Journal of Pharmacognosy and Phytochemistry
Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
www.phytojournal.com
JPP 2021; Sp 10(1): 227-232
Received: 15-11-2020
Accepted: 02-12-2020

Prabhat Kumar
Research Scholar, Division of Agricultural Statistics, ICAR-IARI, New Delhi, India

SS Patil
Professor of Agricultural Statistics, Applied Mathematics and Computer Science, UAS, Bengaluru, Karnataka, India

Hemamalini HC
Assistant Regional Director, IGNOU, Bengaluru, Karnataka, India

RH Chaudhari
Research Scholar, Division of Agricultural Statistics, ICAR-IARI, New Delhi, India

Rajeev Kumar
Research Scholar, Division of Agricultural Statistics, ICAR-IARI, New Delhi, India

Corresponding Author:
RH Chaudhari
Research Scholar, Division of Agricultural Statistics, ICAR-IARI, New Delhi, India

Efficient classification of sugarcane genomes

Prabhat Kumar, SS Patil, Hemamalini HC, RH Chaudhari and Rajeev Kumar

DOI: https://doi.org/10.22271/phyto.2021.v10.i1Sd.13474

Abstract
A Phylogenetic tree construction to know to the relationship of the ancestral association of species. The genome sequences, outlining the transmission of functional and genetic classification. Analysing the quantitative conduct of phylogenetics in the conservation of biodiversity and the successful heuristics of obtaining an accurate distribution of trees plays a predominant role. The study to know higher accuracy from efficient algorithm to deducing phylogenetic relationship among Sugarcane (Saccharum) species. A sample of 431 Saccharum genome sequences was drawn from NCBI dataset. Efficient algorithms like Maximum Likelihood Estimation (MLE) method and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method were considered to construct the phylogenetic tree. The maximum likelihood with Tamura Nei model, Kimura 2-parameter model and achieves the highest precision, while MLE with Maximum likelihood with Jukes-Cantor model achieves the least. The computational biology of statistically results is justifiable and compares the functional relationship between various models in which error percentage has been reduced. The same algorithms perform on individual species under different models such as maximum likelihood with Kimura 2-parameter model and Tamura Nei model more efficient than others to differentiate the species genomic sequences and group them to correct taxon.

Keywords: MLE (Maximum Likelihood Estimation), Phylogenetic tree, UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

1. Introduction
"Amongst the sugarcane we are safe" say the Chinese with sugarcane (*Saccharum spp*.) symbolizing bravery, independence and protection (DeBernardi, 2009) [5]. The name sugarcane is used to refer to a group of tall perennial tropical grass species which were domesticated for sugar production, and have been classified inconsistently (Paterson *et al*., 2013) [17]. Sugarcane (*Saccharum spp*. hybrids) is the most important sugar-producing crop and raw biofuel material in the world (Zhou *et al*., 2018) [22], which belongs to the grass family Gramineae. The major commercial sugarcane cultivars are all complex interspecies hybrids with allopolyploidy and aneuploidy with a huge genome (Zhou *et al*., 2014; Garsmeur *et al*., 2018) [21, 8]. Sugarcane has been known for more than 2,200 years and it was one of the first plants to inspire humans to develop technology (Goldstein and Mintz, 2015) [9]. Sugarcane (*Saccharum officinarum*) family Gramineae (Poaceae) a widely grown crop in India. It employs over a million people, directly and indirectly contributing to the national exchequer. Sugar cane originated in New Guinea but Sugarcane plants spread along with human migration routes to the Indian subcontinent and other countries. Now cross with some wild sugarcane species produced commercial sugarcane with high sugar contents. Cultivation of sugarcane in India from Vedic period around 1400 to 1000 B.C. It is now widely accepted that India is the original home of *Saccharum* species. It belongs to family Gramineae, class monocotyledons and order glumaceae subfamily panicoidae, tribe andripogoneae, and sub tribe saccharinin.
Genome sequencing is the process of determining the complete DNA sequence of an organism's genome at a single time. This involves sequencing all of a chromosomal DNA as well as DNA confined in the mitochondria and chloroplast. Need for the study of the phylogenetic tree is that in recent days almost all evolutionary relationships are inferred from molecular sequence data. Introduced a space of phylogenetic bushes and described the space among two timber to be the length of the shortest path between them in that area (Billera *et al*., 2001) [3]. Reported phylogenetic inference in building a tree-based at the measured pairwise distance of all species (Fisher and Lindberg, 2006) [7]. As we know that the DNA is the inherited material can analyse easily, quickly, and inexpensive with less time and researchers can also investigate the phylogeny to determine their rates of evolution and can also study the direction of evolution by determining were a disease organism, such as bacterium or virus.

Where in the case, where it is not possible to obtain genetic material, the morphological measurement can be used to infer evolutionary relationships. However, this approach is less reliable than using molecular data because as we know that sometimes the same morphological trait can arise from multiple independent evolutionary lineages and also in bioinformatics phylogenetic tree of proteins and DNA are useful and the multiple sequence alignment, structure conservation, homology detection, and paralogous attributes for sequences are some of the direct applications for phylogenetic trees. Scope and importance on analysis of phylogenetic tree on genome sequences, through phylogenetic, we learn not only how the sequences came to be the way they are today but also general principles that enable us to predict how they will change in the future and phylogenetic is most important because it enriches our understanding of how genes, genomes, species evolve. This is not only of fundamental importance but also extremely useful for numerous applications as follows: Classification of new species based on sequence data gives the most accurate description patterns of relatedness relationships between the ancestors and the phylogenetic is used to assess DNA evidence presented in court cases to inform situations.

2. Material and methods

To acquire the nucleotide sequence of *Saccharum* species were sought over the database NCBI. The data sequences collected in FASTA format and aligned in a mega format. Total 431 sequences were taken from 11 *Saccharum* species for the tree construction and analysis. The 11 *Saccharum* species utilized for the investigation are represented in table1.

Table 1: The 11 *Saccharum* species utilized for the investigation

Sl. No.	Species	Sequence Total
1	*S. arundinaceum*	40
2	*S. asper*	31
3	*S. barberi*	40
4	*S. edule*	40
5	*S. officinarum*	40
6	*S. procernum*	40
7	*S. ravennae*	40
8	*S. robustum*	40
9	*S. sinese*	40
10	*S. sponteneum*	40
11	*S. villosum*	40
Total		431

2.1 Unweighted Pair Group Method with Arithmetic Mean

The UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm, which is mainly used in bioinformatics for constructing the evolutionary tree and where in this UPGMA method is the simple agglomerative or hierarchical clustering method for the development of constructing taxonomic phenograms, which replicate the phenotypic similarities between the operational taxonomical units (OUTs). Phonogram is created by considering at each step, then the nearest two clusters are pooled into a higher-level cluster. Let the clusters be A and B, then the distance between the clusters is taken to be the average of all distances that is “x” in A and “y” in B, hence the mean distance between the cluster:

$$\frac{1}{|A||B|}\sum_{x\in A}\sum_{y\in B} d(x,y)$$

The maximum likelihood and parsimony algorithms, the phylogenetic tree was Built under UPGMA.

2.2 Maximum likelihood algorithm

In a statistical way that the maximum likelihood estimation (MLE) is a method of estimating the parameters of a given observation. Let us assume that there is a sample of $x_1, x_2, x_3, \ldots, x_n$of n independent and identically distributed observations, which are obtained from the distribution of unknown probability density function $f_0(.)$. Summarized that the function f_0 belongs to a certain family of distributions {f $(. |\theta), \theta \epsilon \Theta\}$ (where θ is a vector of parameters) called the parametric model, so that $f_0 = f(. |\theta_0)$. The value θ_0 is the unknown parameter which is referred to as the true value of the parameter vector. It is necessary to estimate the estimator which would be as close to the true value θ_0 as possible. Either or both the observed. Variables x_i and the parameter θ can be the vectors. Before calculating the maximum likelihood estimation, one should specify the joint density function for all the observations of independent and identically distributed sample hence, the joint density function is

$$f(x_1, x_2, x_3, \ldots, x_n/\theta) = f(x_1/\theta) \times f(x_2/\theta) \times, \ldots, f(x_n/\theta)$$

Then the likelihood function that is the log-likelihood is given by,

$$\ln L(\theta;\ x_1, x_2, x_3, \ldots, x_n) = \sum_{i=1}^{n} ln\ f(x_i/\theta)$$

And the average log-likelihood is:

$$\hat{l} = \frac{1}{n} lnL$$

Where, $\hat{l}$ indicates that an estimator, certainly, the $\hat{l}$ estimates the expected log likelihood for the single observation in the model. Hence, by finding the value of θ the maximum likelihood estimates θ_0 can be estimated which maximizes $\hat{l} = (\theta; x)$. Hence, this method of estimation defines a maximum likelihood estimator (MLE) of θ_0.

$$\{\hat{\theta}\} \subseteq \left\{ \underset{\theta\ \epsilon\Theta}{arg\ max}\ \hat{l}(\theta;\ x_1, x_2, x_3, \ldots, x_n) \right\}$$

If maximum exists, then the MLE estimate is the same irrespective of maximizing the likelihood or the log-likelihood function, because the log is a monotonically increasing function.

2.3 Distance-based method

Mainly the distance models are used in constructing a phylogeny as non- parametric distance method. Using the matrix of pair-wise distances, non-parametric distances were initially applied to phonetic data. There are various pair-wise distances, such as Euclidean distance which is applied to discrete morphological characters. For the Constructed phylogenetic tree, the following distance-based models were used: Jukes-Cantor model, Kimura 2-parameter model, Tajima- Nei model and P distance model.

2.3.1 Jukes-Cantor (JC)

JC69 is the simplest substitution model, where it assumes equal base frequencies, that is $\pi_A = \pi_G = \pi_U = \pi_T = 0.25$ and also equal in mutation rates. The one and only parameter of this model is μ . Therefore, μ indicates the overall substitution rate. When we normalize the mean rate to 1 automatically this variable becomes a constant.

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} \\ \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{3}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} \\ \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{3}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} \\ \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}-\frac{1}{4}e^{-\mu t} & \frac{1}{4}+\frac{3}{4}e^{-\mu t} \end{pmatrix}$$

When branch length v is measured in the expected number of changes per site
Then,

$$P_{ij} = \begin{cases} \frac{1}{4}+\frac{3}{4}e^{-\mu t} \; if \; i = j \\ \frac{1}{4}-\frac{1}{4}e^{-\mu t} \; if \; i \neq j \end{cases}$$

It is worth observing that

$$v = \frac{3}{4}e^{-\mu t} = \left(\frac{\mu}{4}+\frac{\mu}{4}+\frac{\mu}{4}\right)t$$

Which stands for the sum of any column (or row) of matrix Q, which is multiplied by time and hence the means expected the number of substitutions in time t (branch duration) for each particular site (per site), when the rate of substitution equals μ.
The Jukes-Cantor estimate of the evolutionary distance between the proportions of p site which differs between the two sequences.

$$\hat{d} = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right) = \hat{v}$$

Where, In the p represents the p-distance is ideal for calculating the jukes-cantor distance correction, whereas it is not much sufficient for finding the evolutionary distances under the complex models.

2.3.2 Kimura 2-parameter model

This model basically differentiates between transitions and transversion, that is a conversion of A to G, *i.e.*, from purine to purine, or C to T, *i.e.*, from pyrimidine to pyrimidine which is called as transition and whereas when the conversion from a purine to pyrimidine or vice versa then it is called as transversion. In Kimura's original description of the model, they α and β were used to denote the rates of these types of substitutions, but it is more usual that the rate of transversions is set to 1 and the K is used to indicate the transition transversion rate ratio (as indicated below). Even the Kimura 2- parameter model assumes that all of the bases are equally frequent.

$$\pi_A = \pi_G = \pi_U = \pi_T = 0.25$$

Rat matrix Q = $\begin{pmatrix} * & k & 1 & 1 \\ k & * & 1 & 1 \\ 1 & 1 & * & k \\ 1 & 1 & k & * \end{pmatrix}$

The Kimura two-parameter distance is given by:

$$K = -\frac{1}{2}\ln\left((1-2p-q)\sqrt{1-2p}\right)$$

Where, p is the proportion of sites which indicates the transitional differences and q is the proportion of sites that show transversional differences.

2.3.3 Tamura-Nei model

In this model by considering the differences in substitution rate between nucleotides and the inequality of nucleotide frequencies, (Tamura-Nei., 1993). Mainly it differentiates between transitional substitutions rates between purines and trans versional substitution rates between pyrimidines. It also assumes equality of substitution rates among nucleotide sites. The Tamura-Nei model is:

	A	T	C	G
A	-	βg_T	βg_C	βg_C
T	βg_A	-	$\alpha_2 g_C$	βg_G
C	βg_A	$\alpha_2 g_T$	-	βg_G
G	$\alpha_1 g_A$	βg_T	βg_C	-

Where,
β = Transitional frequencies between two sequences
α = Transversion frequencies between any two sequences
$d = -k_1 \ln(w_1) - k_2 \ln(w_2) - k_3 \ln(w_3)$
$s = -k_1 \ln(w_1) - k_2 \ln(w_2) - (k_3 - 2g_R g_y)\ln(w_3)$
$v = -2g_R g_y \, ln(w_3)$
$R = s/v$

P_1 and P_2 are the proportions of transitional differences between nucleotides between purines and pyrimidines *i.e.*, is between A to G and T to C respectively, Q is the Proportion of transversional differences, g_A, g_C, g_G, g_T, are the respective frequencies of A, C, G, and T.

where, $g_R = g_A + g_G$, $g_Y = g_T + g_C$, $k_1 = 2g_A g_G / g_R$, $k_2 = 2g_r g_c / g_Y$

$k_3 = 2\left(g_R g_Y - g_A g_G g_Y / g_R - g_r g_c g_R / g_Y\right)$,

$k_4 = 2\left(g_A g_G + g_r g_c + g_R g_Y\right)$

$W_1 = 1 - P_1/K_1 - Q/2g_R$, $W_2 = 1 - P_2/K_2 - Q/2g_Y$, $W_3 = 1 - Q/2g_Y g_R$

Var (d) = $[c_1^2P_1 + c_2^2P_2 + C_4^2Q - (C_1P_1 + C_2P_2 + C_4Q)^2]/L$

Var (s) = $[c_1^2P_1 + c_2^2P_2 + C_5^2Q - (C_1P_1 + C_2P_2 + C_5Q)^2]/L$

Var (v) = $[c_3^2Q(1-Q)]/L$

Var (R) = $[c_6^2P_1 + c_1^2P_2 + C_8^2Q - (C_6P_1 + C_7P_2 + C_8Q)^2]/L$
$c_1 = 1/w_1, \quad c_2 = 1/w_2, \quad c_3 = 1/w_3, \quad c_4 = k_1c_1/2g_R + k_2c_2/2g_Y + k_3c_3/2g_Rg_Y$
$c_5 = k_1c_1/2g_R + k_2c_2/2g_Y + k_3c_3/2g_Rg_Y - c_3, \; c_6 = c_1/v,$
$c_7 = c_2/v$
$c_8 = (c_5 = c_3R)/v$

2.3.4 P distance model
This distance model explains the number of sites at which the two associated sequences different. When the pairwise deletion option for handling gaps and missing data is used, then it is important to realize that this count does not normalize the number of differences based on the number of valid sites associated, only under the condition that if the sequences contain alignment gaps.
Therefore, we recommend that if one use of this distance then we go with the complete-deletion option.

Variances
Var (d) = $n_d(L - n_d)$/ L, Var (s) = $s(L - s)$/ L, Var (v) = $v(L - v)$/ L, R = s/v
Var (R) = $[c_1^2P + c_2^2Q - (C_1P + C_2Q)^2]/L$, $C_1 = 1/s$, $C_2 = -s/v^2$

Where,
d: Transitions + Transversions: Number of different nucleotide sites.
s: Transitions only: Number of nucleotide sites with transitional differences.
v: Transversions only: Number of nucleotide sites with transversional differences
R = *s*/*v*: Transition/transversions ratio.
L: No of valid common sites: Number of compared sites.
P and *Q* are the proportion of sites showing transitional and transversional differences, respectively.
Calculating F – measure through information retrieval approach: The information retrieval methods works on the two-way contingency table prepared form in the following form.

	Predicted class	
Actual class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

For classification works, the term true positives, true negatives, false positives and false negative compare results of the classifier under test with trusted external judgements. The term positives and negatives refer to the observation and true and false refer to the expectation. These can be explained by the given formula:

$$Precision = \frac{TP}{TP + FP} \times 100, \qquad Recall = \frac{TP}{TP + FN} \times 100$$

$$Accuracy = \frac{TP+TN}{TP+FP+FP+FN} \times 100$$

F-Measure (Probabilistic interpretation)
F- Measure is the harmonic mean between the precision and the recall, hence it can be interpreted by considering the two probabilities namely:
1. Precision is considered as the probability of a retrieved document is significant.
2. The recall is considered as the probability of relevant document which is retrieved.

$$F = 2 \times \frac{Precision \times recall}{precision + recall} \times 100$$

Using the given methods, a total of 4 phylogenetic trees were constructed and results are shows in results section. Also, the information retrieval methods are applied to find the best classification among the procedure employed.

3. Results and discussion
Constructing phylogenetic tree: The construction of the phylogenetic tree for the examination includes the acquisition of the sugarcane genome sequences from the NCBI dataset with the FASTA and MEGA format, and the phylogenetic tree constructed for the given algorithms was clarified under the following headings.

Table 2: Classification measures for various distance models.

S. No.	Models	True Positive	False Positive	True Negative	False Negative
1	MLE (Jukes Cantor)	20427	2769	64202	5698
2	MLE (Tamura Nei model)	79936	7145	1322	4693
3	MLE (Kimura 2 Parameter)	65161	5597	19468	2870
4	UPGMA (P distance model)	82092	969	2485	7550

Table 3: Classification measures for various distance models.

S. No.	Models	Accuracy (%)	Precision (%)	Recall (%)	F- Measure (%)
1	MLE (Jukes Cantor)	90.90	88.06	78.18	82.83
2	MLE (Tamura Nei model)	88.54	91.79	94.45	93.10
3	MLE (Kimura 2 Parameter)	90.90	92.08	95.78	93.89
4	UPGMA (P distance model)	90.84	98.83	91.57	95.06

The tree obtained for the aggregate 431 sugarcane genome sequencing of 11 distinctive *Saccharum* species for MLE with Tamura Nei distance measure has appeared in Fig 1. From Table 2, it was observed that they had been classified adversely with the 79936 true positives, 7145 false positives, 1322 true negative and 4693 false negatives. In view of this classification, the evolutionary measures are computed and from the Table 3, it was observed that the accuracy is 88.54 per cent, precision is 91.79 per cent, recall with 94.45 per cent and the harmonic mean between the precision and the recall that is F-measure is 93.10 per cent.

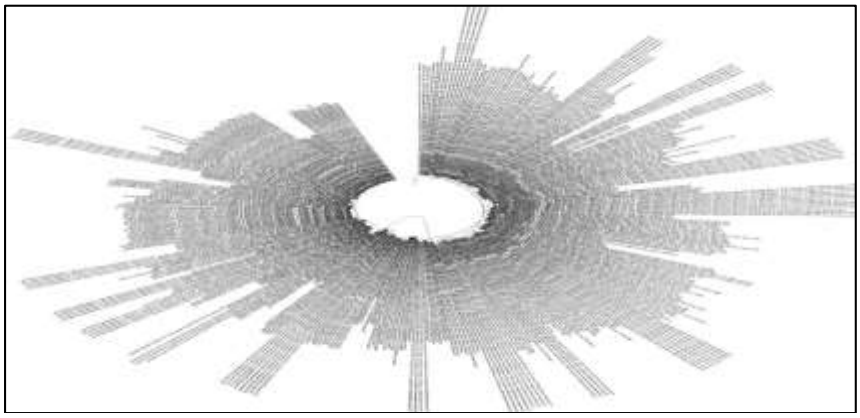

Fig 1: Phylogenetic tree of sugarcane genome sequences of Maximum likelihood algorithm with Tamura Nei model

	Predicted Class	
Actual Class	79936 (TP)	4693 (FN)
	7145 (FP)	1322 (TN)

Table 4: Confusion matrix for individual 11 *saccharum* species under the Maximum likelihood algorithm with Tamura Nei model

S. No.	Species	1	2	3	4	5	6	7	8	9	10	11	TOTAL
1	*Saccharum arundinaceum*	30	0	0	0	1	1	1	0	2	7	0	42
2	*Saccharum asper*	0	23	9	0	0	0	2	0	0	0	0	34
3	*Saccharum barberi*	0	0	38	2	0	0	0	0	0	0	0	40
4	*Saccharum edule*	1	0	0	39	0	0	0	0	0	0	0	40
5	*Saccharum officinarum*	2	0	0	0	37	0	0	0	0	0	0	39
6	*Saccharum procernum*	3	0	0	0	0	29	0	0	1	4	1	38
7	*Saccharum ravennae*	2	4	0	0	0	0	27	2	0	0	0	35
8	*Saccharum robustum*	0	0	0	0	0	0	0	32	4	0	0	36
9	*Saccharum sinese*	11	0	0	0	0	5	0	0	38	2	0	56
10	*Saccharum sponteneum*	2	1	0	0	0	2	2	0	4	21	0	32
11	*Saccharum villosum*	0	0	0	0	0	0	0	0	0	0	39	39
Total		51	28	47	41	38	37	32	36	47	34	40	431

Table 5: Comparing the F- measure of 11 species of sugarcane with different method

Species	Method			
	MLE (Jukes-Cantor)	MLE (Tamura Nei model)	MLE (Kimura 2 Parameter)	UPGMA (P distance model)
Saccharum arundinaceum	77.49	64.51	74.73	75.60
Saccharum asper	65.38	74.18	69.99	67.73
Saccharum barberi	65.62	87.35	85.36	85.40
Saccharum edule	60.97	96.29	96.30	95.00
Saccharum officinarum	57.57	96.09	93.33	86.10
Saccharum procernum	63.36	77.32	76.31	70.26
Saccharum ravennae	72.41	80.59	80.59	78.17
Saccharum robustum	59.37	94.44	91.42	64.28
Saccharum sinese	81.41	73.78	91.42	54.53
Saccharum sponteneum	56.00	63.63	68.84	63.63
Saccharum villosum	76.46	98.73	98.73	96.98

Table 6: Table showing the F-measure for classification of under *Saccharum villosum* various distance models

S. No.	Models	F-measure %
1	MLE (Jukes-Cantor)	72.89
2	MLE (Tamura Nei model)	98.73
3	MLE (Kimura 2 Parameter)	98.73
4	UPGMA (P distance model)	96.97

The present investigation it was inferred that the maximum likelihood with Kimura 2-parameter model and Tamura-Nei model measure gives the best from Table 5, and same outcomes for the different distance measures among every other strategy Maximum likelihood with Jukes-Cantor model give the minimum measure value with least accuracy. In the study of Confusion matrix that is 11×11 contingency table from Table 4 for the individual species under a different model in this model *Saccharum villosum* from Table 6 give the high F-measure value.

4. Conclusion

The examination demonstrated that maximum likelihood with Kimura 2-parameter model and Tamura-Nei model gives the highest results while Maximum likelihood with Jukes-Cantor model give the minimum. F measure will take the greatest qualities with ideal precision, recall, and accuracy esteems. But this computational biology of statistical outcomes demonstrates the best outcomes where we can compare functional relationship per centage between different models in which error per centage can also be reduced.

5. Future line of work

Phylogenetic trees have already witnessed applications in numerous practical domains, such as in conservation biology

(illegal whale hunting), epidemiology (predictive evolution), forensics (dental practice HIV transmission), gene function prediction and drug development. Other applications of phylogenies include multiple sequence alignment, protein structure prediction, gene and protein function prediction, and drug design.

6. References

1. Ababneh F. Models and Estimation for Phylogenetic Trees. Ph.D. Thesis, School of Mathematics and Statistics, Univ. Sydney 2006.
2. Aitken KS, Jackson PA, Mcintyre CL. Construction of a genetic linkage map for *Saccharum officinarum* incorporating both simplex and duplex markers to increase genome coverage. Genome 2007;50(8):742-756.
3. Billera LJ, Holmes S, Vogtmann, K. Geometry of the space of phylogenetic trees. Adv. Appl. Math 2001;27:733-767.
4. Bocci C. Topics on phylogenetic algebraic geometry. J. Comp. Biology, 1991;12(2):204-228.
5. DeBernardi JE. Penang: Rites of Belonging in a Malaysian Chinese Community. Singapore: NUS Press 2009.
6. D'Hont A, Glaszmann JC. Sugarcane genome analysis with molecular markers, a first decade of research. Proc. Int. Soc. Sugarcane Technol 2001;24:556-559.
7. Fisher SW, Lindberg MP. Phylogenetic inference using Genetic Algorithm based Least Squares Methods. M.Sc. Thesis, Aarhus Unv., Denmark 2006.
8. Garsmeur O, Droc G, Antonise R, Grimwood J, Potier B, Aitken K, *et al.* A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nat Commun 2018;9:2638.
9. Goldstein D, Mintz S. The Oxford Companion to Sugar and Sweets. Oxford: Oxford University Press 2015. doi: 10.1093/acref/9780199313396.001. 0001
10. Grivet L, Arruda P. Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr. Opin. Plant Biol 2002;5(2):122-127.
11. Hasan HO, Sayood K. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 2003;19(16):2122-2130.
12. Kaur H, Singh B. Classification and grading sugarcane using multi-class SVM. Int. J. Sci. Res. Publ 2013;3(4):1-5.
13. Kimura M. Simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol. Evol 1980;16(2):111-120.
14. Li W, Jerzy W, Jaromczyki. MSC Trees: a mean-shift based tool kit for cluster analysis of phylogenetic trees. Rep. UT-ORNL-KBRIN Bioinformatics, X Annual Summit, Memphis, USA 2011, 1-3.
15. Patil SS, Dhandra BV. An efficient classification of genomes based on classes and subclasses. Int. J. Comput. Sci. Eng 2010;2(5):1690-1695
16. Patil SS, Veneet Kumar, Vidya Pai, Ashoke R, Patil K. Constructing phylogenetic tree and analysis using information retrieval approach for MYB tfr's of sugarcane genome. IEEE 2015, 19-20.
17. Paterson AH, Moore PH, Tew TL. The gene pool of Saccharum species and their improvement," in Genomics of the Saccharinae, ed. A. H. Paterson (New York, NY: Springer) 2013, 43-71.
18. Saitou N, Tadashi I. Relative Efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution, and Neighbour-joining Methods of Phylogenetic Tree Construction in Obtaining the Correct Tree. Mol. Biol. Evol 1989;6(5):514-525.
19. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol. Biol. Evol 1992;9:678-687.
20. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbour-joining method. *Proc. Natl. Acad. Sci.*, USA 2004;101:11030-11035.
21. Zhu J, Zhou H, Pan Y, Lu X. Genetic variability among the chloroplast genomes of sugarcane (*Saccharum spp*.) and its wild progenitor species Saccharum spontaneum L. Genet. Mol. Res 2014;13:3037-3047.
22. Zhou D, Liu X, Gao S, Guo J, Su Y, Ling H, *et al*. Foreign cry1Ac gene integration and endogenous borer stressrelated genes synergistically improve insect resistance in sugarcane. BMC Plant Biol 2018;18:342.