AGRO- MORPHOLOGICAL CHARACTERIZATION AND RELATIONSHIPS AMOUNG MUSTARD GERMPLASM (*Brassica juncea* [L.] Czern & Coss) IN SRI LANKA: A CLASSIFICATION TREE APPROACH

S. R Weerakoon[1] and S. Somaratne [2]

ABSTRACT

An RT model (RT1) was constructed using 35 agro-morphological characters for 45 mustard (Brassica juncea) accessions. Based on the 'variable importance' of the model RT1, another model (RT2) was developed. These models were developed using classification and regression tree algorithms. The classificatory performance of the RT1 model was compared with RT2 model. RT1 and RT2 models classified the mustard accessions with misclassification rates of 2.3% (98% accuracy) and 4.3% (96% accuracy), respectively. The variable importance of RT1 and RT2 explained that leaf length (LLCM), hypocotyl length (HLCM), hypocotyl-anthocynin coloration (ACH) and leaf width (LWCM) at seedling stage and main inflorescence length (LMICM), silique length (SLMM) and seed yield/plant (SYDIVPG) at maturity stage play an important role in classifying mustard accessions. Comparison of RT1 with RT2 revealed that accuracy of classification made by RT1 is higher in predicting class memberships among mustard accessions. A large degree of variability within and between Sri Lankan mustard accessions has been observed for agro-morphological characters with respect to LLCM, HLCM, ACH, LWCM, LMICM, SLMM and SYDIVPG. The genetic diversity of certain mustard accessions such as Accession Numbers 346, 8658 and 9726 is too high and RT models failed to classify them correctly with acceptable accuracy.

Key words: *Agro-morphological characters, Brassica juncea, Classification Tree Analysis, germplasm*

INTRODUCTION

In Sri Lanka mustard (*Brassica juncea* [L.] Czern & Coss) is used as a condiment in cooking and also as a cooking-oil. In the Indian sub-continent it is an important oilseed crop. Sri Lanka has over sixty mustard accessions in the gene bank of the Plant Genetic Resources Centre (PGRC), Gannoruwa (Plant Genetic Resources Catalogue, 1999). However, a very limited studies have been carried our on genetic diversity and the relationships among these accessions. Estimates of genetic diversity and the relationships between germplasm collections are very important to identify genetically diverse, agronomically superior accessions for the improvement of mustard as an oilseed crop in Sri Lanka. It also enables gene banks to carry out efficient collection and to unambiguously characterize the accessions to avoid confusions arising out of duplications and mishandling.

Many tools are now available for studying variability and the relationships among accessions including total seed protein, isosymes and various types of molecular markers. However,

[1]*Department of Botany, Faculty of Natural Sciences, The Open University of Sri Lanka, Nawala.*

morphological characterization is the first approach in the description and classification of germplasms (Smith and Smith, 1989). There are number of numerical taxonomic analytical methods available for classifying and recognizing the patterns of phenotypic diversity and the relationships between the species and germplasm collections of a variety of crop (Gupta *et al*., 1991; Dias *et al*.,1993; Amurrio *et al*., 1995; Li *et al*., 1995). Since there is very limited studies and records are available, an in depth study is required on the genetic diversity and the relationships among the germplasm collections of local mustard accessions for efficient germplasm manipulation and management.

Preliminary studies conducted on classification of mustard accessions were primarily based on the agro-morphological characters and multivariate statistical analyses (Cluster Analysis (CA), Principle Component Analysis (PCA) and Discriminant Function Analysis (DFA)). These studies revealed that there were uncertainties in classification of local mustard accessions (Weerakoon *et al*., 2005; Weerakoon *et al*., 2007). CA, PCA and DFA indicated that there is a discernible difference in the grouping patterns of local mustard accessions. In CA mustard accessions were fallen within five groups. However, PCA results well separated only six accessions from the rest which is different from results of CA. DFA classified mustard accessions into three groups. Thus grouping patterns of mustard accessions were different under different statistical analytical methods (Weerakoon *et al*., 2007). Though these statistical procedures have long been widely used for classification in various fields of studies, over simplification, ignorance of complex nonlinear interactions etc., are the limitations in accurately classifying the elements of a particular group in concern. Therefore, novel mathematical modeling approaches are required to be employed in the classification and characterization of local mustard germplasm in Sri Lanka.

Thus, the core objectives of the present study were to;

a) Find the applicability of Classification Tree (CT) modeling to study the diversity of the mustard germplam in Sri Lanka

b)Explore the agro-morphological characters and their importance in classification of local mustard accessions.

MATERIALS AND METHODS

Morphological characterization

A total of 45 mustard accessions obtained from PGRC were used in the study. Five (5) seedlings of each accession were planted in plastic trays with standard potting mixture in a green house at the Open University, Nawala. Subsequently, the seedlings were (3-4 leaf stage) transferred to plastic pots of 13 cm in diameter and they were arranged in Complete Randomized Block Design (RCBD). Characterization of accessions was made on different morphological traits observed (Figure 1) from seedling-to-harvesting stage of the crop (Table1). The traits selection and measurement were made according to the International Board for Plant Genetic Resources (IBPGR, 1991), Descriptors for *Brassica* and *Raphanus*: Morphological Descriptors for mustard (MAFF, 1993) and Gupta *et al*., (1991). A total of 35 agro-morphological characters were recorded for each mustard accession, as shown in Table 1.

Table 1

Table 01: Agro-morphological traits used for characterization of local mustard accessions (Rabbani *et al*., 1998) and

Trait designation A. Seedling stage	Code	Trait designation B. Flowering stage	Code	Trait designation C. Maturity stage	Code
Cotyledon Petiole length (cm)	CPLCM	Days to bolting initiation	DBI	Days to maturity	DM
Cotyledon length (cm)	CLCM	Days to first flowering	DFF	Number of leaves /plant	NLP
Cotyledon width (cm)	CWCM			Plant height (cm)	PHCM
Cotyledon width / length ratio	CWDIVCL	Days from bolting to first Flowering	DFB	Number of primary branches/plant	NPB
Hypocotyl length (cm)	HLCM			Length of main inflorescence (cm)	LMICM
Anthocyanin coloration of hypocotyls	ACH	Days from first to last t Flowering	DFFF	Siliques /main inflorescence	SMI
Leaf blade shape	LBS	Leaf petiole length (cm)	LPLCM	Silique length (mm)	SLMM
Location of leaf margins	LLM			Silique width (mm)	SWMM
Number of leaflets	NL	Leaf length (cm)	LLCM	Silique length / width ratio (mm)	SLDIVSW
Number of serrates	NS	Leaf width (cm)	LWCM	Number of seeds /Silique	NSDIVS
Leaf petiole length (cm)	LPLCM			1000-seed weight (g)	SW1000G
Leaf length (cm)	LLCM	Leaf length / width ratio	LLDIVLW	Seed colour	SC
Leaf width (cm)	LWCM			Seed yield/plant (g)	SYDIVEPG
Leaflength / width ratio	LLDIVLW				

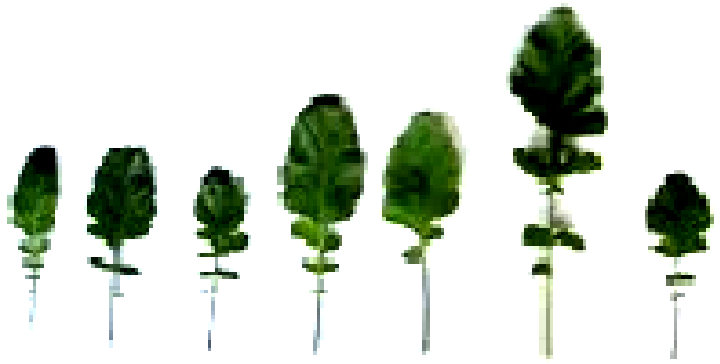

Figure 01: Leaf morphology of different mustard accessions (at seedling stage)

Statistical Analysis

Classification and Regression Trees (CART) was used to construct classification trees for the dataset. In CART modeling, all partitions resulted by all variables are compared with the reduction in heterogeneity or impurity that they provide (Briman *et al*., 1984). In Classification trees, impurity is measured by computing Gini index of diversity. The Gini index of diversity of a node with *n* objects and *c* possible classes is defined as ;

$$Gini = 1 - \sum_{j=i}^{c} \left(\frac{n_j}{n} \right)^2$$

where n_j is the number of object from class *j* in the node.

The selection of correct classification tree was obtained by comparing the cost-complexity parameter which was calculated as follows:

$$R_\alpha = R(T) + \alpha|T|$$

where, |T| is the complexity of the sub-tree (number of terminal nodes), α is the complexity parameter, and R(T) is the re-substitution error (overall misclassification rate). The optimal tree size is determined by cross-validation procedure in which the dataset was randomly divided into 10 subsets. One of the subsets was then used as independent test set while the rest of the subsets were combined and used as training dataset. The tree growing and pruning procedure was repeated 10 times, each time with a different subset as test set. For each size of tree, the prediction error (misclassification rate) was calculated and averaged over all subsets. The prediction error obtained for each sub-tree on the cross validation was matched with the sub-tree of the complete dataset using the α values. The variables in the models were explored by the "variable importance ranking" available in the CART (Salford Systems, 2000). Two models were developed from the dataset. The first model (RT1) used 35 agro-morphological characters. The second model (RT2) constructed after observing the "variable importance" of the RT1 and included 13 agro-morphological characters.

RESULTS AND DESCUSSION

The first classification tree (RT1) showed a minimum cross validated relative error, re-substitution error, and complexity of 0.077 ± 0.012, 0.028, and 0.027 respectively (Table 2). The variable importance of the model RT1 shown in Table 3 indicates that agro-morphological character from LLCM to DM has value above 50%. The rest of the variables indicated less value for the variable importance. The second model (RT2) constructed from the variables chosen from the variable importance of the first model showed minimum cross validated relative error, re-substitution error and complexity of 0.056 ± 0.009, 0.028 and 0.027 respectively (Table 2). Both models indicated more or less similar values for the minimum cross validated relative error, re-substitution error, and complexity. The both models included 29 terminal nodes and more or less similar complexity values and the errors. However, the number of agro-morphological characters used was different, 35 in model RT1 and 13 in RT2.

The variable importance values of RT1 and RT2 are given in Table 3 indicated that RT1 has primary splitters such as LLCM – DM with scores higher than 50%. Meanwhile RT2 showed that variables such as LLCM –LLFCM with score higher than 50%. Comparison of variable importance scores of RT1 and RT2 revealed that LLCM has greater predictive power in classification of local mustard accession.

Table 02: Number of trees developed for RT Model 1 and RT Model 2 with their number of terminal nodes, cross-validated relative errors, re-substitution errors, and complexities. ** - Optimal cost, * - minimum cost.

Tree number / Model Number	Number of terminal nodes	Cross-validated relative error	Re-substitution error	Complexity
***RT* 1**				
1**	30	0.052 ± 0.013	0.000	-1.000
2*	29	0.077 ± 0.012	0.028	0.027
3	28	0.098 ± 0.011	0.059	0.030
4	1	$1.000 \pm 8.93\ 10^{-5}$	1.000	0.034
RT 2				
1**	30	0.028 ± 0.009	0.000	-1.000
2*	29	0.056 ± 0.009	0.028	0.027
3	28	0.087 ± 0.009	0.059	0.030
4	1	$1.000 \pm 8.93\ 10^{-5}$	1.000	0.034

Table 03: Variable importance of RT model 1 and RT model 2.

RT model 1		RT model 2	
Variable	**Score (%)**	**Variable**	**Score (%)**
LLCM	100.00	LLCM	100.00
HLCM	84.25	LWCM	87.50
LMICM	77.37	LMICM	84.97
ACH	77.14	HLCM	79.58
LWCM	77.12	SLMM	76.26
LPLCM	68.20	SYDIVPG	71.43
CWCM	61.47	DM	63.74
SYDIVPG	60.91	LPLCM	59.36
LLFCM	55.49	CPLCM	58.40
LBS	55.39	ACH	56.22
SLMM	54.34	CWCM	52.96
CPLCM	52.14	LLFCM	51.49
DM	52.11	LBS	48.10

The accession numbers those classified successfully (100%) by the models are not given the Table 4. and accession numbers with certain amount of misclassification rates were given in the Table 4. The majority of the accession numbers listed in the table are correctly classified at the rate of 90% and accession umbers 346 and 9726 were classified with a misclassification of 30%. Further, accession number 8658 also misclassified with a rate of 20%. In addition, from comparison of correctly classified and misclassified accession numbers in Table 4, it is clear that certain accession numbers are overlapping each other with respect to the agro-morphological characters used in the characterization.

As far as the total number of mustard accession numbers is concerned, 73% and 26% of the accession numbers were correctly classified by the RT1 model at the rates of 100% and 90%, respectively. Alternatively, RT2 model, 56%, 36%, 3% and 6% of the accession numbers were classified at the rates of 100%, 90%, 80% and 70%, respectively. As a whole both models correctly classified the accession numbers with the rate of 93%.

Table 04: Classificatory performance of RT model 1 and RT model 2.

Accession No. / Model No.	Correctly classified (%)	Misclassified (%)	Misclassified into Acc. No.
RT1			
1353	90	10	7781
2180	90	10	1256
501	90	10	508
5181	90	10	9725
721	90	10	5088
7700	90	10	7789
8658	90	10	1847
9725	90	10	5181
RT 2			
1353	90	10	2180
1814	90	10	346
346	70	30	8852
501	90	10	508
5181	90	10	9725
721	90	10	5088
747	90	10	2310
7700	90	10	1381
7792	90	10	7814
8658	80	10	7789
8831	90	10	1256
9725	90	10	5181
9726	70	30	8852

Genetic erosion and habitat destruction by modern agricultural practices have increased the importance of germplasm characterization of plant materials. In order to ensure the efficient and effective utilization of crop germplasm, its characterization is imperative.

Previous preliminary studies conducted to assess the genetic divergence of local mustard (*Brassica juncea* [L.] Czern & Coss) genotypes in Sri Lanka with 30 mustard accessions using numerical analyses of 35 agro-morphological characters revealed that there were ambiguities in classification of mustard accessions (Weerakoon *et al.*, 2005: Weerakoon *et al.*, 2007). The present results revealed that classification tree models developed to predict the memberships of the mustard accession numbers are much accurate. The comparison of the models revealed that, the number of agro-morphological characters could be reduced without a considerable impact on the model

predictive performances. The RT1 and RT2 models classified the mustard accessions with misclassification rates of 7% (93% accuracy) and 10% (90% accuracy), respectively. The variable importance (> 60%) of RT2 indicated that the length of leaf (100%), leaf width (88%), hypocotyl length (80%) at seedling stage and length of main inflorescence (85%), silique length (76%) and days to maturity (64%), at maturity stage were important in classifying local mustard accessions. However, comparison of performance of RT1 model with RT2 model indicated that accuracy of classification made by RT1 is higher than that of RT2 in predicting class memberships among mustard accessions. It has been suggested that there are 35 agro-morphological characters for classification of mustard accession (Rabani *et al*., 1998). However, the models developed in this study indicated that fewer agro-morphological characters are sufficient in successful classification of mustard accessions with minimum cost, labor, and time. Further, it is recommended that use of Multivariate Regression Tree (MRT) analysis to improve the performance of RT models.

A large degree of variability within and between Sri Lankan mustard accessions has been observed for agro-morphological characters with respect to LLCM, HLCM, ACH, LWCM, SLMM and SYDIVPG. In certain mustard accessions such as 346, 8658 and 9726 the genetic diversity is too high that the RT models also failed to classify them correctly with acceptable accuracy. This reveals that agro-morphological traits such as LLCM, HLCM, ACH, LWCM, SLMM and SYDIVPG are also in limited value in characterization of certain Sri Lankan mustard accessions. This may be due to the fact that these accessions would have reflected the ecological provenance of the mustard accessions. The diversity as indicated by RT models using agro-morphological traits in Sri Lankan mustards agrees with the finding of the previous studies carried out by Weerakoon *et al*. (2005) and Weerakoon *et al*. (2007) and it provides opportunities for selection and breeding.

CONCLUSIONS

The present study revealed that on the basis of agro-morphological traits, there is a wide genetic variation in Sri Lankan mustard accessions which are stored in the gene banks at PGRC. Although 77% of the Sri Lankan mustard accessions are characterized by the agro-morphological traits such as LLCM, HLCM, ACH, LWCM, SLMM and SYDIVPG, there is a limitation in characterization of mustard accessions (*ca.* 23%). It is suggested that further studies should be carried out on the agro-morphological characterization and/or molecular characterization of mustard accessions in order to trace this genetic diversity. In addition, it is worthy to study mustard accessions with their agro-ecological zonal origins, thus the effect of genotype *versus* environment interaction in expression of the phenotypes could be ascertained.

ACKNOWLEDGEMENT

The authors are thankful to the gene bank of the Plant Genetic Resources Centre (PGRC), Gannoruwa for providing mustard accessions used in this study.

REFERENCES

Amurrio, J. M., Ron de A. A. and A. C. Zeven (1995). Numerical taxonomy of Iberian pea landraces based on quantitative and qualitative characters. Euphytica, 82, pp. 195-205.

Breiman, L., Friedman, J., Olshen, R., and C. Stone. (1984). Classification and Regression Trees. Wadsworth & Brooks, Pacic Grove, CA.

Dias, J. S., Monteiro, A. A. and M. B. Lima (1993). Numerical taxonomy of Portugese Tronchuda cabbage and Galega landraces using morphological characters. Euphytica, 69, pp. 51-68.

Gupta, V. P., Sekhon, M. S. And D. R. Satija (1991). Studies on genetic diversity, heterosis and combining ability in Indian mustard (*Brassica juncea* [L.] Czern & Coss). Indian Journal of Genetics, 51, pp. 448-453.

IBPGR (1990). Descriptors for *Brassica* and *Raphanus*. International Board for Plant Genetic Resources, Rome, pp. 51.

Li, Y., Wu, S. and Y. Cao (1995). Cluster analysis of an International collection of foxtail millet (*Setaria italica* (L.) P. Beauv.). Euphytica, 83, pp. 79-85.

MAFF (1993). Morphological descriptors for mustard (*Brassica juncea*). Ministry of Agriculture, Forestry and Fisheries, Government of Japan.

Plant Genetic Resources Catalogue Passport Information (1999). Plant Genetic Resources Centre, Gannoruwa, Sri Lanka.

Rabbani, M. A., Iwabuchi, A., Murakami, Y. and T. Suzuki (1998). Phenotypic variation and the relationships among mustard (*Brassica juncea* L.) germplasm from Pakistan. Euphytica, 101, pp. 357-366.

Salford Systems. (2000). CART for Windows, Version 4.0, San Diego, CA.

Smith, J. S. C. and O. S. Smith (1989). The description and assessment of distance between lines of maiza. II. The utility of morphological, biochemical, and genetic descriptors and a scheme for the testing of distinctiveness between inbred lines. Maydica, 34, pp. 151-161.

Weerakoon, S. R., Somaratne, S., Wimalasooriya, R. and M. C. M. Iqbal. (2005). Phenotypic variation and the relationships among mustard (*Brassica juncea* L.) germplasm from Sri Lanka. OUSL Silver Jubilee Academic Sessions, pp. 67-73.

Weerakoon, S. R., Iqbal, M. C. M., Somaratne, S., Peiris, P. K. D. and W. S. R. Wimalasuriya (2007). Delimitation of local mustard (Brassica juncea) germplasm in Sri Lanka and improvement of their nutritive quality. 12th International Rapeseed Congress, Wuhan, China, 5, pp. 75-78.