IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Phylogenetic Tree Construction of Bacterial Species using Clustering Algorithms In MEGA 7

A. Sharma[1*], R.S.Thakur[2], S. Jaloree [3]

[1] Department of Computer Science, Anand Vihar College For Women,Bhopal,India
[2] Department of Computer applications, MANIT, Bhopal, India
[3] Department of Computer Science, SATI ,Vidisha,India

Corresponding Author: akansha_dolly@yahoo.com, Tel.: +00-942405-59333

DOI: https://doi.org/10.26438/ijcse/v7i5.11541157 | Available online at: www.ijcseonline.org

Accepted: 19/May/2019, Published: 31/May/2019

Abstract— A phylogenetic tree is a tree that shows the transformative similarity and dissimilarity among various biological species. The biological species may be human species or bacterial species. The comparative analysis of phylogenetic tree is useful in various areas. In this paper phylogenetic tree is constructed for various bacterial species of Rhizobium by using MEGA7 software. MEGA is molecular evolutionary genetics analysis user friendly software for framing sequence alignments and phylogenetic tree construction. This paper also infer us about the how different algorithms like UPGMA, Neighbour joining are implemented effectively on the bacterial species of Rhizobium.

Keywords—UPGMA,Neighbour joining,Phylogenetic tree.

I. INTRODUCTION

In the bioinformatics, large volume of biological data is generated through experiments. Before the data is analyzed and classified properly, it is not useful and interpretative. The phylogenetic tree construction helped the researchers to understand the biological data [1]. A phylogenetic tree is the study of similar and dissimilar property among the species. This tree also represents the relation between intra species. It is the evolutionary tree or graph that shows new relationships among species & it works on genetic closeness [2]. It enhances our understanding of how genes and species evolve. It is also helpful for future prediction. This tree can discover morphological as well as chemical characters of genes. The methods which are used to construct the phylogenetic tree are broadly divided into two main categories- a) distance based method b) character based method

A phylogenetic tree study is useful in almost all areas of

- Medical
- Agriculture
- Industry

Application of Phylogenetic tree broadly divided in the areas[3].

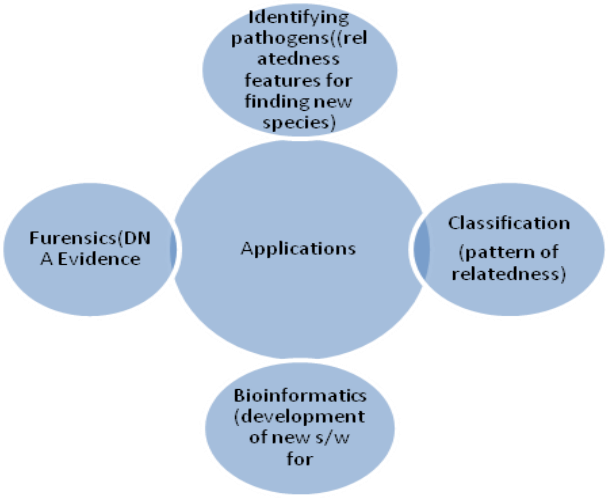
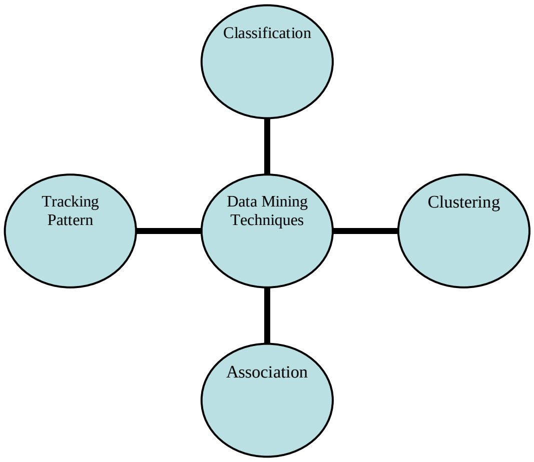

Identifying pathogens((rel atedness features for finding new species)
Furensics(DN A Evidence
Applications
Classification (pattern of relatedness)
Bioinformatics (development of new s/w for

Figure.1

Data Mining

Data mining build models by using the process of uncovering patterns and trends in large database [4]. Data mining has many techniques to uncover the information [5]. Data mining techniques are



Fiure.2

The present manuscript tries to resolve problems of the researchers regarding phylogenetic tree constructions by various distance based Algorithms used [6]. The readers of this manuscript will find some way out in comparing the relationships between given bacterial samples in respect of efficiency of the software (MEGA7)[7]. The manuscript also provides biologist of agriculture field to get some future pathway in understanding the genetical construction of the given bacterial species.

The article is structured in five different sections in order to study and understand easily about the research undertaken. Section I composed of the Introduction of topic. Section II contains the Related work guiding reader about some past researches. Section III briefs a Methodology applied for generating Data and problem solving activity. Section IV contains Results obtained during the process and discussed in as easy possible manner. Section V Conclude and draws attention of the reader about Limitations and Future scope of the particular study

II. RELATED WORK

D.V Chandra shekar et.al (2013) chooses the best algorithm for analysis of clustering methods for biomedical domain. In this work biomedical data extract from UCI machine learning repository. They use different attributes on breast cancer Wisconsin data set like clump thickness, uniformity of cell shape, cell size etc. Finally they conclude that EBM produces better results.

Mahapatro et.al (2012) constructed the phylogenetic tree for DNA sequence using different clustering methods. In this work they use three different clustering algorithms named K-mean medoid and DBSCAN. They conclude that the DBSCAN is performing better in many respects in future.

J.Yang et.al (2010) believes that the phylogenetic tree can construct based on the minimum spanning tree of the complete graph. In this work we compare the phylogenetic tree using minimum spanning method with the neighbor joining method in phylip and they conclude that the minimum spanning tree does not needs multiple sequence alignment.

Jeffrey rizzo et.al (2007) give a review of phylogenetic tree construction. They discuss common approaches algorithm like UPGMA and neighbor joining, Maximum parsimony and the alternative method ATO (ant colony optimization. They conclude that the ATO seems to be the most promising of the two traditional methods

III. METHODOLOGY

The phylogenetic tree of Azetobacter and rhyzobium species is generated with the help of MEGA 7 software. The genomic data of Azetobacter and rhizobium species in the form of nucleotide is obtained from NCBI. NCBI has a big data set of nucleotides having different sizes of almost all species [8]. There are linear and circular sequences of nucleotide having different accession numbers in NCBI.

Steps followed for phylogenetic analysis

1. NCBI (open bank) is used as a data resource of nucleotide sequences. 30 linear nucleotide
2. Sequences of azetobacter and rhizobium species were chosen randomly with size constraint.
3. The selected nucleotide sequences are downloaded in fasta format
4. The downloaded nucleotide sequence file is added in MEGA7 software for further analysis.
5. Alignment of the sequences in MEGA7 was done by the tool called muscle.
6. From the different clustering algorithms available in MEGA 7 to make phylogenetic tree. Some of them were used taking care of various factors used to analyze the phylogenetic tree.
7. Phylogenetic tree was generated by using UPGMA algorithm and neighbor joining algorithm
8. Selection of nucleotide sequences, protein coding and standard genetic codes was done for each and every algorithm.
9. Analysis of generated phylogenetic trees for relationships interpretations of rhizobium species was done.

IV. Results and Discussion

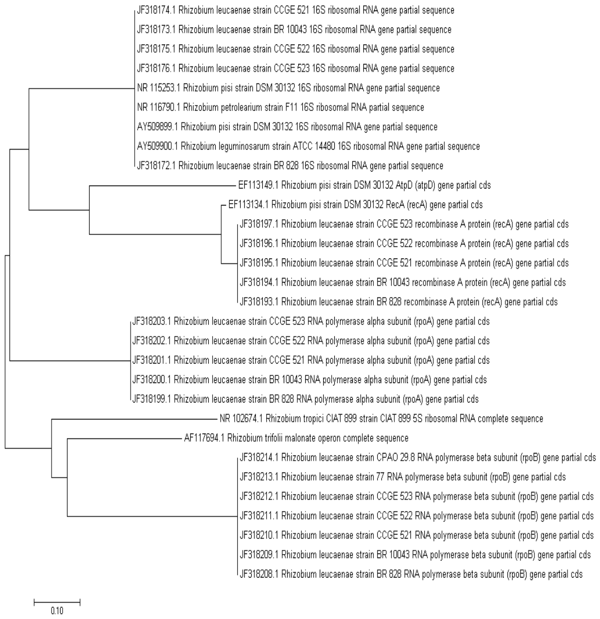

JF318174.1 Rhizobium leucaenae strain CCGE 521 16S ribosomal RNA gene partial sequence
JF318173.1 Rhizobium leucaenae strain BR 10043 16S ribosomal RNA gene partial sequence
JF318175.1 Rhizobium leucaenae strain CCGE 522 16S ribosomal RNA gene partial sequence
JF318176.1 Rhizobium leucaenae strain CCGE 523 16S ribosomal RNA gene partial sequence
NR 115253.1 Rhizobium pisi strain DSM 30132 16S ribosomal RNA gene partial sequence
NR 116790.1 Rhizobium petrolearium strain F11 16S ribosomal RNA partial sequence
AY509899.1 Rhizobium pisi strain DSM 30132 16S ribosomal RNA gene partial sequence
AY509900.1 Rhizobium leguminosarum strain ATCC 14480 16S ribosomal RNA gene partial sequence
JF318172.1 Rhizobium leucaenae strain BR 828 16S ribosomal RNA gene partial sequence
EF113149.1 Rhizobium pisi strain DSM 30132 AtpD (atpD) gene partial cds
EF113134.1 Rhizobium pisi strain DSM 30132 RecA (recA) gene partial cds
JF318197.1 Rhizobium leucaenae strain CCGE 523 recombinase A protein (recA) gene partial cds
JF318196.1 Rhizobium leucaenae strain CCGE 522 recombinase A protein (recA) gene partial cds
JF318195.1 Rhizobium leucaenae strain CCGE 521 recombinase A protein (recA) gene partial cds
JF318194.1 Rhizobium leucaenae strain BR 10043 recombinase A protein (recA) gene partial cds
JF318193.1 Rhizobium leucaenae strain BR 828 recombinase A protein (recA) gene partial cds
JF318203.1 Rhizobium leucaenae strain CCGE 523 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318202.1 Rhizobium leucaenae strain CCGE 522 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318201.1 Rhizobium leucaenae strain CCGE 521 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318200.1 Rhizobium leucaenae strain BR 10043 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318199.1 Rhizobium leucaenae strain BR 828 RNA polymerase alpha subunit (rpoA) gene partial cds
NR 102674.1 Rhizobium tropici CIAT 899 strain CIAT 899 5S ribosomal RNA complete sequence
AF117694.1 Rhizobium trifolii malonate operon complete sequence
JF318214.1 Rhizobium leucaenae strain CPAO 29.8 RNA polymerase beta subunit (rpoB) gene partial cds
JF318213.1 Rhizobium leucaenae strain 77 RNA polymerase beta subunit (rpoB) gene partial cds
JF318212.1 Rhizobium leucaenae strain CCGE 523 RNA polymerase beta subunit (rpoB) gene partial cds
JF318211.1 Rhizobium leucaenae strain CCGE 522 RNA polymerase beta subunit (rpoB) gene partial cds
JF318210.1 Rhizobium leucaenae strain CCGE 521 RNA polymerase beta subunit (rpoB) gene partial cds
JF318209.1 Rhizobium leucaenae strain BR 10043 RNA polymerase beta subunit (rpoB) gene partial cds
JF318208.1 Rhizobium leucaenae strain BR 828 RNA polymerase beta subunit (rpoB) gene partial cds
0.10

Fig.1 Phylogenetic tree of rhizobium species by neighbor joining algorithm in mega7 software.

The neighbor joining method shows metamorphic past. The tree generated infers comparative structure of rhizobium species. The tree with the sum of branch length = 2.42834891 is generated Maximum likelihood method is used for computing the evolutionary distances, and are represented by the units of the number of base substitutions per site. 30 nucleotide sequences are evaluated. Positions having gaps and missing data are removed. There are total of 45 positions in the final dataset in all. Evolutionary analyses is conducted in MEGA7 .

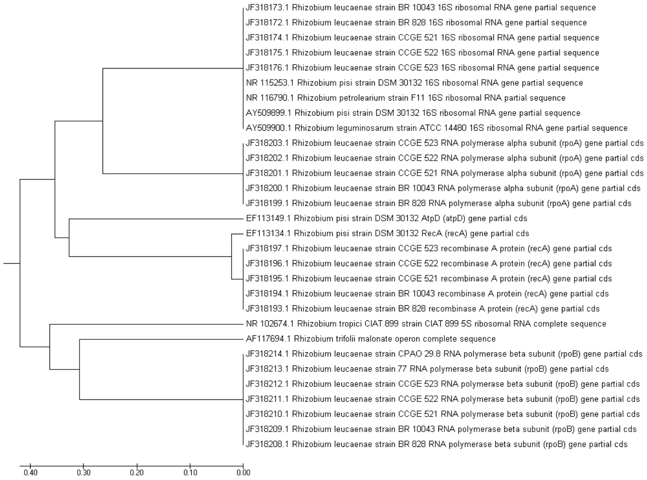

JF318173.1 Rhizobium leucaenae strain BR 10043 16S ribosomal RNA gene partial sequence
JF318172.1 Rhizobium leucaenae strain BR 828 16S ribosomal RNA gene partial sequence
JF318174.1 Rhizobium leucaenae strain CCGE 521 16S ribosomal RNA gene partial sequence
JF318175.1 Rhizobium leucaenae strain CCGE 522 16S ribosomal RNA gene partial sequence
JF318176.1 Rhizobium leucaenae strain CCGE 523 16S ribosomal RNA gene partial sequence
NR 115253.1 Rhizobium pisi strain DSM 30132 16S ribosomal RNA gene partial sequence
NR 116790.1 Rhizobium petrolearium strain F11 16S ribosomal RNA partial sequence
AY509899.1 Rhizobium pisi strain DSM 30132 16S ribosomal RNA gene partial sequence
AY509900.1 Rhizobium leguminosarum strain ATCC 14480 16S ribosomal RNA gene partial sequence
JF318203.1 Rhizobium leucaenae strain CCGE 523 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318202.1 Rhizobium leucaenae strain CCGE 522 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318201.1 Rhizobium leucaenae strain CCGE 521 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318200.1 Rhizobium leucaenae strain BR 10043 RNA polymerase alpha subunit (rpoA) gene partial cds
JF318199.1 Rhizobium leucaenae strain BR 828 RNA polymerase alpha subunit (rpoA) gene partial cds
EF113149.1 Rhizobium pisi strain DSM 30132 AtpD (atpD) gene partial cds
EF113134.1 Rhizobium pisi strain DSM 30132 RecA (recA) gene partial cds
JF318197.1 Rhizobium leucaenae strain CCGE 523 recombinase A protein (recA) gene partial cds
JF318196.1 Rhizobium leucaenae strain CCGE 522 recombinase A protein (recA) gene partial cds
JF318195.1 Rhizobium leucaenae strain CCGE 521 recombinase A protein (recA) gene partial cds
JF318194.1 Rhizobium leucaenae strain BR 10043 recombinase A protein (recA) gene partial cds
JF318193.1 Rhizobium leucaenae strain BR 828 recombinase A protein (recA) gene partial cds
NR 102674.1 Rhizobium tropici CIAT 899 strain CIAT 899 5S ribosomal RNA complete sequence
AF117694.1 Rhizobium trifolii malonate operon complete sequence
JF318214.1 Rhizobium leucaenae strain CPAO 29.8 RNA polymerase beta subunit (rpoB) gene partial cds
JF318213.1 Rhizobium leucaenae strain 77 RNA polymerase beta subunit (rpoB) gene partial cds
JF318212.1 Rhizobium leucaenae strain CCGE 523 RNA polymerase beta subunit (rpoB) gene partial cds
JF318211.1 Rhizobium leucaenae strain CCGE 522 RNA polymerase beta subunit (rpoB) gene partial cds
JF318210.1 Rhizobium leucaenae strain CCGE 521 RNA polymerase beta subunit (rpoB) gene partial cds
JF318209.1 Rhizobium leucaenae strain BR 10043 RNA polymerase beta subunit (rpoB) gene partial cds
JF318208.1 Rhizobium leucaenae strain BR 828 RNA polymerase beta subunit (rpoB) gene partial cds
0.40
0.30
0.20
0.10
0.00

Fig. 2 Phylogenetic tree of rhizobium species by UPGMA algorithm in mega7 software.

The UPGMA method also helps in analyzing metamorphic past. The phylogenegtic tree generated gives a comparative structure of rhizobium species .The tree with the sum of branch length = 2.47744462 is generated . UPGMA method is used for computing the evolutionary distances, and are represented by the units of the number of base substitutions per site. 30 nucleotide sequences are evaluated . Positions having gaps and missing data are removed. There are total of 45 positions in the final dataset in all. Evolutionary analyses is conducted in MEGA7 .

While conducting this research it was found that during the construction of phylogenetic trees of rhizobium species by neighbourjoining and UPGMA method in MEGA7 .The evolutionary distances computed by UPGMA method was almost ultra-metric in nature were as computation done with neighbourjoining method is of additive nature, almost 0.5 times better in neighbour joining as compared to UPGMA method. This shows better similarity and dissimilarity representation in neighbour joining method while using MEGA7 for the particular species of rhizobium.

V. Conclusion and Future Scope

To make new data mining method for biological data, MEGA7 Software package proofs to be useful and reliable. MEGA7 implementation on given bacterial species provides an easy to use comparison based Method. However a comparison of efficiency explains that the neighbour joining method implementation is more accurate than UPGMA Method in MEGA7 for the specific biological dataset. This study gives a way ahead to further go for more detailed and specific go through of comparison analysis and can lead to a development of new model for relationship alignment.

Acknowledgment

I would like to express my deep gratitude to Dr.R.S Thakur and Dr.Shailesh Jaloree, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to thank Dr. Kiran Singh for her guidance in understanding biological background. I would also like to extend my thanks to the technicians of the laboratory of the computer science department of SATI Vidisha for their help in offering me the resources in running the program.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

References

[1] J.Rizzo, E.C.Rouchka,"Review of phylogenetic tree construction",Bioinformatics laboratory technical report series,pp.1-7,2007.

[2] G.Mahapatro,D.Mishra,k.Shaw,S.Mishra, T.Jena,"Phylogenetic tree construction for DNA Sequences using clustering methods", In the

proceedings of 2012, International conference on modeling optimization and computing,

[3] D.V.Chandra shekhar, V.V.J.R.Krishnaiah, Y.S.Babu, “Association of data mining and biolomedical domain:choosing the best algorithm for analysis”, International journal of data mining and emerging technologies, vol.3,issue.1, pp.40-46, 2013.

[4] S Hussain,”Survey on current trends and techniques of data mining research”, Londan journal press, Vol.17, Issue.1 ,2017.

[5] D.Patel, R.Modi, K. Sarvakar, ”A comaritive study of clustering data mining:Techniques and research challenges”, IJLTEMAS, Vol.3, Issue.9,2014.

[6] A.Joshi, R.Kaur, ”A Review:comparative study of various clustering techniques in data mining”, International journal of advanced research in computer science and software enginnering, Vol.3, issue.3, 2013.

[7] K.Tamura, D.Peterson, N.Peterson,G.stecher, M.Nei, S.Kumar, “MEGA 5:Molecular evolutionary genetics using maximum likelihood,evolutionary distance,and maximum parsimony methods”, Mol.Biol.Evol., vol-28, issue-10,pp.2731-2739,2011.

[8] W.P. Hanage, C.Fraser,B.G.Spratt, “Sequences,sequence clusters and bacterial species”, Philosophical transactions of the royal society B”, Vol.361,pp.1917-1927,2006.

Authors Profile

Mrs. Akansha Sharma pursed Bachelor of Science from B.U. Bhopal and Masters in computer application from RGPV Bhopal,india. She is currently pursuing Ph.D. and working as Assistant Professor in Anand Vihar College for women, Bhopal. She has published 5 research papers in reputed National journals . She has published 1 online paper in springer link. Her main research work in implementation of data mining algorithm in biological data.

Dr.R.S Thakur is curently working as a Associate Profesor in MANIT, Bhopal . He persued M.C.A and M.Tech (C.S.E) from RGPV University. He completed his Ph.D from RGPV,bhopal. He is being indulged in research and academics since last 20 years. He published approximately 200 research papers in various national and international journals of repute.

Dr.Shailesh Jaloree is currently working as a Professor and HOD in Dept. of Applied Mathematics at SATI Vidisha,M.P. He is a post graduate and Ph.D in Maths having 20 years of experience in research and academics.