9,450.69	14,407.57	5,712.98
2,715.73	1,074.70	11,112.96	9,830.70	6,033.40	0.00	10,610.95	9,830.90	14,866.59	3,395.06
10,569.25	10,948.58	10,937.18	9,828.88	10,468.09	10,610.95	0.00	9,823.76	15,380.93	10,261.79
9,715.73	10,043.85	9,989.04	616.03	9,450.69	9,830.90	9,823.76	0.00	14,139.86	9,359.76
14,561.70	15,084.86	15,666.94	14,135.38	14,407.57	14,866.59	15,380.93	14,139.86	0.00	14,414.54
3,654.09	3,491.32	10,601.29	9,361.76	5,712.98	3,395.06	10,261.79	9,359.76	14,414.54	0.00

Table 4.2

Minkowski distance neither satisfies ultrametric condition nor it satisfies the four-point condition.

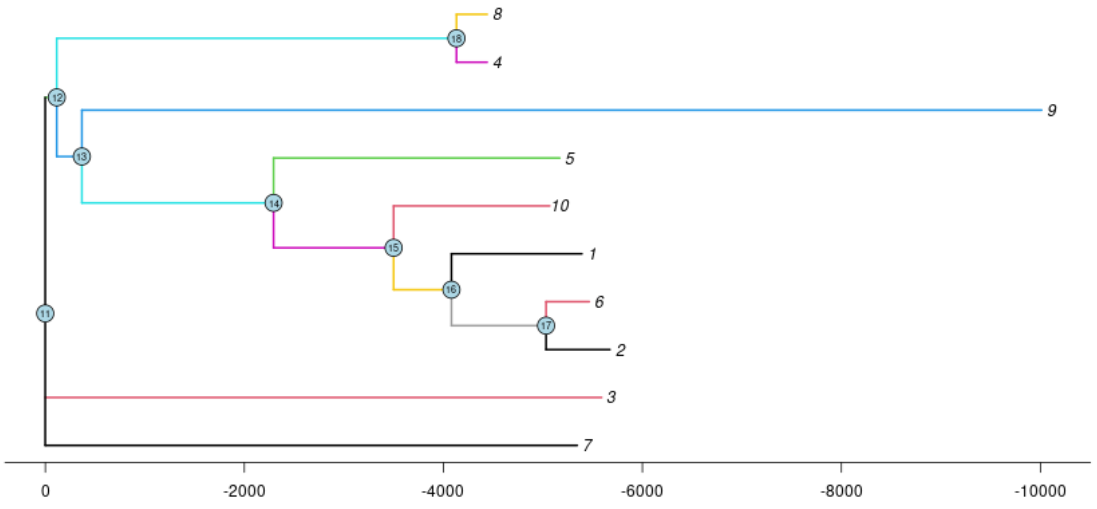


Figure 4.2: Tree Topology using Minkowski distance

Interpretation of the phylogenic tree topology:

1. The sequences 8 and 4 evolve from a common ancestor node 18. This means that sequences 8 and 4 have the highest degree of kinship among all other sequences.

2. The sequences 6 and 2 evolve from a common ancestor node 17.
3. Sequence 1 and node 17 evolve from a common ancestor node 16.
4. Sequence 10 and node 16 evolve from a common ancestor node 15.
5. Sequence 5 and node 15 evolve from a common ancestor node 14.
6. Sequence 9 and node 14 evolve from a common ancestor node 13.
7. Node 18 and node 13 evolve from a common ancestor node 12.
8. Node 12 and sequence 3 evolve from a common ancestor node 11.
9. Node 11 and sequence 7 evolve from a common ancestor node.

Earth Mover's Distance

- The Earth Mover's Distance (EMD), also known as the Wasserstein distance, is a distance metric that can be used to compare sequences of DNA, RNA, or protein. Unlike Hamming distance, which only considers the number of differences between two sequences, EMD takes into account the magnitude of the differences and can handle insertions and deletions (indels) in the sequences.

- The basic idea behind EMD is to treat the sequences as probability distributions and calculate the minimum amount of "work" (i.e. moving mass) required to transform one distribution into the other. In the context of sequence data, we can think of the distributions as representing the frequency of occurrence of each symbol (e.g. A, C, G, T for DNA) at each position in the sequence.

- To calculate the EMD between two sequences, we first convert them into probability distributions by dividing the frequency of each symbol at each position by the total frequency at that position. We can then use a linear programming algorithm to find the minimum amount of work required to transform one distribution into the other.

- The EMD is calculated as the sum of the products of the amount of mass moved and the distance it is moved, summed over all pairs of symbols and

positions. The distance between two symbols is typically defined as the Hamming distance between their binary encodings (i.e. whether or not they are the same), but other distance metrics can also be used. The EMD between two sequences is then the total work required to transform one distribution into the other.

- EMD can be a powerful tool for comparing sequences of DNA, RNA, or protein, especially when there are indels or other structural differences between the sequences. However, it can also be computationally expensive, especially for long sequences or large datasets, and may require specialized software or algorithms to calculate.

Earth Mover's distance matrix for genome sequence data

0.0000	0.0276	0.0100	0.0166	0.0295	0.0257	0.0171	0.0162	0.0646	0.0038
0.0276	0.0000	0.0223	0.0394	0.0048	0.0038	0.0105	0.0390	0.0418	0.0238
0.0100	0.0223	0.0000	0.0266	0.0242	0.0204	0.0119	0.0261	0.0547	0.0109
0.0166	0.0394	0.0266	0.0000	0.0366	0.0356	0.0328	0.0005	0.0813	0.0157
0.0295	0.0048	0.0242	0.0366	0.0000	0.0038	0.0124	0.0361	0.0447	0.0257
0.0257	0.0038	0.0204	0.0356	0.0038	0.0000	0.0086	0.0352	0.0456	0.0219
0.0171	0.0105	0.0119	0.0328	0.0124	0.0086	0.0000	0.0323	0.0485	0.0171
0.0162	0.0390	0.0261	0.0005	0.0361	0.0352	0.0323	0.0000	0.0808	0.0152
0.0646	0.0418	0.0547	0.0813	0.0447	0.0456	0.0485	0.0808	0.0000	0.0656
0.0038	0.0238	0.0109	0.0157	0.0257	0.0219	0.0171	0.0152	0.0656	0.0000

Table 4.3

Earth Mover's distance neither satisfies ultrametric condition nor it satisfies the four-point condition.

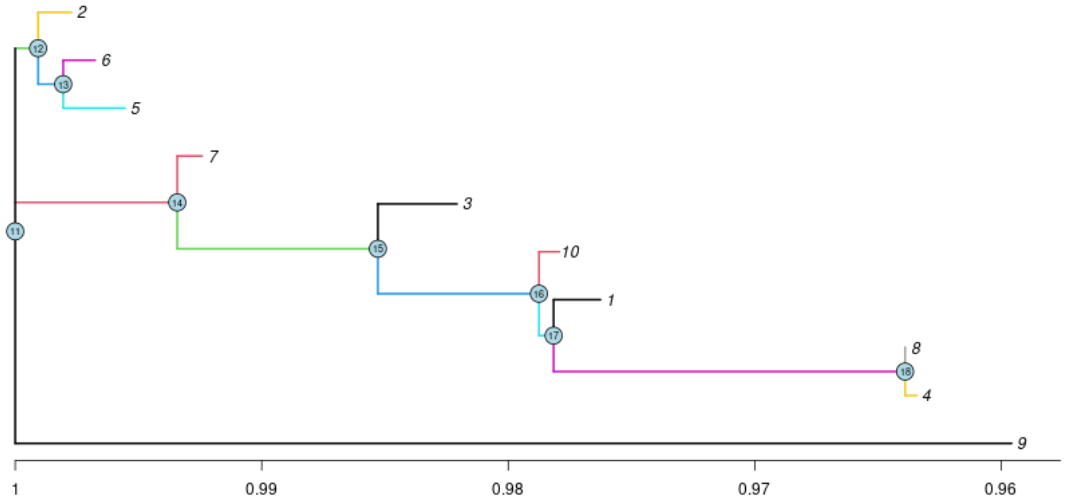


Figure 4.3: Tree Topology using Earth Mover's distance

Interpretation of the phylogenic tree topology:

1. The sequences 8 and 4 evolve from a common ancestor node 18. This means that sequences 8 and 4 have the highest degree of kinship among all other sequences.

2. Sequence 1 and node 18 evolve from a common ancestor node 17.

3. Sequence 10 and node 17 evolve from a common ancestor node 16.

4. Sequence 3 and node 16 evolve from a common ancestor node 15.

5. Sequence 7 and node 15 evolve from a common ancestor node 14.

6. Sequences 6 and 5 evolve from a common ancestor node 13.

7. Sequence 2 and node 13 evolve from a common ancestor node 12.

8. Node 12 and node 14 evolve from a common ancestor node 11.

9. Node 11 and sequence 9 evolve from a common ancestor node.

Question 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 € (A,C,G,T). Obtain the estimates of one step transition probability matrix. Are these Markov chains ergodic? Justify your answer.

Solution:

In Markov chain analysis, ergodicity refers to the property that the chain satisfies certain conditions that allow it to converge to a unique stationary distribution, regardless of its starting state. Checking for ergodicity is an important step in Markov chain analysis, as it ensures that the chain is well-behaved and that the results of any calculations or simulations based on the chain are valid.

One way to check for ergodicity is to examine the transition matrix of the Markov chain. The transition matrix is a square matrix that describes the probabilities of moving from one state to another in a single time step. To be ergodic, the Markov chain must satisfy two conditions:

The chain must be irreducible, meaning that it is possible to move from any state to any other state in a finite number of steps.

The chain must be aperiodic, meaning that the time required to return to a state must not be periodic (i.e. it cannot be a multiple of some fixed time period).

Now, we estimate the TPM for all the sequences using MLE

For sequence 1:

	a	c	g	t
a	0.1815	0.2952	0.2538	0.2696
c	0.2	0.3093	0.2119	0.2789
g	0.1869	0.3152	0.2629	0.235
t	0.1622	0.2409	0.3171	0.2798

Table 6.1

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 1 is ergodic.

For sequence 2:

	a	c	g	t
a	0.1895	0.2972	0.2415	0.2719
c	0.1997	0.3077	0.2069	0.2857
g	0.1881	0.3122	0.2648	0.2349
t	0.1656	0.2389	0.3204	0.275

Table 6.2

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 2 is ergodic.

For sequence 3:

	a	c	g	t
a	0.2	0.2822	0.2489	0.2689
c	0.191	0.331	0.2168	0.2612
g	0.1851	0.3154	0.2688	0.2306
t	0.1677	0.234	0.3149	0.2834

Table 6.3

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 3 is ergodic.

For sequence 4:

	a	c	g	t
a	0.1837	0.2736	0.2489	0.2939
c	0.198	0.3393	0.1936	0.2691
g	0.1822	0.301	0.2733	0.2435
t	0.1747	0.2177	0.3253	0.2823

Table 6.4

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 4 is ergodic.

For sequence 5:

	a	c	g	t
a	0.1673	0.2951	0.2688	0.2688
c	0.1989	0.3299	0.2021	0.2691
g	0.1971	0.3153	0.2602	0.2273
t	0.1794	0.2343	0.3124	0.2739

Table 6.5

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 5 is ergodic.

For sequence 6:

	a	c	g	t
a	0.1991	0.2873	0.2403	0.2733
c	0.1991	0.3085	0.2073	0.285
g	0.1867	0.309	0.2654	0.2388
t	0.1672	0.2413	0.3194	0.2721

Table 6.6

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 6 is ergodic.

For sequence 7:

	a	c	g	t
a	0.192	0.2816	0.2455	0.2809
c	0.1916	0.307	0.2131	0.2883
g	0.1847	0.3225	0.2661	0.2267
t	0.1686	0.2407	0.3129	0.2778

Table 6.7

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 7 is ergodic.

For sequence 8:

	a	c	g	t
a	0.1846	0.2735	0.2496	0.2922
c	0.1992	0.338	0.1919	0.2708
g	0.1827	0.3013	0.2719	0.2441
t	0.1743	0.2199	0.3255	0.2804

Table 6.8

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 8 is ergodic.

For sequence 9:

	g	c	g	t
a	0.1703	0.3544	0.2198	0.2555
c	0.1707	0.3372	0.2267	0.2654
g	0.1777	0.3535	0.2559	0.2129
t	0.1736	0.2868	0.2698	0.2698

Table 6.9

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 9 is ergodic.

For sequence 10:

	a	c	g	t
a	0.1945	0.271	0.245	0.2895
c	0.1898	0.3189	0.2097	0.2816
g	0.1919	0.3083	0.2701	0.2297
t	0.1683	0.2302	0.3217	0.2798

Table 6.10

The above TPM is irreducible and is aperiodic hence the Markov chain for sequence 10 is ergodic.

Q.7 Check Whether in each of your sequence there is a CpG island

CpG islands

CpG islands are regions of DNA that have a high density of CpG dinucleotides, which are two nucleotides (C and G) that occur consecutively in the DNA sequence.

CpG islands are regions in genomes where the dinucleotide CpG occurs more frequently than expected. They have important functional roles:

- They often containing promoter regions of genes. CpG islands near gene promoters are associated with active gene transcription.
- They tend to remain methylation-free. Since CpG methylation is typically associated with gene silencing, CpG islands often contain actively expressed genes.
- They are associated with housekeeping genes. Genes that are expressed in most cell types (housekeeping genes) tend to have CpG islands in their promoters.
- They may contain enhancer regions. Some CpG islands contain enhancer sequences that can regulate gene expression.
- They help define gene boundaries. CpG islands are often found at exon-intron boundaries and can help distinguish genes from non-coding sequences.

In mammals, CpG islands are often located near gene promoters and play an important role in gene regulation. The identification of CpG islands in a genome sequence can provide insights into the location and activity of genes in the genome.

Below, for the sequences we have considered, we consider the **window size of 200, CpG ratio threshold of 0.1, and CpG observed/expected threshold of 0.2 for the CpG island.**

For the sequence: 4AB097812:

CpG island found at :

- 396 - 396 with size 1
- 398 - 477 with size 80
- 479 - 515 with size 37

For the sequence: 3FJ906896:

CpG island found at :

- 932 - 934 with size 3

For the sequence: 3bAB291953:

CpG island found at :

- 1791 - 1834 with size 44

For the sequence: 4AB074915:

CpG island found at:

- 408 - 413 with size 6
- 416 - 454 with size 39
- 475 - 478 with size 4
- 486 - 486 with size 1
- 526 - 599 with size 74

For the sequence: 3bAB291955:

CpG island found at:

- 1791 - 1834 with size 44

For the sequence: 1AF051351:

CpG island found at:

- 233 - 434 with size 202
- 443 - 605 with size 163
- 607 - 615 with size 9
- 1218 - 1245 with size 28
- 1273 - 1273 with size 1
- 1295 - 1298 with size 4
- 1692 - 1693 with size 2

For the rest of the sequences under study, there are no CpG island(s).

Interpretations

- For the sequence 4AB097812, three CpG islands were identified. The first CpG island starts and ends at position 396 and has a size of 1 base pair. The second CpG island starts at position 398 and ends at position 477, and has a size of 80 base pairs. The third CpG island starts at position 479 and ends at position 515, and has a size of 37 base pairs.

- For the sequence 3FJ906896, one CpG island was identified at positions 932-934, with a size of 3 base pairs.

- For the sequence 3bAB291953, one CpG island was identified at positions 1791-1834, with a size of 44 base pairs.

- For the sequence 4AB074915, five CpG islands were identified. The first CpG island starts at position 408 and ends at position 413, with a size of 6 base pairs. The second CpG island starts at position 416 and ends at position 454, with a size of 39 base pairs. The third CpG island starts at position 475 and ends at position 478, with a size of 4 base pairs. The fourth CpG island starts and ends at position 486, with a size of 1 base

pair. The fifth CpG island starts at position 526 and ends at position 599, with a size of 74 base pairs.

- For the sequence 3bAB291955, one CpG island was identified at positions 1791-1834, with a size of 44 base pairs.

- For the sequence 1AF051351, seven CpG islands were identified. The first CpG island starts at position 233 and ends at position 434, with a size of 202 base pairs. The second CpG island starts at position 443 and ends at position 605, with a size of 163 base pairs. The third CpG island starts at position 607 and ends at position 615, with a size of 9 base pairs. The fourth CpG island starts at position 1218 and ends at position 1245, with a size of 28 base pairs. The fifth CpG island starts and ends at position 1273, with a size of 1 base pair. The sixth CpG island starts at position 1295 and ends at position 1298, with a size of 4 base pairs. The seventh CpG island starts at position 1692 and ends at position 1693, with a size of 2 base pairs.

These results indicate that the genome sequences analyzed contain a variety of CpG islands, ranging in size from 1 to 202 base pairs and with different numbers and positions in each sequence. The locations and sizes of these CpG islands may provide insights into the regulation and function of genes in the genome.