UNDERSTANDING EVOLUTIONARY RELATIONSHIPS AND ANALYSIS METHODS THROUGH MEGA SOFTWARE

Gülşah KEKLİK
gulsahkeklk@gmail.com, https://orcid.org/0000-0002-1775-2773
Çukurova University, Faculty of Agriculture, Department of Animal Science, Division of Biometry and Genetics, Adana, Türkiye

ABSTRACT: *In recent years, genome sequencing has been a highly effective tool to investigate a wide range of biological systems. Numerous steps for the interpretation of raw sequence data require comparative analysis of molecular sequences for each time period to reveal functional genome differences. Molecular Evolutionary Genetics Analysis (MEGA), a user-friendly software developed to analyse DNA and protein sequence data, provides a set of tools to perform these analyses. MEGA enables a wide range of investigations such as combining sequence alignments, constructing evolutionary trees, estimating genetic distances and variations, and uncovering ancestral sequences. There are some bioinformatics methods to estimate the past knowledge of organisms from their current knowledge. The phylogenetic tree is similar to the evolutionary tree in terms of relationships between biological entities, and the relationships between various species, organisms, genes, etc. in the phylogenetic tree structure are shown by dendograms. Analyses can be performed on nucleic acid, protein, DNA and RNA sequences. There are basically two trees: rooted and unrooted phylogenetic trees. The root is represented by the rooted phylogenetic tree type and is not represented in the unrooted tree, but it can be considered as a root tree because some characters are mapped to the data used. The common ancestor is represented by the root of the phylogenetic tree and the branches of the nodes in the tree represent recent biological information. MEGA software was used to estimate the relationships shown in this figure. In this study, phylogenetic tree construction and evolutionary roots detection and analysis methods were investigated through MEGA 11 software programme on a working file containing data of 10 strains (KC710304.1, KC710305.1, KC710306.1, KC710307.1, KC710308.1, KC710309.1, KC710310.1, KC710311.1, KC710312.1 and KC710313.1) of Thermoplasma acidophilum bacteria obtained from NCBI data bank.*

KEYWORDS: *MEGA Software, Phylogenetic Tree, Nucleic Acid Sequences, Protein Sequences*

Date of Submission: 20.11.2023 Date of acceptance: 25.12.2023
https://doi.org/10.5281/zenodo.10446714

1. INTRODUCTION

Understanding the phylogenetic relationships among species is crucial for many studies in biology. An accurate phylogenetic tree is the basis for understanding major transitions in evolution. A well-structured phylogenetic tree is crucial for determining the origin and molecular adaptation of new genes and for reconstructing demographic changes in recently diverged species. Despite the presence of more data than ever before and the use of powerful analysis methods, many challenges exist for reliable tree construction. In order to decrease or eliminate these difficulties, it is an important step to be aware of the existence of various software programmes. being able to use these programmes is useful for understanding evolutionary relationships and correctly interpreting the output of analyses. MEGA software is just one of the programmes used to understand evolutionary history and relationships (Kapli et al., 2020).

1.1. MEGA Software

MEGA is a software programme that enables research such as combining sequence alignments, constructing evolutionary tree structures, estimating genetic distances and variation, and uncovering ancestral sequences. MEGA software is maturing day by day to include a large collection of methods and tools of computational molecular evolution. With the latest update, the software has become supportive and interactive for large datasets, which will further increase the quality of results and the speed of biological discoveries over time (Hall, 2013; Tamura et al., 2021).

The following figure shows the tree explorer obtained from the MEGA 11 software programme (Figure 1).

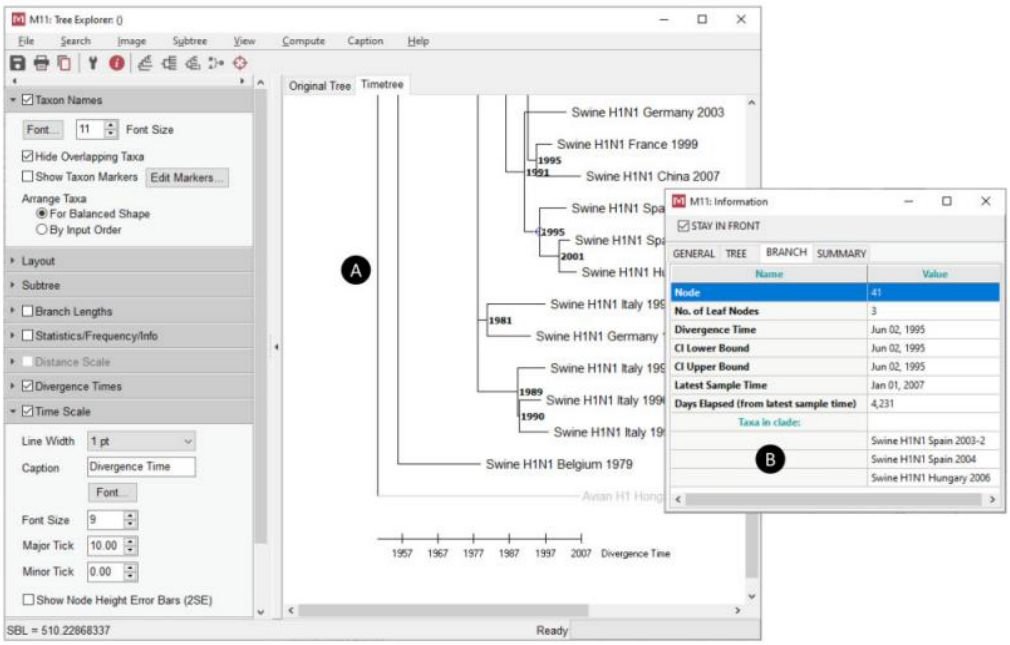


Figure 1. *MEGA's tree explorer screenshot*

MEGA's tree explorer (A) is a feature-rich, versatile phylogeny viewer that provides many interactive exploration and customisation possibilities. In addition to an updated toolbar, there are now options to automatically collapse nodes containing clusters of taxa belonging to the same group, based on user-specified cluster size or branch length difference. The tree information box (B) for time trees has been updated to show branch- and node-specific information (Tamura et al., 2021).

Over a long period of time, the scope and usefulness of MEGA has grown with the addition of new methods, tools and interfaces, which has resulted in a modern and integrated software for comparative sequence analysis (Caspermeyer, 2018).

1.2. Phylogenetic Tree Structure and Terminology

Phylogenies, which aim to reveal the evolutionary ancestor-descendant relationship among all groups of organisms, expressing that these organisms are related to each other by evolving from a common ancestor, that is, they have a single ancestor, are expressed as phylogenetic tree structure, evolutionary tree or tree of life (Sarıçam and Müştak, 2015; Singh, 2015; Woese et al., 1985; Xiong, 2006).

Taxon

Each of the units (family, class, genus, order, order, branch) in a hierarchical order from the genus to the species is called a taxon (Sarıçam and Müştak, 2015).

Ancestry

The branch that expresses the ancestor-descendant relationship and extends to the monophyletic group is called lineage (Sarıçam and Müştak, 2015).

Node Point

Each taxonomic unit formed by the intersection of two neighboring units is called a node (Sarıçam and Müştak, 2015).

Outgroup

A unit that is sharply separated from all other units and does not have a close relationship with others is called an outgroup (Kapli et al., 2020; Sarıçam and Müştak, 2015).

Figure 2 shows an example of a drawing for the identification of outgroups.

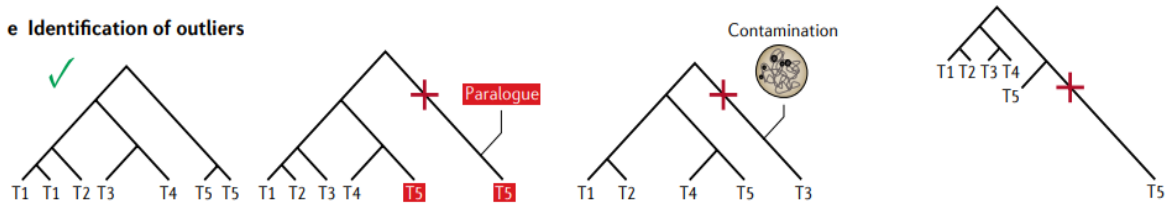


Figure 2. *Identification of outliers*

e. Identification of outgroups: Multiple Sequence Alignment (MSA) can generate the phylogenetic tree for putative orthologues, which can be used to identify other problematic sequences indicated by long branches (Kapli et al., 2020).

Dendogram

The simplest representation, also called a cluster tree, which is formed by combining the Greek words dendro (tree) and gramma (drawing), is expressed as a dendogram.

1.2.1. Unrooted and Rooted Trees

Phylogenetic trees can be constructed in rooted and unrooted forms. Unrooted trees do not have an evolutionary direction because their common ancestors are not known, although the relationship of each unit with the other unit is observed. Rooted trees, on the other hand, have an evolutionary direction because they originate from a common ancestor and allow more information to be obtained compared to unrooted phylogenetic trees (Xiong, 2006).

Below, a visualisation of the unrooted and rooted phylogenetic tree structure is given (Figure 3).

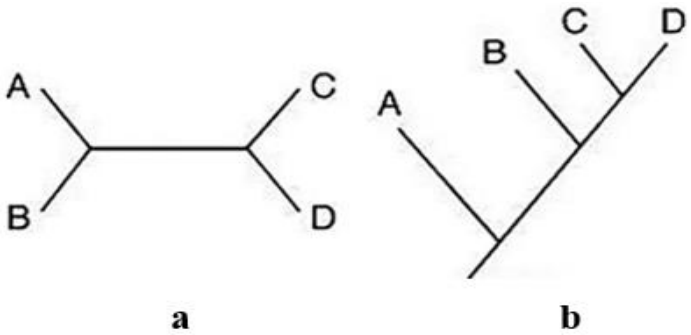


Figure 3. *Unrooted (a) and rooted phylogenetic tree (b)*

2. MATERIALS AND METHODS

In this study, evolutionary analyses of the microorganism *T. acidophilum*, an archaeon first isolated from a self-heating coal waste pile at 2 pH and 59 °C (Ruepp et al., 2000), were carried out using MEGA 11 software programme. Data obtained from 10 strains (KC710304.1, KC710305.1, KC710306.1,…,KC710312.1 and KC710313.1) of the bacterium were obtained from the NCBI database. Evolutionary analysis and evolutionary history were generated using the Maximum Likelihood (ML) method and the General Time Reversible (GTR) model (Nei and Kumar, 2000; Tamura et al., 2021). A bootstrap consensus tree extracted from 50 replicates was constructed to

represent the evolutionary history of the taxa analysed (Felsenstein, 1985). Branches corresponding to sections replicated in fewer than 50 bootstrap replicates were collapsed. The initial tree(s) for heuristic search were obtained by applying the Neighbour-Joining (NJ) method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. The analysis included 10 nucleic acid sequences with codon positions 1, 2, 3 and non-coding, and the final dataset contained a total of 602 positions.

2.1. Maximum Likelihood Method

ML, one of the most widely used statistical estimation methods, is known as Fisher's invention. It is a powerful and widely used tool for obtaining the best estimate of a parameter from observed data, evaluating the quality of a model and making decisions based on data (Richards, 1961). To understand Maximum Likelihood, it is necessary to know what likelihood is. Likelihood is the probability of the observation of the data considering a set of parameter values. Maximum Likelihood is a method of obtaining parameter values μ and σ that maximise the likelihood of the observed data. The likelihood function is as given below:

$$L(\mu, \sigma \mid data) = (1 / (\sigma * sqrt(2\pi)) \wedge n * exp(-1/2 * \Sigma((x\_i - \mu) / \sigma) \wedge 2) \quad (1)$$

n is the number of data points, x_i is the $i-$th data point and Σ is the sum of all data points (Anonymous 1).

2.2. General Time Reversible Model

GTR model, first described by Simon Tavaré in 1986, is the most general possible neutral, independent, finite domain, time reversible model. Many patterns of beneficial substitution are time-reversible. In the analysis of real biological data, there is often access to sequences of present-day species rather than access to sequences of ancestral species. When a model is time-reversible, it does not matter which species is ancestral. Instead, the phylogenetic tree can be rooted using any species. Each of the species will eventually descend from each other with the same probability (Tavaré, 1986).

A model can be said to be time reversible if and only if it has the following properties:

$$\pi_i Q_{ij} = \pi_j Q_{ji} \quad (2)$$

2.3. Neighbour-Joining Method

In bioinformatics, N-J is a bottom-up agglomerative clustering method for the construction of phylogenetic trees created by Naruya Saitou and Masatoshi Nei in 1987 (Saitou and Nei, 1987). N-J takes as input a distance matrix specifying the distance between pairs of taxa. The algorithm starts with a completely unsolved tree whose topology corresponds to a star network and repeats the following steps until all branch lengths are known:

1. Based on the given distance matrix, calculate a matrix Q,

2. For the smallest $Q(i, j)$, find a pair of different taxa i and j ($i \neq j$). Create a new node connecting taxa i and j and connect the new node to the centroid node.

3. Calculate the distance of each taxon in the pair to that new node.

4. Calculate the distance of each of the taxa out of this couple to the new node.

5. Replace the merged neighbour pair with the new node and reinitialise that algorithm by using the distances calculated in an earlier step.

With n taxa, the nxn matrix Q is calculated as follows:

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k) \quad (3)$$

2.4. Maximum Composite Likelihood Approach

The composite likelihood is an inference function obtained by multiplying a collection of component likelihoods. Each component is a conditional or marginal density, and the result of the derivative of the composite log-likelihood is an unbiased prediction equation. Assume an m-dimensional vector random variable Y, having the probability density function $f(y;\ Q)$ for several uncertain p-dimensional parameter vector $Q \in \Theta$. Show $\{A_1, \ldots, A_K\}$ a set of the marginal or contingent events which are related likelihoods $\mathcal{L}_k(Q;y) \infty f(y \in A_k;\ Q)$.

$$\mathcal{L}_C(Q;y) = \prod_{k=1}^{K} \mathcal{L}_k(Q;y)^{w_k} \quad (4)$$

Where w_k are the non-negative weights to be selected and this can be disregarded if all weights are equal (Varin et al., 2011).

3. RESULTS

The first analyses of the study started by reading the nucleic acid sequences obtained from 10 strains of *T. acidophilum* bacteria in MEGA 11 software programme and making them ready for alignment.

In the figure below, the alignments of the nucleic acid sequences are shown in the screenshot taken from the programme (Figure 4).

Figure 4. *Nucleic acid sequence alignment*

In Figure 5, information on nucleotide compositions is given in the screenshot taken from the MEGA 11 software programme. It was observed that the highest and lowest base numbers belonged to microorganisms with GenBank accession numbers KC710312.1 and KC710305.1, correspondingly.

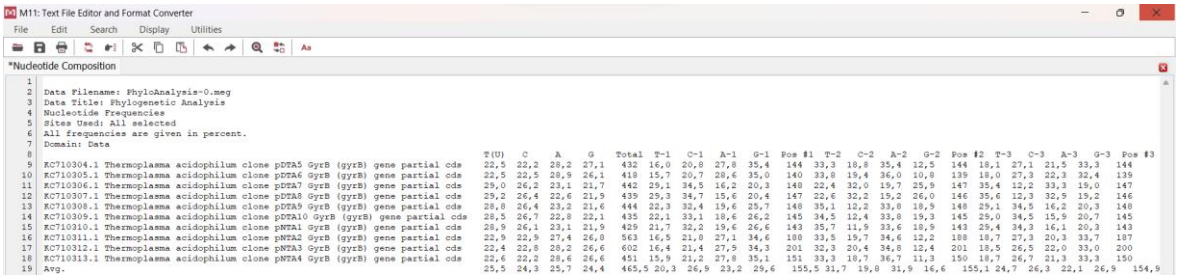


Figure 5. *Nucleotide compositions*

Figure 6 shows the "original tree" image of the nucleic acid sequences. As a result of the taxonomic classification, which is basically divided into two main groups, it is clearly seen that microorganisms with GenBank accession numbers KC710308.1 and KC710309.1, KC710305.1 and KC710313.1, KC710306.1 and KC710307.1, KC710310.1 and KC710311.1 are close to each other in binary groups. It is observed that microorganisms with GenBank accession numbers KC710304.1 and KC710312.1 are outgroups.

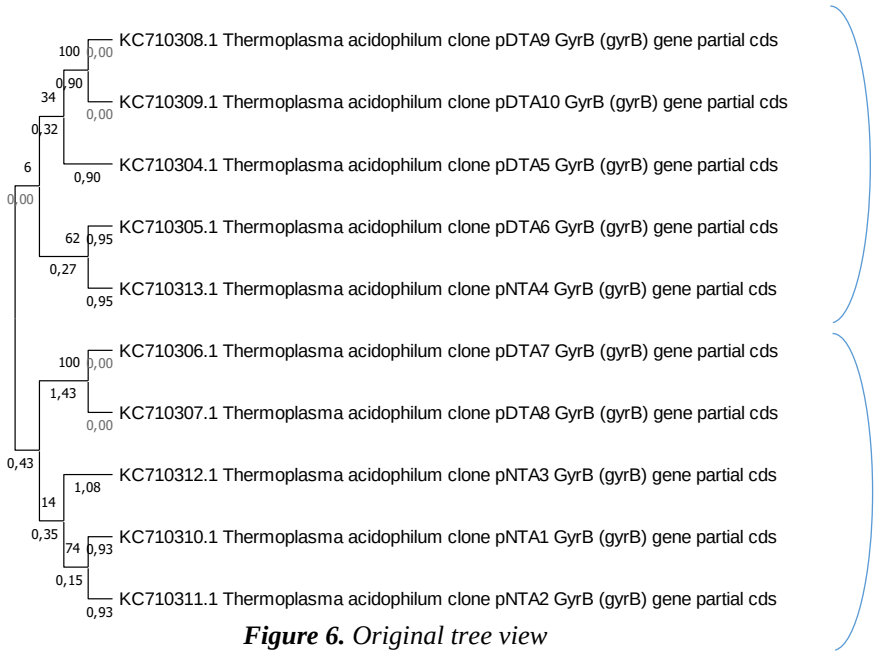


Figure 6. *Original tree view*

Figure 7 shows the "Bootstrap consensus tree" image. The conclusions made for Figure 6 are also valid for Figure 7.

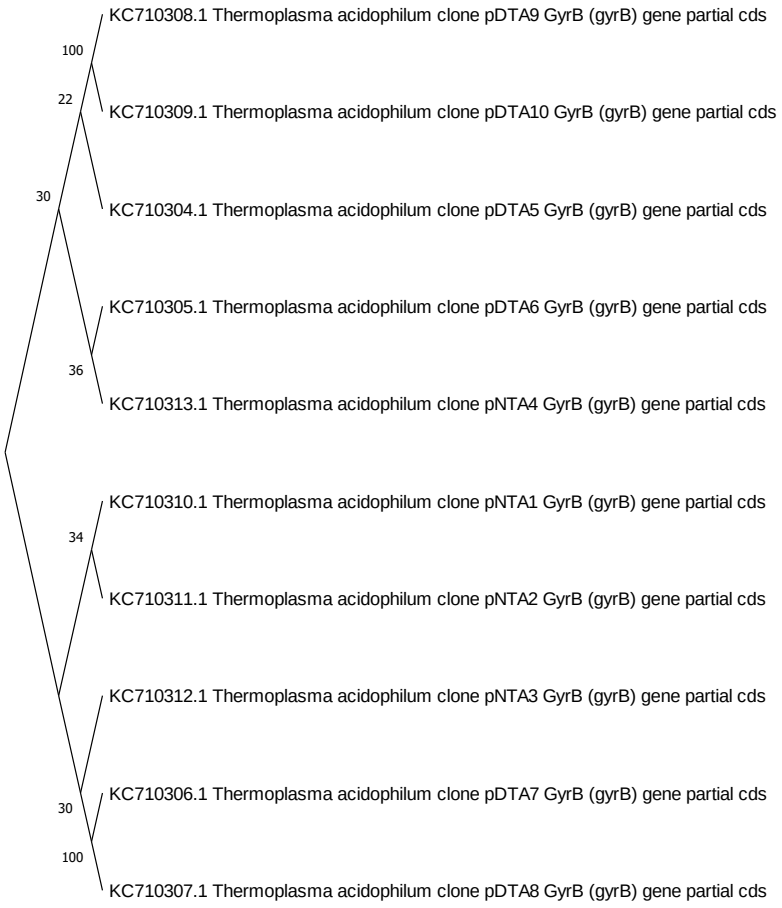
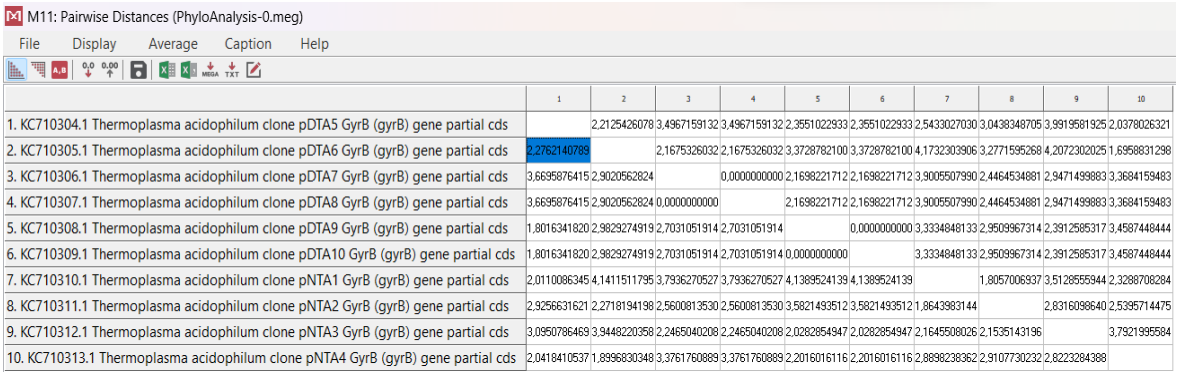


Figure 7. *Bootstrap consensus tree image*

Figure 8 shows the result screen with the values of pairwise distances. Looking at the result screen, it was observed that the distance between microorganisms with GenBank accession numbers KC710305.1 and KC710312.1 was approximately 4.21 and this value was the highest value observed

on the result screen. The lowest distance was between KC710306.1 and KC710307.1 and microorganisms with GenBank accession numbers KC710308.1 and KC710309.1 and this distance was observed as 0.00.


M11: Pairwise Distances (PhyloAnalysis-0.meg)

File Display Average Caption Help

	1	2	3	4	5	6	7	8	9	10
1. KC710304.1 Thermoplasma acidophilum clone pDTA5 GyrB (gyrB) gene partial cds		2,2125426078	3,4967159132	3,4967159132	2,3551022933	2,3551022933	2,5433027030	3,0438348705	3,9919581925	2,0378026321
2. KC710305.1 Thermoplasma acidophilum clone pDTA6 GyrB (gyrB) gene partial cds	2,2762140789		2,1675326032	2,1675326032	3,3728782100	3,3728782100	4,1732303906	3,2771595268	4,2072302025	1,6958831298
3. KC710306.1 Thermoplasma acidophilum clone pDTA7 GyrB (gyrB) gene partial cds	3,6695876415	2,9020562824		0,0000000000	2,1698221712	2,1698221712	3,9005507990	2,4464534881	2,9471499883	3,3684159483
4. KC710307.1 Thermoplasma acidophilum clone pDTA8 GyrB (gyrB) gene partial cds	3,6695876415	2,9020562824	0,0000000000		2,1698221712	2,1698221712	3,9005507990	2,4464534881	2,9471499883	3,3684159483
5. KC710308.1 Thermoplasma acidophilum clone pDTA9 GyrB (gyrB) gene partial cds	1,8016341820	2,9829274919	2,7031051914	2,7031051914		0,0000000000	3,3334848133	2,9509967314	2,3912585317	3,4587448444
6. KC710309.1 Thermoplasma acidophilum clone pDTA10 GyrB (gyrB) gene partial cds	1,8016341820	2,9829274919	2,7031051914	2,7031051914	0,0000000000		3,3334848133	2,9509967314	2,3912585317	3,4587448444
7. KC710310.1 Thermoplasma acidophilum clone pNTA1 GyrB (gyrB) gene partial cds	2,0110086345	4,1411511795	3,7936270527	3,7936270527	4,1389524139	4,1389524139		1,8057006937	3,5128555944	2,3288708284
8. KC710311.1 Thermoplasma acidophilum clone pNTA2 GyrB (gyrB) gene partial cds	2,9256631621	2,2718194198	2,5600813530	2,5600813530	3,5821493512	3,5821493512	1,8643983144		2,8316098640	2,5395714475
9. KC710312.1 Thermoplasma acidophilum clone pNTA3 GyrB (gyrB) gene partial cds	3,0950786469	3,9448220358	2,2465040208	2,2465040208	2,0282854947	2,0282854947	2,1645508026	2,1535143196		3,7921995584
10. KC710313.1 Thermoplasma acidophilum clone pNTA4 GyrB (gyrB) gene partial cds	2,0418410537	1,8996830348	3,3761760889	3,3761760889	2,2016016116	2,2016016116	2,8898239362	2,9107730232	2,8223284388	

Figure 8. *Pairwise distances*

4. CONCLUSION AND DISCUSSION

Revealing evolutionary relationships is only possible with the correct construction of phylogenetic tree structures. In this direction, it is very important to classify the obtained tree structures in a taxonomic way, to form and interpret the existing internal, external and subgroups. There are other programmes other than MEGA software for phylogenetic tree structure research. Numerous steps are required for the interpretation of raw sequence data and comparative analysis of molecular sequences to reveal genome differences. MEGA, a user-friendly software developed to analyse nucleic acid and protein sequence data, stands out in that it provides a set of tools to perform these analyses.

Within the scope of this study, evolutionary relationships and evolutionary history were investigated by using MEGA 11 software programme. As a result of these investigations; it was observed that microorganisms belonging to accession numbers KC710304.1, KC710305.1, KC710308.1, KC710309.1 and KC710313.1 and accession numbers KC710306.1, KC710307.1, KC710310.1, KC710311.1 and KC710312.1 are related taxa within themselves. There are mainly two main groups and many ingroups.

Many recent studies have shown that MEGA is a commonly used software for identifying evolutionary relationships and ancestral history. “Assessing the Phylogenetic Relationships of *Baliospermum solanifolium* [*B. montanum* (Willd.) Müll. Arg.] in South India through DNA Barcoding (Karthik et al., 2023)”, “Phylogenetics Analysis of Blackwater Fishes in Peat Swamp Forest Using Cytochrome B Gene (Sout et al., 2023)”, “Phylogenetic Analysis of the Partial Sequences of the Env and Tax BLV Genes Reveals the Presence of Genotypes 1 and 3 in Dairy Herds of Antioquia, Colombia (Úsuga-Monroy et al., 2023)” and “Molecular Characterization and Phylogenetic Analysis of Newcastle Disease Viruses Isolated in Southern Angola, 2016-2018 (Henriques et al., 2013)” are several examples of these studies.
By looking at the original or Bootstrap consensus tree images, it is easier to quickly identify closely related and distantly related groups. It is important to confirm the results with various distance calculations in order to prevent incomplete or incorrect evaluations in bioinformatics studies. As with many software programmes, new features and analysis methods are added to MEGA every day. In order to analyse the researches with more effective methods and to create more qualified analysis outputs, following the current versions of the software and installing them on computers will be a useful way to prevent problems that will occur in the functioning of *in silico* studies.

The aim of this study is to construct and interpret evolutionary trees and genetic distances based on data obtained from strains of the bacterium *T. acidophilum*. Undoubtedly, there has been a recent awareness of the use of bioinformatics tools. The study is also very important for researchers to use bioinformatics tools to understand evolutionary relationships and to acquire good analytical skills by understanding the statistical methods used in this field. The importance of *in silico* studies in bioinformatics has increased due to some important reasons such as experimental methods taking a

long time and being quite costly. In the coming years, there will be a need for a large number of studies involving phylogeny research and many researchers who will build their careers in bioinformatics to fill this gap.

REFERENCES

[1]. Anonymous 1: https://medium.com/techdevathe/maximum-likelihood-36256cebc8a0 [Last access date: 14.12.2023].

[2]. Caspermeyer, J. (2018), MEGA Software Celebrates Silver Anniversary, Molecular Biology and Evolution, 35(6), 1558–1560.

[3]. Felsenstein, J. (1985), Confidence Limits on Phylogenies: An Approach Using The Bootstrap, Evolution, 39(4), 783-791.

[4]. Hall, B. G. (2013), Building Phylogenetic Trees From Molecular Data with MEGA, Molecular Biology and Evolution, 30(5), 1229-1235.

[5]. Henriques, A. M., Neto, A., Fagulha, T., Almeida, V., and Fevereiro, M. (2023), Molecular Characterization and Phylogenetic Analysis of Newcastle Disease Viruses Isolated in Southern Angola, 2016-2018, Infection, Genetics and Evolution, 113, 105481.

[6]. Kapli, P., Yang, Z., and Telford, M. J. (2020), Phylogenetic Tree Building in The Genomic Age, Nature Reviews Genetics, 21(7), 428-444.

[7]. Karthik, S., Basker, S., Dheeban Shankar, P., Senthil Kumar, N., Sarathbabu, S., and Saravanan, K. Assessing the Phylogenetic Relationships of Baliospermum solanifolium [B. montanum (Willd.) Müll. Arg.] in South India through DNA Barcoding.

[8]. Nei, M., and Kumar, S. (2000), Molecular Evolution and Phylogenetics, Oxford University Press, New York.

[9]. Richards, F. S. (1961), A Method of Maximum-Likelihood Estimation, Journal of the Royal Statistical Society: Series B (Methodological), 23(2), 469-475.

[10]. Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes, H. W., ... , Baumeister, W. (2000), The Genome Sequence of The Thermoacidophilic Scavenger *Thermoplasma Acidophilum*. Nature, 407(6803), 508-513.

[11]. Saitou, N., and Nei, M. (1987), The Neighbor-Joining Method: a New Method for Reconstructing Phylogenetic Trees, Molecular Biology and Evolution, 4(4), 406-425.

[12]. Sarıçam, S., and Müştak, H. K. (2015), Filogenetik Ağaçlandırma Metotları, Etlik Veteriner Mikrobiyoloji Dergisi, 26(2), 58-64.

[13]. Singh, G. B. (2015), Fundamentals of Bioinformatics and Computational Biology, Cham: Springer International Publishing, 159-170.

[14]. Sout, N. M., Othman, R., Suliman, N. A., Murgaya, D., Harmin, S. A., and Zan, M. S. M. (2023, June), Phylogenetics Analysis of Blackwater Fishes in Peat Swamp Forest Using Cytochrome B Gene, In AIP Conference Proceedings (Vol. 2625, No. 1), AIP Publishing.

[15]. Tamura, K., Stecher, G., and Kumar, S. (2021), MEGA11: Molecular Evolutionary Genetics Analysis Version 11, Molecular Biology and Evolution, 38(7), 3022-3027.

[16]. Tavaré S. (1986), "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences" (PDF). Lectures on Mathematics in the Life Sciences. 17, 57–86.

[17]. Úsuga-Monroy, C., Díaz, F. J., González-Herrera, L. G., Echeverry-Zuluaga, J. J., and López-Herrera, A. (2023), Phylogenetic Analysis of the Partial Sequences of the *env* and *tax* BLV Genes Reveals the Presence of Genotypes 1 and 3 in Dairy Herds of Antioquia, Colombia, VirusDisease, 1-15.

[18]. Varin, C., Reid, N., and Firth, D. (2011), An overview of composite likelihood methods. Statistica Sinica, 5-42.

[19]. Woese, C. R., Stackebrandt, E., Macke, T. J., and Fox, G. E. (1985), A Phylogenetic Definition of The Major Eubacterial Taxa, Systematic and Applied Microbiology, 6(2), 143-151.

[20]. Xiong, J. (2006), Essential Bioinformatics, Cambridge University Press.