CHAPTER IV.

PHYLOGINY

1. Comparaison de séquences

The first question a biologist asks when obtaining a sequence is: "Is there one or more sequences in the database that resemble mine?" The answer to this question requires defining the similarity between sequences. Aligning two sequences is the basis of this comparison.

Sequence comparison is the most used computational task by biologists. It involves determining to what extent two genomic sequences are similar. Thus, if two sequences are highly similar, and one is known to be coding, there can be a hypothesis that the other might also be coding. A biologist who obtains a new sequence first seeks to browse these databases to find similar sequences and transfer the associated knowledge to the new one. By comparing sequences from the genomes of current species, it is also possible to reconstruct phylogenetic trees that represent evolutionary history.

Confused by the diversity of life, one of the earliest biological activities of humans was classification. Biologists have long been involved in the effort to create a hierarchical classification of all species consistent with their evolutionary relationships, also known as the tree of life. This has made tree-building a central activity for biologists, as well as a way to understand the functional similarities between organisms. Evolution requires three basic elements: reproduction, with variation, and selection.

2. Phylogenetic Data

The term "phylogeny" was coined by Ernst Haeckel, derived from the Latin words “fulon” (tribe, race) and “genus” (birth, origin), thus meaning the common ancestor (origin) of a group of genes or other sequences. Phylogeny is based on the principle of comparing specific characteristics for a group of individuals. These characteristics are generally homologous and belong to contemporary organisms.

We can divide the data used to construct phylogenetic trees into two distinct groups:

- Phenotypic data
- Molecular data such as DNA or protein sequences

2.1. Phenotypic Data

This includes observable characteristics (at different states: morphological, biochemical, and physiological) and binary patterns (such as presence/absence of a specific characteristic). In the case of bacteria, for instance, characteristics may include:

- Biochemical and enzymatic traits
- Antigenic traits
- Antibiotic sensitivity
- Phage sensitivity
- Electrophoretic profiles of enzyme systems, etc.

2.2. Molecular Data

In this case, the data are biological sequences such as nucleic acids, including sequences of specific genes, mRNA, RFLPs, microsatellites, SNPs, IGS (rRNA and mitochondria), ITS (rRNA and mitochondria), cytochrome C sequences, elongation factor alpha sequences, or enzymatic or structural protein sequences.

The most commonly used markers for constructing phylogenetic trees are:

- 16S rRNA: Bacteria
- 18S rRNA, actin, EF1, RPB1: Eukaryotes
- 18S rRNA, RBCL: Plants

Traditionally, phylogenetic trees were constructed by comparing phenotypic characteristics, referred to as a phenogram. This approach still plays a dominant role in analyzing data such as fossils. However, today, phylogenetic trees are primarily based on the multiple alignments of nucleotide or amino acid sequences, referred to as a phylogram, and this is known as molecular phylogeny.

3. Constructing a Phylogenetic Tree

3.1. The Distance Matrix

Evolutionary distance is defined as the percentage of nucleotide or amino acid substitutions. It is estimated using several models, such as p-distance, Poisson, Dayhoff, Jones-Taylor-

Thomson (JTT), etc. The distance between sequences is calculated pairwise to produce a distance matrix (Table 1).

Table 1. Estimation of evolutionary divergence between chloroplast protein sequences from 10 plant species.

	1	2	3	4	5	6	7	8	9	10
1. Synechocys										
2. Odontella	0.387									
3. Porphyra	0.305	0.326								
4. Cyanophora	0.304	0.366	0.291							
5. Euglena	0.496	0.493	0.469	0.474						
6. Marchantia	0.402	0.421	0.371	0.366	0.457					
7. Pinus	0.432	0.459	0.414	0.407	0.486	0.193				
8. Nicotiana	0.435	0.462	0.409	0.412	0.491	0.204	0.187			
9. Zea	0.455	0.478	0.429	0.432	0.500	0.241	0.224	0.123		
10. Oryza	0.454	0.478	0.430	0.432	0.500	0.241	0.223	0.122	0.025	

3.2. Phylogenetic Tree Topology

Different methods for constructing phylogenetic trees vary in the evolutionary assumptions they make and the algorithms they use. These methods can be grouped into two categories:

- **Distance-based methods:** Genetic distances (percentage of nucleotide or amino acid substitutions, for example) are measured between all pairs of sequences. These methods are fast and yield good results.
- **Character-based methods:** These focus on phenotypic traits that can take on more than two states. This group includes "parsimony" methods and "maximum likelihood" methods.

For distance-based methods (which are relevant to this course), the first step is to choose a distance criterion between the future leaves of the tree (individuals or OTUs). For example, if these individuals are DNA sequences, the distance between them can be defined as the number of differing nucleotides. To determine this, multiple alignment is performed. Then, methods like UPGMA (unweighted pair group method with arithmetic mean) or NJ (Neighbor-Joining) can be used to deduce the tree's topology. If the individuals were studied based on morphological, physical, and biochemical traits, the distances would result from similarity coefficients.

Distance-based methods use two distinct algorithms to construct dendrograms:

3.2.1. The UPGMA Method

UPGMA uses a sequential clustering algorithm in which relationships are identified in order of their similarity, and the tree is reconstructed step by step following this order. The two closest individuals (OTUs) are identified first, and this group is then treated as a single individual. The process continues, identifying the next closest individual, until only two groups remain. This algorithm calculates an ultrametric tree.

The UPGMA method proceeds through the following steps:

- **Step 1:** In the distance matrix (denoted as dij), find the taxa i and j for which the distance dij is the smallest. First, cluster the two OTUs with the smallest distance.
- **Step 2:** Place the root (theoretical ancestor of the two chosen OTUs) equidistant from the two OTUs i and j, meaning d = dij/2. This distance equals the branch length of the clade that groups individuals i and j.
- **Step 3:** Create a new set including i and j.
- **Step 4:** Calculate the distance between the new group (ij) and each other taxon (k) using the formula: (dki + dkj) / 2.
- **Step 5:** From this new matrix, repeat the process from Step 1.

3.2.2. The NJ method

This method, developed by Saitou and Nei (1987), attempts to correct the UPGMA method to allow for different mutation rates along branches (non-ultrametric tree). The distance matrix accounts for the average divergence of each individual with other taxa. The tree is then constructed by connecting the closest individuals in this new matrix.

The NJ method proceeds as follows:

- **Step 1:** Calculate the net divergence r(i) for each of the N OTUs compared to the others.
- **Step 2:** Calculate the new distance matrix using the following formula:

$$M(i,j)=d(i,j)−(r(i)+r(j)N−2)M(i,j) = d(i,j) - \left(\frac{r(i) + r(j)}{N-2}\right)M(i,j)=d(i,j)−(N−2r(i)+r(j))$$

- **Step 3:** Choose the nearest neighbors, that is, the two OTUs with the smallest M(i,j)M(i,j)M(i,j). These two OTUs form a new node uuu.
- **Step 4:** Calculate the distance of each of the two OTUs to node uuu.

$$S(i,u)=d(i,j)2+r(i)−r(j)2(N−2)S(i,u) = \frac{d(i,j)}{2} + \frac{r(i) - r(j)}{2(N-2)}S(i,u)=2d(i,j)+2(N−2)r(i)−r(j)$$

Therefore:

$$S(j,u)=d(i,j)−S(i,u)S(j,u) = d(i,j) - S(i,u)S(j,u)=d(i,j)−S(i,u)$$

- **Step 5:** Calculate the distances between node uuu and all the other OTUs.
- **Step 6:** Create a new matrix and repeat the process from Step 1.

4. Evaluation of a Phylogenetic Tree

After successfully constructing the phylogenetic tree, the next step is to evaluate the topology of the tree. This process can be performed using two evaluation methods, namely the bootstrap method and the internal branch test.

4.1. The Bootstrap Method

The basic concept of the bootstrap method is to evaluate the tree topology by constructing phylogenetic trees based on a number of repeated pseudo-data sets. Nodes in the tree that show bootstrap values greater than 70% are generally considered consistent.

4.2. The Internal Branch Test

This test is calculated using the bootstrap procedure and is based on the length of internal branches. It is valid only for NJ trees. In this test, the confidence in the length of internal branches being non-zero is assessed.

5. Examples of Phylogenetic Trees

The evolutionary distance values obtained earlier in the distance matrix (chloroplast protein sequences from 10 plant species, Table 2) are projected into space, allowing the construction of the phylogenetic tree using:

- The NJ method and bootstrap test (Figure 1),
- The NJ method and internal branch test (Figure 2),
- The UPGMA method and bootstrap test (Figure 3).

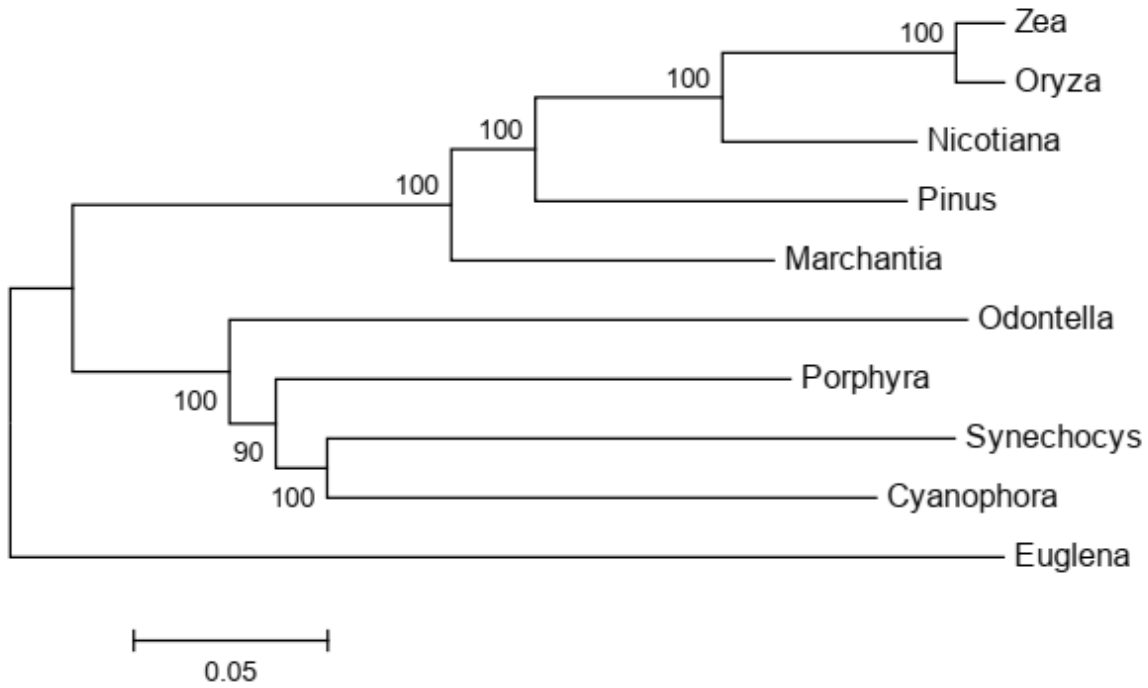

Figure 1. Phylogenetic tree of chloroplast protein sequences from 10 plant species using the NJ method with bootstrap test, created using MEGA6 software.

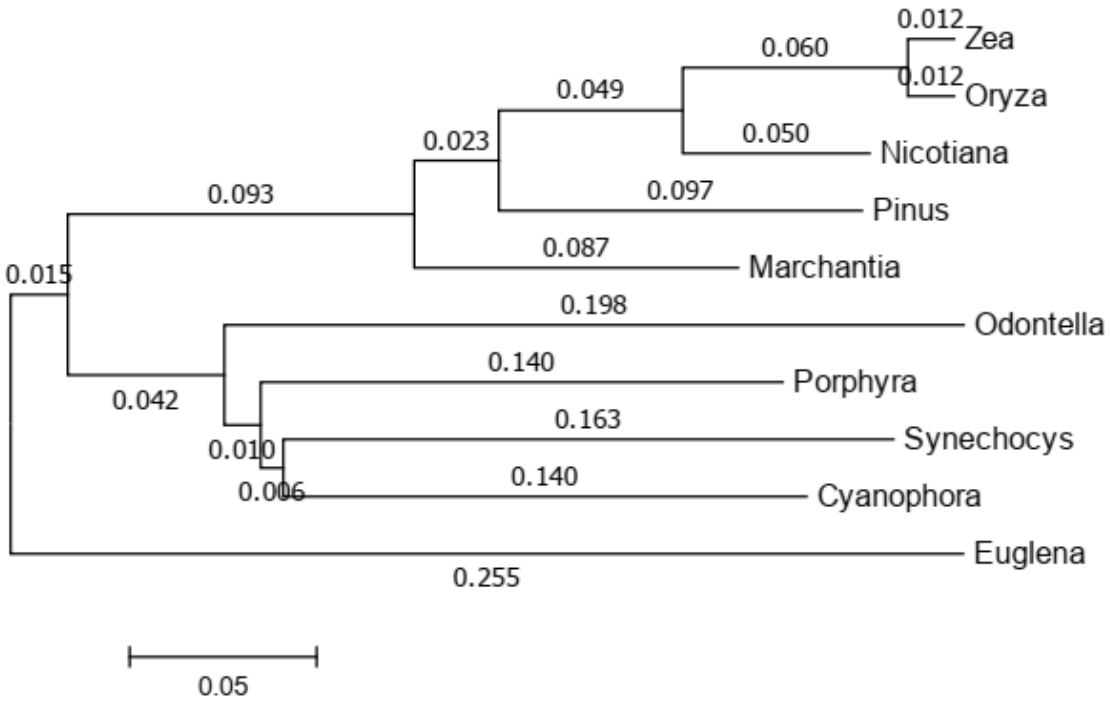

Figure 2. Phylogenetic tree of chloroplast protein sequences from 10 plant species using the NJ method with internal branch test, created using MEGA6 software.

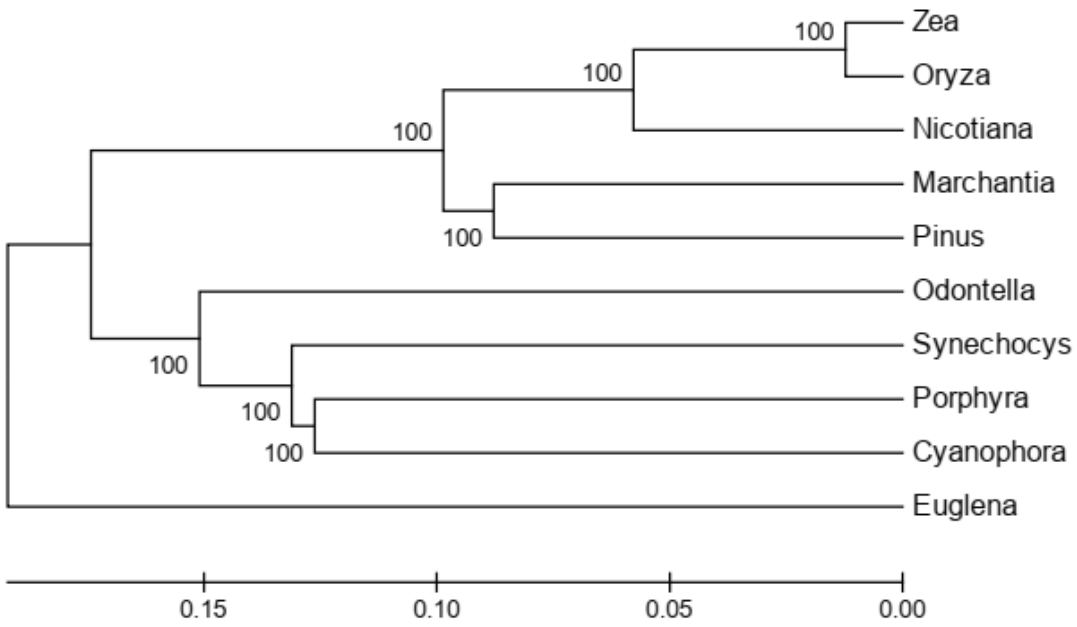

Figure 3. Phylogenetic tree of chloroplast protein sequences from 10 plant species using the UPGMA method with bootstrap test, created using MEGA6 software.

6. Study Exemple

Here, we present an example of a study conducted on the variability and phylogeny of OXA-48 protein structures from class D carbapenemases in *Klebsiella pneumoniae* (Boubendir and Mostakim 2019).

Carbapenem antibiotics are considered the last line of defense in treating infections caused by multidrug-resistant *Enterobacteriaceae* producing Extended-Spectrum Beta-Lactamases (ESBL). The emergence of *Klebsiella pneumoniae* OXA-48, in particular, is constantly expanding and poses a significant public health issue. The objective of this study is to analyze the variability and phylogeny of the amino acid structures of *K. pneumoniae* OXA-48 from various global locations.

Data on the amino acid structures of *K. pneumoniae* OXA-48 were collected during May 2019 from the Protein Data Bank (PDB). Protein sequence alignment was performed using the

Clustal Omega program, available on the UniProt database. Phylogenetic analysis and dendrograms were conducted using the MEGA software version 6.0.

Out of 58 retrieved structures, 8 representative OXA-48 variants were selected for this study (Table 2). The alignment demonstrated that conserved motifs are generally well-preserved, except for two mutations, S70G and S70A, observed in the two chains 5HAQ and 5HAP from the United States (Figure 4). However, the OXA-181 and OXA-245 variants exhibited mutations distant from the active sites. Compared to OXA-48, the OXA-181 variant showed four substitutions at Thr104Ala, Asn110Asp, Glu168Gln, and Ser171Ala, while OXA-245 had a single amino acid substitution at Glu125Tyr.

Phylogenetic analysis revealed three distinct clusters (Figure 5). The first cluster consisted of four OXA-48 structures (Canada, Norway, United States, and Italy) and one OXA-245 structure (Norway). The second cluster included two OXA-48 structures from the United States, while the third cluster was formed by a single OXA-181 structure from Norway.

The results of this study confirm a similar evolutionary trend of OXA-48 structures worldwide. Current data on *K. pneumoniae* OXA-48 structures are limited to restricted geographic areas and need to be expanded to provide a more accurate picture of molecular changes and the evolution of antibiotic resistance.

Table 2: OXA-48 variants of *Klebsiella pneumoniae* collected from PDB during May 2019: variant name, PDB ID, country of origin, and references.

Nom du variant	Ipdb	Pays d'origine	Références
OXA-48	3HBR	Italie	8
OXA-48	4WMC	USA	23
OXA-48	5HAQ*	USA	29
OXA-48	5HAP**	USA	29
OXA-48	5FAQ	Canada	19
OXA-48	5QA4	Norvège	3
OXA-181	5OE0	Norvège	2
OXA-245	5OE2	Norvège	2

*: mutant - S70G, **: mutant - S70A

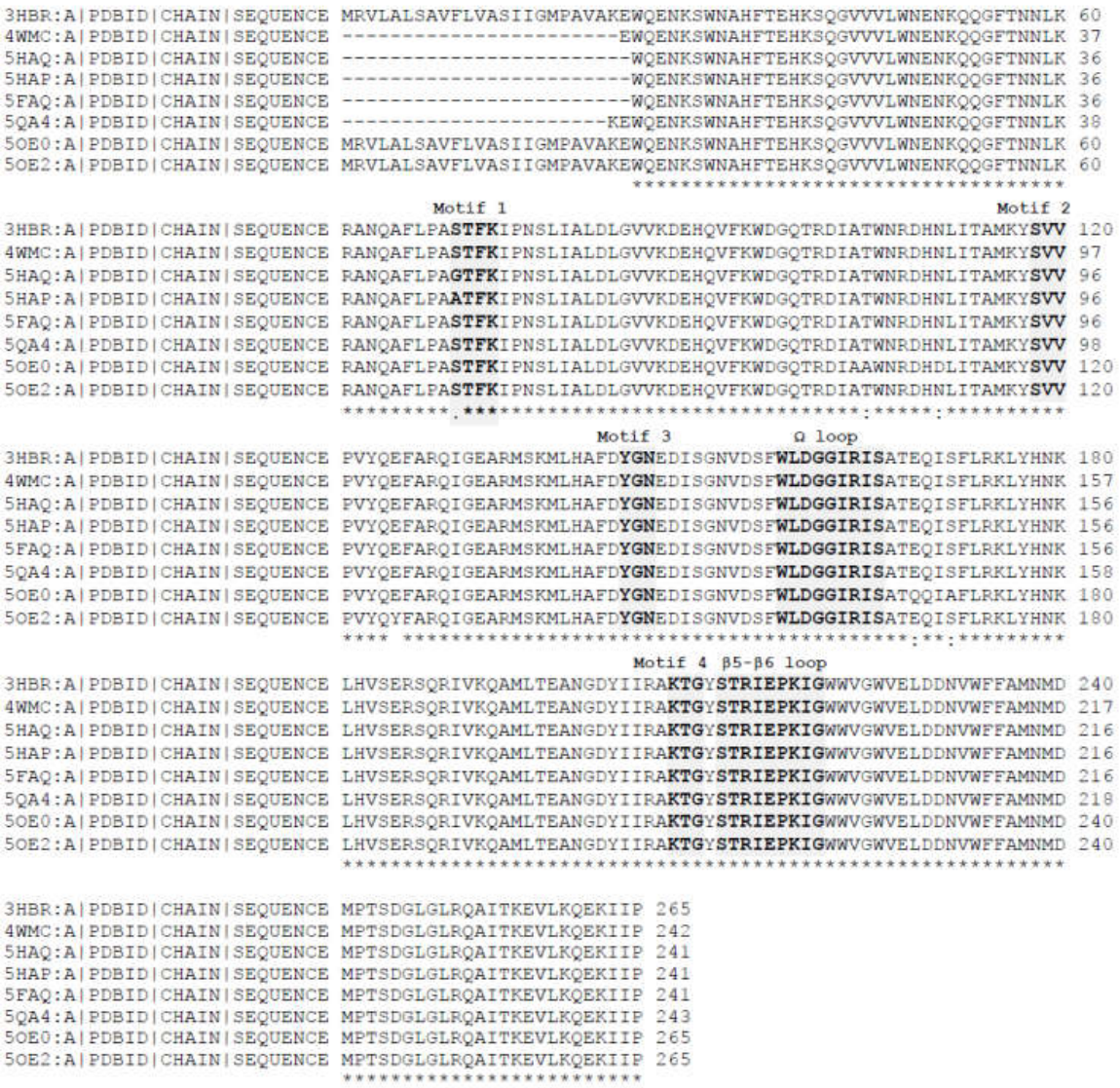

Figure 4: Alignment of 8 representative amino acid structures of *Klebsiella pneumoniae* OXA-48 from different regions of the world: OXA-48 (3HBR/Italy, 4WMC, 5HAQ, and 5HAP/USA, 5FAQ/Canada, 5QA4/Norway); OXA-181 (5OE0) and OXA-245 (5OE2)/Norway. Stars indicate identical residues among all amino acid sequences. Amino acids in motifs that are well-conserved (even with possible variation) are indicated in gray. The numbering follows the DBL system.

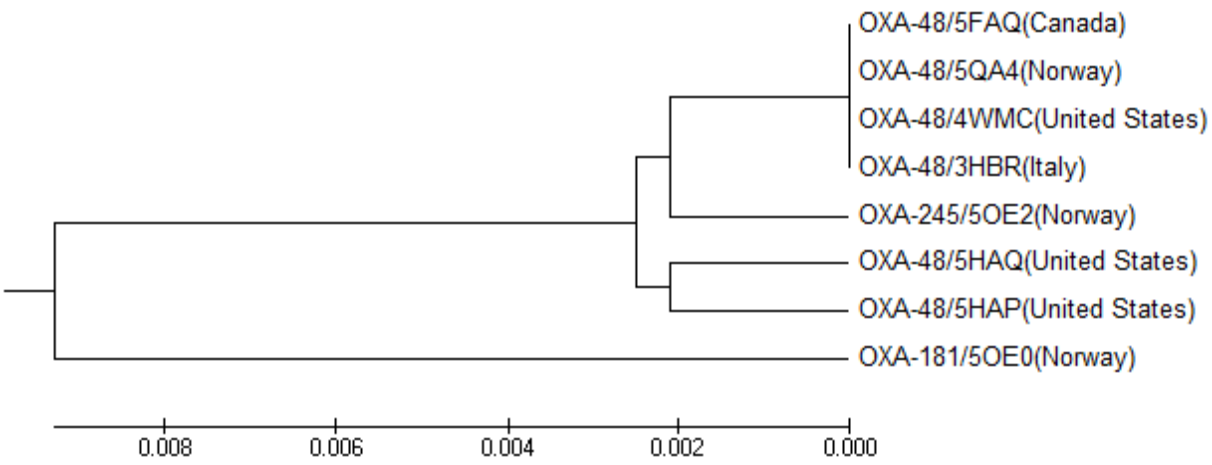

Figure 5: Dendrogram obtained from 8 representative variants of *Klebsiella pneumoniae* OXA-48 amino acid structures from different global locations. The evolutionary history is inferred using the UPGMA method. Evolutionary distances are calculated using the Poisson correction method. Evolutionary analysis was conducted using MEGA software version 6.0.