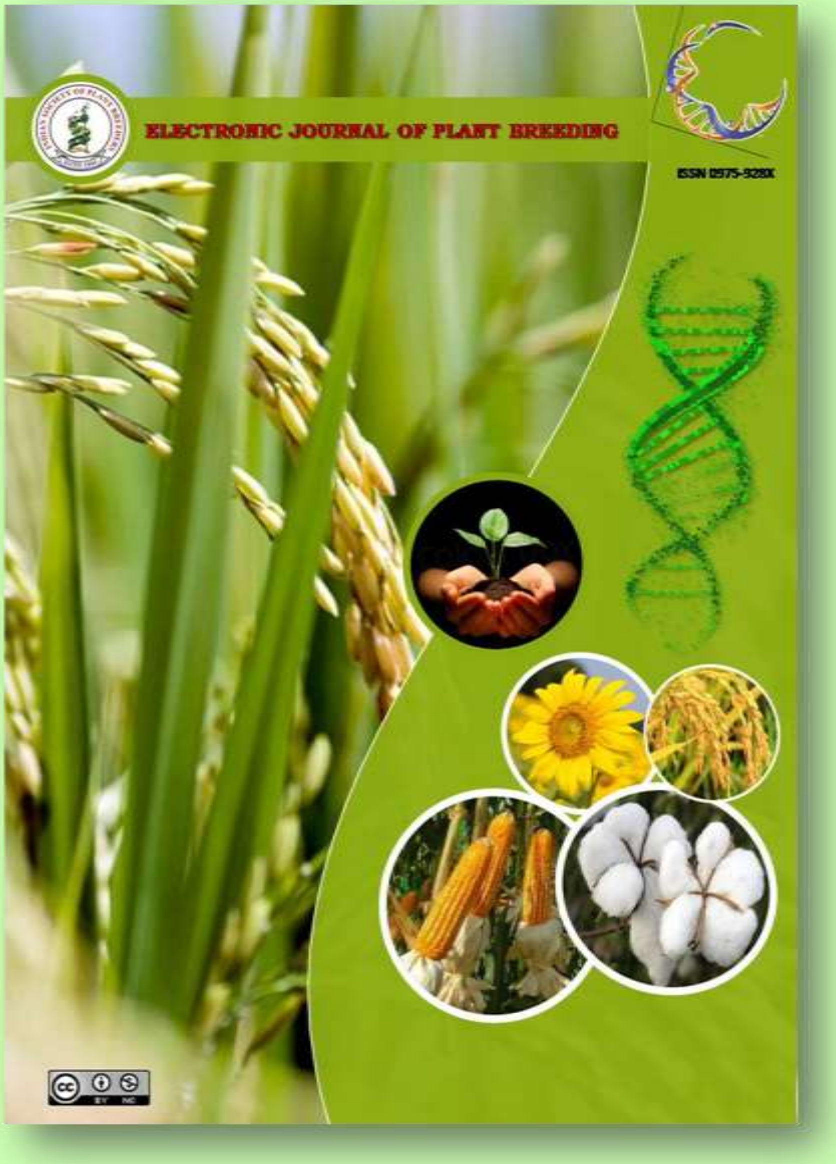
Genetic diversity analysis of sesame – A bayesian clustering approach

R. Nivedha, M. R. Duraisamy, Patil Santosh Ganapathi and S. Manonmani



ISSN: 0975-928X

Volume: 10

Number:2

EJPB (2019) 10(2):748-753

DOI:10.5958/0975-928X.2019.00098.X

https://ejplantbreeding.org

Research Article

Genetic diversity analysis of sesame – A bayesian clustering approach

R. Nivedha[1], M. R. Duraisamy[2*], Patil Santosh Ganapathi[3] and S. Manonmani[4]
[1]Agricultural Statistics, Tamil Nadu Agricultural University, Coimbatore.
[2]Professor (Mathematics), Tamil Nadu Agricultural University, Coimbatore.
[3]Assistant Professor (Agricultural Statistics), Tamil Nadu Agricultural University, Coimbatore.
[4]Professor (Plant Breeding and Genetics), Tamil Nadu Agricultural University, Coimbatore.
[*]**E-Mail**: mrd7@tnau.ac.in

(Received: 10 May 2019; Revised: 11 Jun 2019; Accepted: 11 Jun 2019)

Abstract
Diversity in plant genetic resources (PGR) provides opportunity for plant breeders to develop new and improved cultivars with desirable characteristics *viz*., high yield, pest and disease resistance, photosensitivity and high oil quality. Genetic diversity is a ubiquitous feature of all species in nature. Therefore, different genotypes of sesame were used for diversity analysis. Different clustering techniques were widely used for the analysis of diversity. In this paper, Bayesian hierarchical clustering algorithm is applied which can be interpreted as a novel fast bottom-up approximate inference method. Finally, this method clusters the genotypes into various groups with their corresponding genotypes in respective clusters.

Keywords
Sesame, Clustering, Bayesian hierarchical clustering, Diversity analysis, R software.

Introduction

Oilseed crops are the second most important determinant of agricultural economy, next only to cereals within the segment of field crops. The self-sufficiency in oilseeds achieved through “Yellow Revolution” during early 1990’s, could not be sustained beyond a short period. Despite being the fifth largest oilseed crop producing country in the world, India is also one of the largest importers of oilseeds (NFSM, 2015-16). The demand-supply gap in the edible oils has necessitated huge imports accounting for 60 per cent of the country’s requirement import of 14.01 million tonnes (DGCIS-Department of commerce, 2016-17).

Sesame is referred as “Queen of oilseeds” because of its stability due to rich antioxidants and less rancid properties. Sesamum is drought tolerant, can grow even where other crops fail (Ram *et al.*, 1990). This is due to its well-developed root system. The main constraints to sesame production are associated with its highly variable performance and low seed production. The problem of low seed production could be solved if heterosis in sesamum hybrids is utilised giving nearly 100-500% seed yield (Yol *et al.*, 2010). The seed contains 50 - 60% oil which has first-rate stability due to the presence of natural antioxidants such as sesamol, sesamin and sesamolin (Brar and Ahuja, 1980). Utilisation of sesamum germplasm is another alternative method for crop improvement. Genes for characters like higher yield, high oil quality, photosensitivity, pest and disease resistance *etc.*, may be identified from germplasm and introduced into cultivars. This is done with the help of genetic diversity analysis. Morphological, biochemical, and molecular markers can be used for crop-genetic diversity determination (Stuber *et al.*, 1992).

Clustering is an operation of multidimensional analysis that consists in partitioning a collection of objects. A *partition* is a division of a set (collection) into subsets, such that each object belongs to one and only one subset for that partition (Legendre and Rogers, 1972). One of the most important goals of unsupervised learning is to discover meaningful clusters in data. Clustering algorithms strive to discover groups, or clusters, of data points which belong together because they are in some way similar.

Despite the wide applicability of clustering methods, one should remember that no single family of methods can answer all questions raised in diversity analysis. Given a sufficiently large group of objects, clustering methods should be able to recognize clusters of similar objects while ignoring the few intermediates that often persist between clusters. Indeed, one cannot expect to find discontinuities when clustering sampling sites unless the physical environment is itself discontinuous, or unless sampling occurred at opposite ends of a gradient, instead of within the gradient (Whittaker, 1962).

Bayesian Hierarchical Clustering (BHC) algorithm which efficiently addresses many of the drawbacks of traditional hierarchical clustering algorithms. The goal of BHC is to construct a hierarchical representation of the data, incorporating both finer to coarser grained clusters, in such a way that we can also make predictions about new data points, compare different hierarchies in a principled manner, and automatically discover interesting levels of the hierarchy to examine. BHC can also be viewed as a fast way of performing approximate inference in a Dirichlet Process Mixture model (DPM), one of the cornerstones of nonparametric Bayesian Statistics. The objective of this study is to find the genetic diversity of sesame genotypes using Bayesian Hierarchical Clustering approach.

Materials and Methods

This study uses the evaluated data on 100 genotypes of sesame during 2018-19. This quantitative data was collected from Department of Oilseeds, TNAU, Coimbatore. The experiment was laid out in randomized complete block design with two replications for more than 100 genotypes and the following measures were observed on each genotype *viz*., Plant height (cm), number of primary branches per plant, days to 50 % flowering, days to 50 % maturity, capsules per plant and seed yield per plant (g) understand agronomic practices.

Heller and Ghahramani (2005) proposed the following Bayesian hierarchical clustering algorithm. Bayesian hierarchical clustering algorithm uses marginal likelihoods to decide which clusters to merge and to avoid over fitting. Manhattan distance, Euclidean distance measure etc., are all types of physical distance measures. For BHC, a probabilistic measure is used. This means that the distance is probability that data elements should be present in a particular cluster rather than a fixed distance determining clustering. This is important in determining on clusters iteratively merge within the hierarchical structure to form new clusters. Bayesian refers to the posterior probability calculation which is used to define the algorithms distance measure.

Let $D = \{x^{(1)}, \dots, x^{(n)}\}$ signify the entire data set of n elements. $D_i \subset D$ is the set of data elements belonging to cluster i, the set of data points at the leaves of subtree T_i. The algorithm is initialized with n trees, $T_i : i = 1 \dots . n$ each containing a single data point $D_i = \{x^{(i)}\}$. The algorithm proceeds by iteratively merging clusters. This is done by combining subtrees $T_i \; and \; T_j$ into a new tree T_k where the set of data elements in T_k is $D_k = D_i \cup D_j$.

For each merger, there are two hypotheses to be compared. The first, H_k^1 , is that all the data elements belonging to D_k were created independently and identically form the same probabilistic model $p(x|\theta)$, where θ are known as parameters. For example, the Normal Gaussian distribution, which has mean and covariance parameters $\theta = (\mu, \Sigma)$. For the type of data, the model selection needs to be appropriate. Next to model, there is the need for prior $p(\theta|\beta)$, where β are hyperparameters. we now have the ingredients to compute the probability of the data D_k under H_1^k:

$$p(D_k|H_1^k) = \int p(D_k|\theta)\, p(\theta|\beta) d\theta \qquad (1)$$

The alternate hypothesis, would be that the data in D_k has tow or more clusters in it. H_2^k is simply a product over the subtrees

$$p(D_k|H_2^k) = \int p(D_i|T_i)\, p(D_j|T_j) \qquad (2)$$

Combining the probability of the data under hypotheses $H_1^k \; and \; H_2^k$ weighted by the prior that all points in D_k belong to one cluster, $\pi_k \overset{\text{def}}{=} p(H_1^k)$, we can obtain the marginal probability of the data in tree T_k:

$$p(D_k|T_k) = \pi_k\, p(D_k|H_1^k) + (1 - \pi_k)\, p(D_i|T_i)p(D_j|T_j)$$
(3)

Let the posterior probability of the merger hypothesis, H_1^k, be denoted by r_k. The Bayes Theorem is given by

$$p(B|A) = \frac{p(A)p(B|A)}{p(B)} \qquad (4)$$

Applying Bayes Theorem Eq. (4) gives the posterior probability for the merged hypothesis

$$r_k = \frac{\pi_k\, p(D_k|H_1^k)}{p(D_k|T_k)} \qquad (5)$$

A probability $r_k \geq 0.5$ mentions that a merger between clusters is more likely than not.

The general algorithm

input: data $D = \{x^{(1)}, \dots, x^{(n)}\}$,model $p(x|\theta)$, prior $p(\theta|\beta)$

```
initialize: number of clusters $c = n$, and $D_i = \{x^{(i)}\}$ for $i = 1 \dots . n$
while $c > 1$ do
    Find the pair $D_i$ and $D_j$ with the highest probability of the merged hypothesis:
        $r_k = \frac{\pi_k p(D_k|H_1^k)}{p(D_k|T_k)}$
  Merge $D_k \leftarrow D_i \cup D_j$ ,$T_k \leftarrow (T_i,T_j)$
  Delete $D_i$ and $D_j$, $c - 1$
end while
output: Bayesian mixture model where each tree node is a mixture component
```

The tree can be cut at points where $r_k < 0.5$

The coding used in R programming language to find the clusters:

```
R > setwd ("D:/") ## Set the directory to import the data
R > ## Or mention the file path of the dataset
R > data<-read.csv ("data.csv",header = TRUE)
R > head(data)
R>## result is given in Table .1
R > x <-scale (data [, -c (1,2)])
R > is.matrix (x)  # check whether the dataset is in matrix form   [1] TRUE
R > x.id<-c (rep (1:100, each = 1), rep (1:100, each = 1))
R > meansumsq <- meancss (x, x.id)
R > # define a function to estimate the model hyperparameters
R > optimfunc <- function (phi) {
 +    -loglikelihood (x.mean = meansumsq$mean, x.css = meansumsq$css,
 +    repno = meansumsq$repno, transformed.par = phi, var.select = FALSE) }
R > # estimate the hyperparameters and save them into xinit.tpar
R > xinit.tpar <- optim(rep(0, 5), optimfunc, method = "BFGS")$par
R > #define another function to estimate the hyperparameters of the variable selection model
R > optimfunc<-function (phi) {
 +    -loglikelihood (x.mean = meansumsq$mean, x.css = meansumsq$css,
 +    repno = meansumsq$repno, transformed.par = c (xinit.tpar [1:4], phi)) }
R > # write the hyperparameters into x.tpar
R > x.tpar <- c (xinit.tpar [1:4], optim (rep(0, 2), optimfunc, method = "BFGS") $par)
R > bclust.obj <- bclust (x, rep.id = x.id, transformed.par = x.tpar)
```

Results and Discussion

The bclust package (Nia and Davison, 2012) available in R software implements the algorithm presented. This uses marginal distributions to estimate the model hyper parameters using empirical Bayes, thereby produces a fully automatic method. bclust package is applied to the sesame dataset on 100 genotypes having 2 replications with 6 variables. From the results of the dendrogram (Fig.1.) the Bayesian clustering method clusters the 100 genotypes into 5 groups with 11, 28, 33, 18, 10 genotypes in respective clusters and specify the days to 50 % flowering as the important variable.

The Bayesian hierarchical clustering algorithm has several advantages. (Lowing *et al.*, 2017) (1) The algorithm can be interpreted as a novel fast bottom-up approximate inference method for a Dirichlet process (2) It describes a probabilistic model of the data which is used to compute the predictive distribution of a test point and the probability of being in any of the existing clusters in the tree (3) Bayesian hypothesis testing is used to decide which merges are advantageous and to output the recommended depth of the tree.

References

Brar, GS, and KL Ahuja. 1980. "Sesame: its culture, genetics, breeding and biochemistry." *Annual reviews of plant sciences*.

Heller, Katherine A, and Zoubin Ghahramani. 2005. "Bayesian hierarchical clustering." Proceedings of the 22nd international conference on Machine learning.

https://commerce.gov.in/writereaddata/uploadedfile/MOC_636281140249481285_annual_report_16_17_eng.pdf

https://www.nfsm.gov.in/StatusPaper/NMOOP2018.pdf.

Legendre, Pierre, and David J Rogers. 1972. "Characters and clustering in taxonomy: a synthesis of two taximetric procedures." *Taxon*:567-606.

Lowing, Nicholas, Ryan Bomalaski, and Debasis Mitra. 2017. "Bayesian Hierarchical Clustering."

Nia, Vahid Partovi, and Anthony C Davison. 2012. "High-Dimensional Bayesian Clustering with Variable Selection: The R Package bclus." *Journal of Statistical Software* 47 (ARTICLE):1-22.

Ram, Raghav, David Catlin, Juan Romero, and Craig Cowley. 1990. "Sesame: New approaches for crop improvement." *Advances in new crops. Timber Press, Portland*:225-228.

Stuber, Charles W, Stephen E Lincoln, DW Wolff, T Helentjaris, and ES Lander. 1992. "Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers." *Genetics* **132** (3):823-839.

Whittaker, Robert H. 1962. "Classification of natural communities." *The Botanical Review* 28 (1):1-239.

Yol, Engin, Emre Karaman, Seymus Furat, and Bulent Uzun. 2010. "Assessment of selection criteria in sesame by using correlation coefficients, path and factor analyses." *Australian Journal of Crop Science* **4** (8):598.

Table 1. Sample of data set

Genotype	Replication	Days to 50 % Flowering	Days to Maturity	Plant Height	Primary Branches per plant	Capsule per plant	Yield per plant
1	1	60	105	140	6	116	2.3
2	1	47	97	110	6	152	1.87
3	1	47	99	99	8	199	1.02
4	1	46	103	95	5	62	5.76
5	1	47	103	100	6	45	6.03

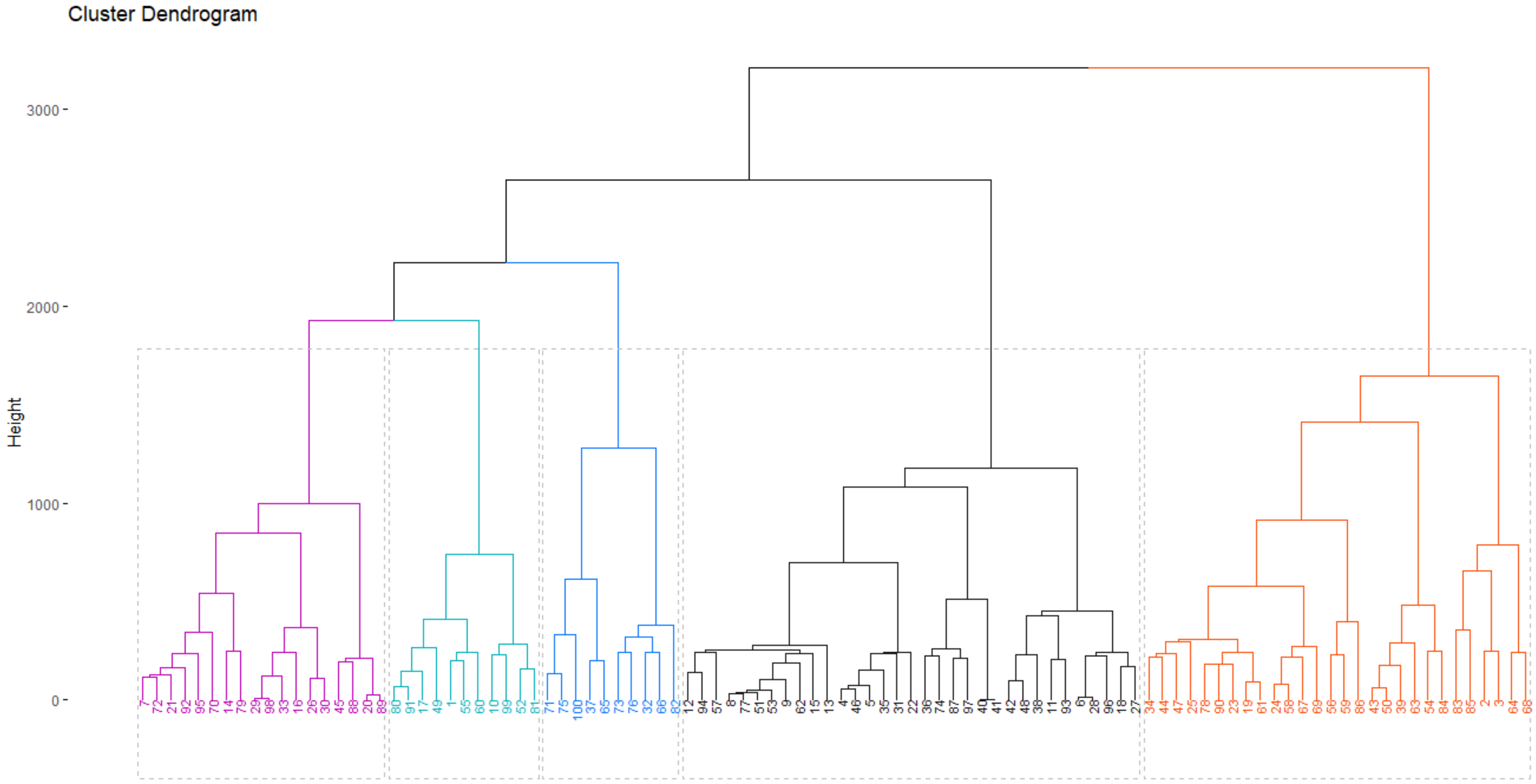


Fig. 1. Cluster dendrogram of different genotypes

INDIAN SOCIETY OF PLANT BREEDERS

ESTD 1995

https://ejplantbreeding.org