International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Computational Analysis of Distance based Phylogenetic Tree for Azotobacter Species

Akansha Sharma, R.S Thakur, Shailesh Jaloree

Abstract: ***Phylogenetic tree is a pictorial representation of evolutionary relationships between organisms. It is important method to analyze the biological data. Phylogenetic trees are based on two methods : Distance based and Character based. Phylogenetic tree are used comparative analysis of any organism like human Beings, Animals, Bacteria, Viruses and Fungi's etc. In this paper we compare 12 different nucleotide sequences of Azotobacter species having linear DNA of 999 BP as maximum size using substitution model and phylogenetic model. In this study two different models name P-Distance and Jukes cantor model are used and helped in finding UPGMA or Neighbour joining method efficiency in evaluating the similarity and dissimilarity of bacterial species. This paper gives influence in reconciliation of Azotobacter species to produce phylogram with informative branch lengths. This further leads to analyze and understand various expressive characters of Azotobacter in agriculture field.***

Index Terms**: *Phylogenetic tree, UPGMA, Neighbour joining, P-Distance, Jukes cantor***

I. INTRODUCTION

A phylogenetic tree is a branching tree that gives similarity and dissimilarity of organisms based on physical and genetic characteristics[1]. Phylogenetic tree can be represented in two ways rooted tree and unrooted tree. Rooted tree gives direction tree with the unique node called root node and the unrooted tree give the relation of leaf nodes without ancestral root. It is very useful in the field of bioinformatics [2]. When clustering algorithms of data mining is included in the phylogenetic tree, it gives more detailed or accurate relatedness. UPGMA and Neighbour joining clustering algorithms are mostly used to make distance based phylogenetic tree[3][4]. There are so many software like phylip and MEGA for constructing and analyzing the phylogenetic tree [5]. In MEGA software, there are so many models which can be used to analyze the tree [6]. Some models are P-distance, Jukes Cantor, Poisson etc. The MEGA software has the ability to make distance matrix based as Character based matrix on different models [7]. We are interested in comparing distance based matrices and worked over it. Data mining is the method of extracting information for the patterns and models from large broad datasets [8]. Before implementing data mining algorithms, Some preprocessing steps have to be done like data cleaning process, data selection and transformation process[9]. Association, Classification, Clustering are the different methods which are applied in analysis of pattern in data mining. To draw the conclusion of biological data or analyze the biological data, data mining method are commonly used and they play important role. There are various clustering or classification algorithms that are implemented in bioinformatics [10]. V.Sohpal et.al using different substitution models to compare the capsid protein of HHV to find relationship between proteins. They also analyze the effect of poisson correction with shape parameter. They give the computational analysis of phylogenetic tree for both distance and character based tree. Finally they conclude that taxon separation of proteins give maximum similarities and can use as a drug for human therapy. A.Smith et.al gives the methods for aligning, synthesizing and analyzing rooted phylogenetic trees within a graph, called a tree alignment graph (TAG). This model can be used for large scale analysis and resolve the problem for find common node and edges. 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.

II. METHODOLOGY

Phylogenetic tree construction and analysis have four phases: 1. Select the sequences and download in fasta format. 2. Align the sequences from any selected method that are used in multiple sequence alignment. 3. Construct the distance matrix for different parameters. 4. Tree building and evaluation In Phylogenetic tree analysis, the important source of data is NCBI (National Center for biotechnology Information).NCBI has different databases like gene bank for DNA sequences and many more . It is having thousands of nucleotides sequences for almost all organisms in two different classifications linear and non linear with different sizes, which are useful in the construction of phylogenetic tree[11]. In this paper we are using 12 different nucleotide sequences of Azotobacter chrococcum strains with whole genome shotgun sequences with different nodes in them. They all are the family of Azotobacter chrococcum species. But some where they are different from each other at molecular genetic level. Following nucleotide sequence are downloaded in Fasta format. Among many soft wares available to make a phylogenetic tree, we have chosen MEGA Software that constitutes of all features to make align, make distance matrix, tree evaluation and so on[12]. Here in this study P-Distance and Jukes cantor model were used as a different parameters and compare it with result.

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Mrs Akansha Sharma, Computer Science Dept., Anand Vihar college for women, Bhopal,India
Dr.R.S. Thakur, MCADept. MANIT, Bhopal, India.
Dr. Shailesh Jaloree, Department of mathematics, SATI, Vidisha, India.

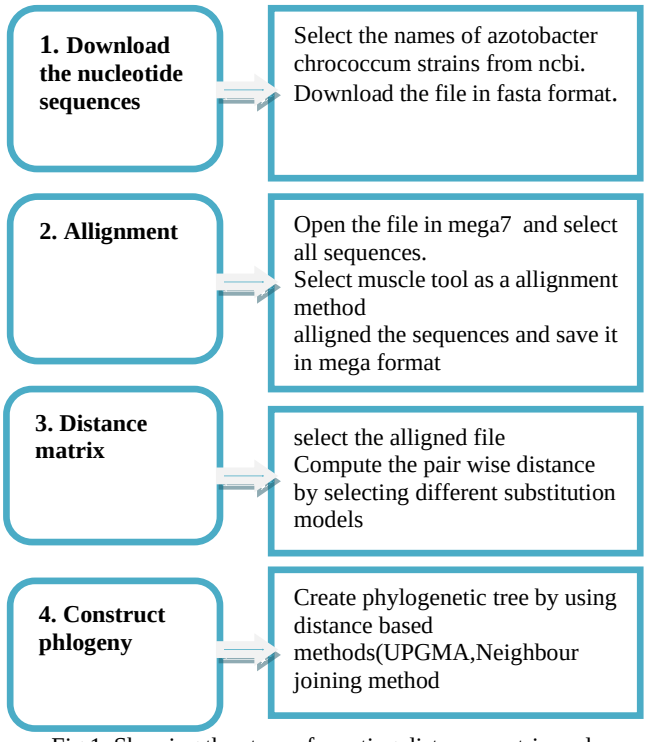

Fig 1. Showing the steps of creating distance matrix and phylogenetic tree

III. RESULT

A. *Substitution models*

Distance measures are the important tool by which evolution is studied at molecular level, phylogenetic reconstruction. In this study azotobacter chrococcum has undergone estimation by using P-distance and jukes cantor model for analyzing the evolutionary divergence of base pair sequences of whole genome. Table.1 shows the data of distance against 12 whole genome sequence of azotobacter chrococcum for P-distance model. In continuation of the study table.2 gives the data for distance against 12 whole genome sequence of azotobacter chrococcum for jukes cantor model. The study gives a critical evaluation of distance between Azotobacter chroococcum strain ATCC 9043 Node 115 and Azotobacter chroococcum strain ATCC 9043 Node 113 has minimum distance in both the substitution models. The distance divergence among these two strains is 0.255 and 0.312 for the P-distance and jukes cantor model respectively(Table.3). On the other side Azotobacter chroococcum strain ATCC 9043 Node 106, Node 104 and Node 111 shows maximum distance divergence of 0.617 and 1.297 with Azotobacter chroococcum strain ATCC 9043 Node 110 in both the models respectively.All tables.1,2and 3 shows that results are qualitatively consistent irrespective of models used. Lower divergence and the higher divergence between the strains signifies the similar and dissimilar relationship among them.

Table 1. P-Distance matrix

Species No./Name	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
[1]Azotobacter_chroococcum_strain_ATCC9043NODE117											
[2]Azotobacter_chroococcum_strainATCC9043NODE114	0.539										
[3]Azotobacter_chroococcum_strainATCC9043NODE110	0.596	0.553									
[4]Azotobacter_chroococcum_strainATCC9043NODE109	0.504	0.511	0.553								
[5]Azotobacter_chroococcum_strainATCC9043NODE106	0.574	0.511	**0.617**	0.546							
[6]Azotobacter_chroococcum_strainATCC9043NODE104	0.582	0.468	**0.617**	0.546	0.553						
[7]Azotobacter_chroococcum_strainATCC9043NODE101	0.518	0.447	0.539	0.475	0.489	0.447					
[8]Azotobacter_chroococcum_strainATCC9043NODE111	0.539	0.567	**0.617**	0.560	0.567	0.582	0.553				
[9]Azotobacter_chroococcum_strainATCC9043NODE113	0.511	0.539	0.546	0.525	0.546	0.496	0.489	0.539			
[10]Azotobacter_chroococcum_strainATCC9043NODE115	0.553	0.546	0.553	0.553	0.539	0.553	0.553	0.560	**0.255**		
[11]Azotobacter_chroococcum_strainATCC_9043_NODE_112	0.567	0.525	0.596	0.596	0.589	0.553	0.546	0.532	0.525	0.567	
[12]Azotobacter_chroococcum_strain_ATCC_9043_NODE_103	0.589	0.511	**0.617**	**0.617**	0.603	0.454	0.496	0.553	0.532	0.553	0.574

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 2. Jukes cantor model

Species No/Name	[1]	[2	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	
[1]Azotobacter_chroococcum_strain_ ATCC9043NODE117												
[2]Azotobacter_chroococcum_strain ATCC9043NODE114	0.951											
[3]Azotobacter_chroococcum_strain ATCC9043NODE110	1.186	1.003										
[4]Azotobacter_chroococcum_strain ATCC9043NODE109	0.835	0.857	1.003									
[5]Azotobacter_chroococcum_strain ATCC9043NODE106	1.089	0.857	**1.297**	0.977								
[6]Azotobacter_chroococcum_strain ATCC9043NODE104	1.120	0.734	**1.297**	0.977	1.003							
[7]Azotobacter_chroococcum_strain ATCC9043NODE101	0.879	0.679	0.951	0.753	0.793	0.679						
[8]Azotobacter_chroococcum_strain ATCC9043NODE111	0.951	1.059	**1.297**	1.031	1.059	1.120	1.003					
[9]Azotobacter_chroococcum_strain ATCC9043NODE113	0.857	0.951	0.977	0.902	0.977	0.813	0.793	0.951				
[10]Azotobacter_chroococcum_strain ATCC9043NODE115	1.003	0.977	1.003	1.003	0.951	1.003	1.003	1.031	**0.312**			
[11]Azotobacter_chroococcum_strain ATCC_9043_NODE_112	1.059	0.902	1.186	1.120	1.152	1.003	0.977	0.926	0.902	1.059		
[12]Azotobacter_chroococcum_strain _ATCC_9043_NODE_103	1.152	0.857	**1.297**	1.059	1.221	0.697	0.813	1.003	0.926	1.003	1.089	

Table:-3 Comparision of Models

B. Distance Based Phylogenetic tree

Model	Lower divergence	Higher divergence	Species of Lower divergence	Species of higher divergence
P-Distance	0.255	0.617	Node 115-Node113	Node 103-Node 109,Node111,Node104, Node 106
Jukes Cantor	0.312	1.297	Node 115-Node 113	Node 103-node 111,104,106

This phylogenetic method measures the distance between pair of sequences observing dissimilarities as well as similarities computed on the basis of sequence alignment. The primary thought of distance based method is homology among the sequences and are summed up by help of tree branches. The neighbor joining and UPGMA Method based on clustering algorithm are used in this paper for 12 whole genome sequence of azotobacter chrococcum species. The tree shown in fig 2 and fig 3 is output of neighbour joining methodology and for which the p-distance and jukes cantor distance are shown in table.1 and table.2 respectively. On comparing the fig2 and fig.3 with table1 and table2, distance matrix the two closest operational taxonomic units(OTUs) i.e Azotobacter chroococcum strain ATCC 9043 Node 115 and Azotobacter chroococcum strain ATCC 9043 Node 113 has the lowest taxon separation and 100% , 99% similarity in neighbour joining tree. The tree shown in fig 4 and fig 5 is an observation of UPGMA method for which the p-distance and jukes cantor distance are shown in table.1 and table.2 respectively. On comparative analysis of the fig4 and fig.5 with table1 and table2, distance matrix the two closest OTUs i.e Azotobacter chrococcum strain ATCC 9043 Node 115 and Azotobacter chrococcum strain ATCC 9043 Node 113 has the lowest taxon separation and 100% similarity in UPGMA tree.

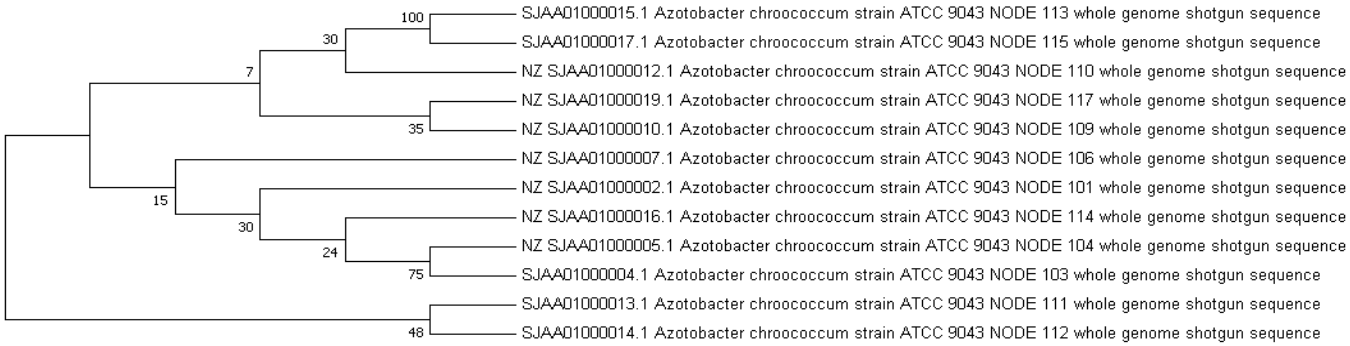

Fig2. Neighbour joining method for P distance model (bootstrap consenses)

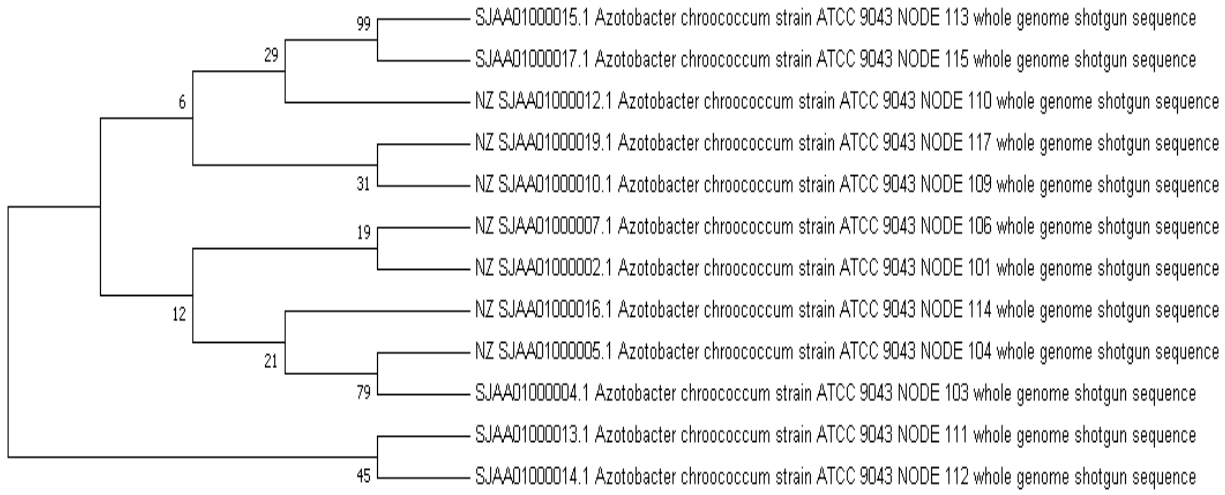

Fig3. Neighbour joining method for Jukes cantor modele (bootstrap consenses)

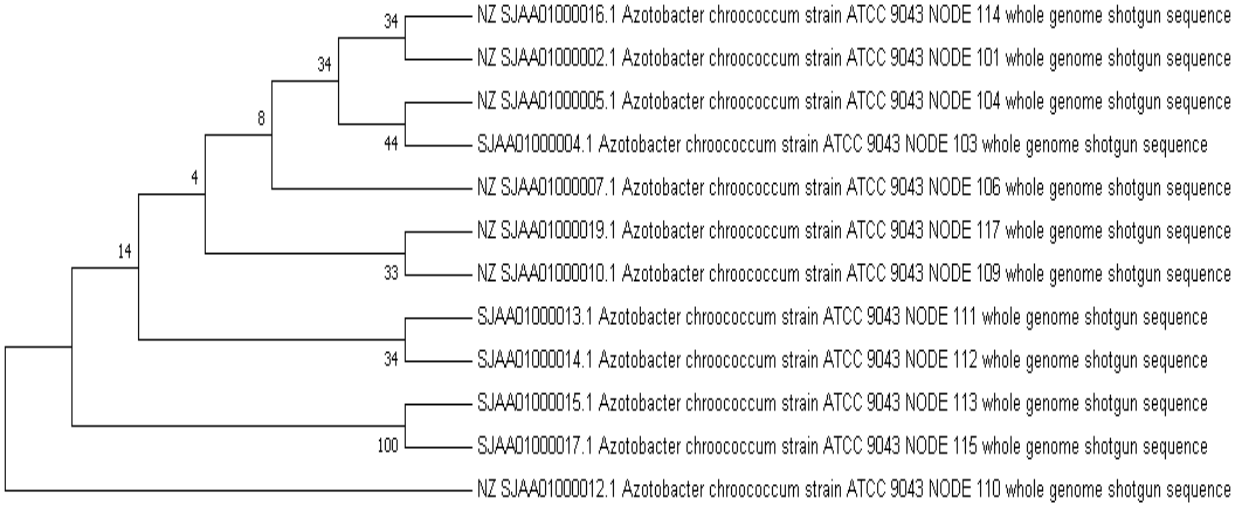

Fig4. UPGMA method for P distance model (bootstrap consenses)

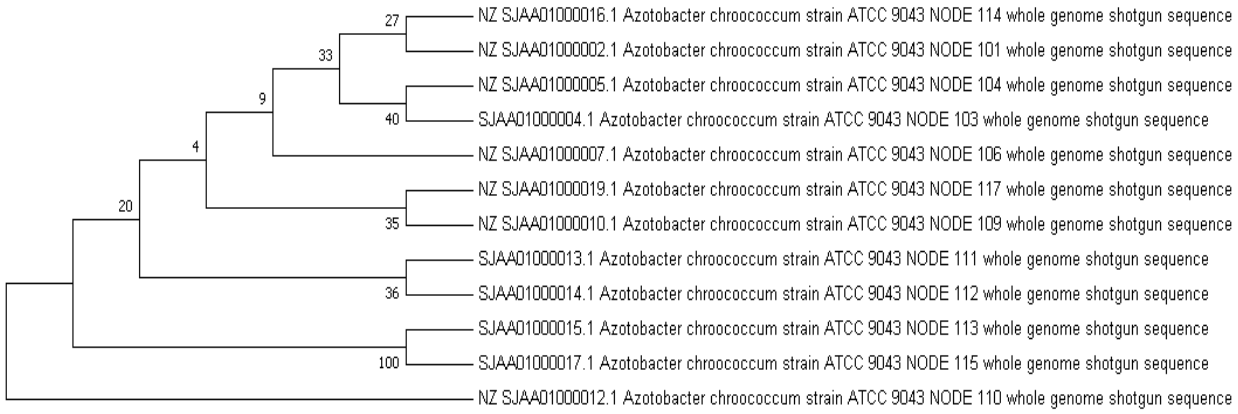

Fig5. UPGMA method for Jukes cantor model (bootstrap consenses

Retrieval Number: J90060881019/19©BEIESP
DOI: 10.35940/ijitee.J9006.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

IV. DISCUSSION

Substitution model proofs to be a powerful tool for reconstructing distance based phylogenetic trees . The present study support this argument with proficiency and authenticity. The study shows that UPGMA method is efficient to obtain correct tree on the basis of distance data for Azotobacter chrococcum. Several lines of evidence suggest that both the UPGMA and NJ estimates are robust in calculating evolutionary distances.

V. CONCLUSION

The research of specific biological dataset of Azotobacter chrococcum reveals the importance of different methods used to analyze evolutionary a likeness or differences . Both the methods used In this present study gives us a presumption that they are able to recreate distance based phylogenetic trees for the given bacterial species with utmost efficiency leading to proficient analysis of the relationships among the bacterial species

REFFERENCE

1. J.Rizzo, E.C.Rouchka,"Review of phylogenetic tree construction", Bioinformatics laboratory technical report series,pp.1-7,2007.
2. Krause,N.N.Diaz,A.Goesmann,S.Kelley,T.W.Nattkemper,"Phylogene tic classification of short environmental DNA fragments", Nucleic acids Res,Vol.36, pp. 2230-2239, 2008
3. J.Felsenstein, "Confidance limits on phylogenies: an approach using bootstrap", Evolution,Vol.39, pp. 368-376, 1981.
4. 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,
5. A.stamatkis, "Phylogenetics:applications,software and challenges", Cancer genomics and proteomics, vol.2, pp. 301-306, 2005.
6. V.K.Sohpal, A.Dey, A. Singh ,"Computational analysis of distance and character based phylogenetic tree for capsid proteins of human herpes virus", Data mining in Genomics & Proteomics, Vol.4,pp. ,2013.
7. S.Kumar, G.Stecher, K.Tamura, "MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets", Molecular biology and evolution, Vol.33, pp. 1870-1874, 2016.
8. S.Hussain,"Survey on current trends and techniques of data mining research", Londan journal press, Vol.17, Issue.1 ,2017.
9. S.Sahu,S.Sharma,S.Gondhalkat, "A brief overview on data mining survey", IJCTEE, Vol.1, Issue.3, pp. 114-119, 2011.
10. D.Patel, R.Modi, K. Sarvakar, "A comaritive study of clustering data mining:Techniques and research challenges", IJLTEMAS, Vol.3, Issue.9,2014.
11. F.P.Martins,O.S.Paula, "NCBI mass sequence downloader large dataset downloading made easy", SoftwareX, Vol.5, pp. 80-83, 2016.
12. 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.

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 is a member of the CSI, IEEE, ACM, IAENG, ISTE, GAMS and IACSIT.

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