Brazilian Journal of Poultry Science
Revista Brasileira de Ciência Avícola

ISSN 1516-635X Apr - Jun 2005 / v.7 / n.2 / 79 - 84

Evaluation of Genetic Divergence among Lines of Laying Hens using Cluster Analysis

■ Author(s)

Barbosa L[1]
Regazzi AJ[2]
Lopes PS[3]
Breda FC[1]
Sarmento JLR[1]
Torres RA[3]
Torres Filho RA[4]

[1] Graduate Student - Genética e Melhoramento - UFV
[2] Professor - Departamento de Informática - UFV
[3] Professor - Departamento de Zootecnia - UFV
[4] Geneticist - Globoaves

■ Mail Address

Leandro Barbosa
Rua da Conceição, 171
Bom Jesus
36.570-000. Viçosa, MG
Phone: +55 +31 3891-7316

E-mail: leandromoco@bol.com.br

■ Keywords

Laying hens, Mahalanobis distance, single linkage, Tocher's optimization method

■ Acknowledgements

The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the scholarships granted to JRL Sarmento, FC Breda and L Barbosa.

Arrived: january / 2005
Approved: may / 2005

ABSTRACT

Cluster analysis was used to investigate the genetic divergence among five lines of laying hens. The following traits were evaluated: body weight at 40, 48 and 56 weeks of age; egg weight at 40, 44, 52 and 60 weeks of age; and laying rate from 40 to 60 weeks of age. Three groups were formed when data were analyzed by the single-linkage hierarchical method using squared Mahalanobis distance (D^2) as dissimilarity measures: one group comprised lines 3 and 5, the second group line 1, and the third group comprised lines 2 and 4. Using Tocher's optimization method, only two groups were formed: one group comprised lines 3, 5 and 1, and the second comprised lines 2 and 4. This evidences the disagreement between the methods over the evaluation of genetic divergence. The trait that contributed mostly to the genetic divergence was body weight at 48 weeks of age.

INTRODUCTION

Poultry genetic improvement programs are based on the genetic variability of individuals, which may be changed by introducing new genotypes in the flocks. Crosses are also frequently used in these programs, and the success of this practice is dependent on the genetic divergence between the parents. Parents showing high productivity indexes and great genetic diversity may produce a progeny that is more productive and shows great genetic variability (Piassi *et al.*, 1995b).

Studies on the genetic diversity of plants and livestock have increased after the 70's, simultaneously to the sudden increase in the availability of informatics resources (Sakaguti *et al.*, 1996).

Genetic divergence studies may be used to evaluate the behavior of genotypes in different environments, to evaluate the superiority of some genotypes over others, to identify divergent genotypes that may be used as parents in breeding programs and to relate genetic divergence with heterosis (Piassi, 1994).

Techniques of multivariate analysis such as clustering analysis have been successfully employed as a means to identify divergent genotypes and better utilize the advantages provided by heterosis. The objective of this model is to join the parents (or any other kind of sample objects) in many clusters based on a classification criterion, so that there is within-group homogeneity and between-group heterogeneity. Alternatively, techniques of clustering analysis divide a group of objects into many subgroups, according to a criterion of similarity or dissimilarity. Such process involves basically two steps. The first one is related to estimating a measure of similarity (or dissimilarity) between parents, whereas the second is related to clustering the objects into subgroups based on a joining technique (Cruz *et al.*, 2004).

Studies on genetic divergence have used the mean Euclidian distance or the standardized squared Mahalanobis distance as dissimilarity measures (Cruz *et al.*, 2004). Although the latter has been preferred, it can be estimated only when the residual covariance matrix is available. Therefore, the advantage of the standardized squared Mahalanobis distance (D^2) over the mean Euclidian distance is that it considers the correlation among the chosen traits.

The second step of the clustering analysis consists in choosing a joining method. The most used methods in studies of genetic diversity are the hierarchical and optimization methods. The Tocher's method is the most used among the optimization methods (Piassi, 1994).

In the present study, the objective was to evaluate the genetic divergence among five lines of laying hens by clustering analysis, using both hierarchical and optimization methods.

MATERIAL AND METHODS

Data used herein were collected from August 1997 to August 1999. Five lines of laying hens (Leghorn) were used and had been developed at Universidade Federal de Viçosa, MG, Brazil. The lines were obtained by artificial insemination in an open house with cages. The evaluated traits were body weight at 40 weeks (BW40), 48 weeks (BW48) and 56 weeks of age (BW56); egg weight at 40 (EW40), 44 (EW44), 52 (EW52) and 60 weeks of age (EW60); and laying rate from 40 to 60 weeks of age (LP). Two hundred and fifty birds were evaluated (50 per line).

Clustering analyses were conducted considering the standardized Mahalanobis D^2 distance as dissimilarity measure and, as joining methods, the nearest-neighbour hierarchical method and the Tocher's optimization method.

The standardized squared Mahalanobis distance between the parents i and i' is calculated by the formula $D_{ii'} = (\bar{X}_i - \bar{X}_{i'})R^{-1}(\bar{X}_i - \bar{X}_{i'})$, where R is the residual covariance matrix, and $\bar{X}_i - \bar{X}_{i'}$ are p-dimensional vectors of the means of the parents i and i', respectively (Mahalanobis, 1936; Cruz *et al.*, 2004).

In the hierarchical method of the nearest-neighbour, the parents (laying hen lines) were grouped based on the closest D^2 distances using a procedure that is repeated at different levels until a dendogram or a tree diagram is established. In this case, optimum numbers of groups are not important, since the interest lies on the resulting tree and its branches. Delimitations of the dendogram were visually established, so that the points of high-level changes were identified and considered as the limits of the number of parents for a single group (Cruz & Carneiro, 2003).

In the hierarchical method, it is important to evaluate the adjustment between the original distance coefficient matrix (phenetic matrix ? ph) and the matrix that results from the clustering process (cophenetic matrix - c). Thus, the cophenetic correlation coefficient (CCC) proposed by Sokal & Rohlf (1962) was used, which is estimated based on the following formula:

$$CCC = r_{coph} = \frac{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(c_{jj'}-\bar{c}\right)\left(ph_{jj'}-\overline{ph}\right)}{\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(c_{jj'}-\bar{c}\right)^2}\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(ph_{jj'}-\overline{ph}\right)^2}}$$

where:

$$\bar{c}=\frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}c_{jj'} \qquad \text{and} \qquad \overline{ph}=\frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}ph_{jj'}$$

n = number of objects.

The higher the CCC value, the lower is the distortion caused by the grouping. In practice, it should be considered that dendograms with CCC lower than 0.7 indicate that the clustering method is inadequate to resolve the information from a dataset (Rohlf, 1970).

The criterion adopted in the Tocher's method, mentioned by Rao (1952), is that the mean of the dissimilarity measures within each group must be lower than the mean distances between any pair of groups. The inclusion of a parent in a specific group is defined by comparing the increase in the mean value within a group with the maximum distance value (q) within the group of closest distances from each parent.

Initially, the parent pair (laying lines) with the closest distance $D^2_{ii'}$ was identified and, if the distance was not higher than the established limit (criterion), the first grouping was formed. Afterwards, according to the adopted criterion, it was evaluated if it would be possible to include other parents in the group or if it would be necessary to form other groups, following the same criterion. The mean increase in the within-group distance was determined as the ratio between the distance of the group and the parent to be inserted in it and the number of parents in that group, following the same criterion.

The criterion used in cluster formation is schematically shown by:

$$D^2_{(ij)k} = D^2_{ik} + D^2_{jk} \quad or \quad D^2_{(ijk)l} = D^2_{(ij)l} + D^2_{kl}$$

If $\frac{D^2(Cluster)i}{g} \leq \theta \Rightarrow$ the parent (line) should be included in the cluster;

If $\frac{D^2(Cluster)i}{g} > \theta \Rightarrow$ the parent (line) should not be included in the cluster;

so that:
q = limit for addition;
j, *k*, *l* = parents in the cluster;
i = parent to be included, or not, in the group;
g = number of parents that comprises the cluster that is being formed

This method is different from the hierarchical methods because the formed clusters are mutually exclusive or, in the light of the group theory, because the original group of objects is sub-divided into disjoint subgroups, whose intersection is an empty set and the union of the subgroups reconstitutes the original group (Cruz, 1990).

The consistency in the clustering pattern obtained by the optimization method was evaluated by means of discriminant analysis based on the methodology described by Anderson (1958). Further information concerning the discriminant analysis is given by Mardia *et al.* (1997), Johnson & Wichern (1998), Khattree & Naik (2000) and Cruz & Carneiro (2003).

The relative importance of each trait in the genetic divergence was evaluated according to the methodology described by Singh (1981).

RESULTS AND DISCUSSION

Trait means of the five laying hen lines are shown in Table 1.

The pairwise standardized squared Mahalanobis distances (D^2) are shown in Table 2. The maximum D^2 distance was seen between lines 2 and 5 ($D^2 = 0.0054$) and the minimum distance between lines 3 and 5 ($D^2 = 0.0016$), showing that lines 3 and 5 are the most similar and lines 2 and 5 are the most divergent.

According to Silveira Neto (1986), it is worth noting that, for any dissimilarity measure, values are comparable only if within the same study and it is worthless to compare similarity results between objects or samples that were not included in its determination.

Table 1 - Means of the traits evaluated in five lines of laying hens.

Lines /Traits	1	2	3	4	5
LR (egg number/ day)	0.65	0.67	0.67	0.64	0.64
EW40 (g)	57.56	57.82	54.84	56.98	57.24
EW44 (g)	59.00	59.42	56.76	59.83	59.07
EW52 (g)	60.96	60.16	59.02	62.15	61.39
EW60 (g)	60.91	60.11	57.71	60.23	60.03
BW40 (g)	1613	1506	1484	1571	1551
BW48 (g)	1593	1492	1471	1542	1567
BW56 (g)	1615	1511	1483	1567	1583

LR - laying rate from 40 to 60 weeks, EW40 - egg weight at 40 weeks of age (wk), EW44 - egg weight at 44 wk, EW52 - egg weight at 52 wk, EW60 - egg weight at 60 wk, BW - body weight at 40 wk, BW48 -body weight at 48 wk, BW56 - body weight at 56 wk.

Table 2 - Standardized squared Mahalanobis distances (**D^2**) between the five lines of laying hens.

Lines	1	2	3	4	5
1	0	0.0037	0.0038	0.0049	0.0031
2		0	0.0052	0.0035	0.0054
3			0	0.0050	0.0016
4				0	0.0040
5					0

Since it is desirable to have pairwise information in the clustering procedure, a relatively large number of estimates of dissimilarity measures is produced, making it impossible to recognize homogeneous groups only by visual examination of the estimates (Johnson & Wichern, 1998).

The lines have been clustered according to the closest D^2 distances, given that the nearest-neighbour hierarchical method was used. Table 3 shows the summary of the nearest-neighbour method, enabling better visualization of the cluster formation.

The dendogram obtained by the nearest-neighbour method is shown in Figure 1. The degree of similarity between parents, between the parents and similar clusters or between two different groups can be seen in the dendogram. In this case, the clusters are subjectively defined, based on the highest-level differences and the previous knowledge of the evaluated parents (Regazzi, 2002).

Table 3 - Summary of the nearest-neighbour method

Step	Joining	Distance	Distance in %
1	3; 5	0.0016	43.25
2	3, 5; 1	0.0031	83.37
3	2; 4	0.0035	92.75
4	3, 5, 1; 2, 4	0.0037	100

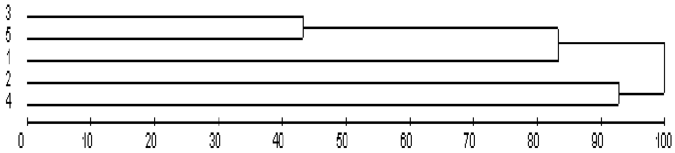

Figure 1 - Dendogram obtained by the nearest-neighbour method with D2 distance (in percentage) of the evaluated lines.

The analysis of the dendogram that represents the genetic divergence between lines obtained using the nearest-neighbour method (Figure 1) shows three distinct clusters. Taking into consideration the changes in the degree of similarity to produce the clusters, the first one comprises lines 3 and 5, the second comprises line 1 and the third, lines 2 and 4.

The produced dendogram can largely simplify the original information and some distortions on the similarity patterns between the groups may result. Thus, the last step in the joining process is to evaluate if the results are adequate (Cruz & Carneiro, 2003).

A criterion used to evaluate whether the results are adequate is the cophenetic correlation coefficient (CCC), which should preferably be higher than 0.70 (Rohlf, 1970). In the present study, CCC was 0.848, indicating that the clustering method was adequate to resolve the information from the dataset.

Nevertheless, dendograms should be carefully employed. Although results are traditionally displayed in tree diagrams, which have the advantage of being easily interpretable, the number of optimal clusters, by inference, may be not so easily visualized.

The used algorithms produce clusters that constitute a proposition about the basic and unknown organization of the data. Nevertheless, the difficulty lies on the determination of the ideal number of clusters. According to Cruz *et al.* (2004), a great number of clusters is usually evaluated, and based on an optimization criterion, the most convenient is selected.

In the optimization methods, the parental group (set) is partitioned into non-empty subgroups that are mutually exclusive by means of the maximization or minimization of a pre-established measure (Cruz *et al.*, 2004). Therefore, the differences among optimization techniques result from the methods that are used for the initial partition and the aggregation criterion used for optimization. Thus, a great number of optimization techniques uses as aggregation criterion the minimization of the residue of the dispersion matrix within groups and the minimization of the determinant, among others. Nevertheless, a method largely used in animal breeding and frequently reported is the Tocher's method, quoted by Rao (1952).

In the clustering analysis by the Tocher's optimization method, only two clusters were produced (cluster I = lines 1, 3 and 5; and cluster II = lines 2 and 4). Therefore, the nearest-neighbour hierarchical method and the Tocher's optimization method showed divergent results concerning group partition, corroborating previous results reported by Sakaguti *et al.* (1996).

Table 4 presents the mean distances established by the Tocher's optimization method, both within and between groups. The mean distance within clusters was always smaller than the distance between clusters.

Table 4 - Mean distance within- and between-clusters.

Cluster	I	II
I	0.0028	0.0047
II		0.0035

It is possible to evaluate the differences between the studied traits using the means of the clusters formed in the clustering analysis. Table 5 presents trait means for each cluster produced with the five studied laying hen lines.

Table 5 - Trait means of the clusters formed by five lines of laying hens.

Trait	Cluster I	Cluster II
LR (number eggs/day)	0.59	0.62
EW40 (g)	56.67	57.71
EW44 (g)	58.13	59.77
EW52 (g)	60.20	61.39
EW60 (g)	59.73	60.53
BW40 (g)	1539	1530
BW48 (g)	1552	1512
BW56 (g)	1565	1554

LR - laying rate from 40 to 60 weeks, EW40 - egg weight at 40 weeks of age (wk), EW44 - egg weight at 44 wk, EW52 - egg weight at 52 wk, EW60 - egg weight at 60 wk, BW - body weight at 40 wk, BW48 - body weight at 48 wk, BW56 - body weight at 56 wk.

Considering the variables directly related to the productivity in laying hens, cluster II (lines 2 and 4) was superior, i.e., laying hens from cluster II showed higher laying rate and egg weight, besides lower body weight, when compared to the laying hens from cluster I (lines 1, 3 and 5). It is worth noting that lower body weight indicates lower feed intake for maintenance. The high laying rate, the intermediate size of the egg and the small body weight are variables that characterize the best lines when the cost:benefit ratio is used as a classification criterion (Piassi *et al.*, 1995a).

The negative genetic correlations between some of these traits prevent fast improvement of the productive performance of bird flocks, even if the flocks are submitted to constant selection. In laying hen lines, Braccini Neto et al. (1997) reported that body weight showed desirable correlation with sexual maturity, egg weight, laying rate and egg mass, but undesirable correlation with feed efficiency. It was also pointed out that, if the target of the selection program is to increase body weight, the result should be a more precocious and productive line, but with poorer feed efficiency.

Table 6 presents the relative importance of each trait in the genetic divergence according to the methodology of Singh (Lengh, 1981). Body weight at 48 weeks of age (BW48) had the highest relative contribution for divergence, i.e., 24,44%.

Table 6 - Relative importance of traits ($S_{.j}$) in the genetic divergence of five lines of laying hens.

Trait	$S_{.j}$	$S_{.j}$ (%)
LR (number eggs/day)	0.003978	9.8951
EW40 (g)	0.007196	17.8988
EW44 (g)	0.005135	12.7718
EW52 (g)	0.007490	18.6302
EW60 (g)	0.001658	4.1229
BW40 (g)	0.004915	12.2251
BW48 (g)	0.009826	24.4416
BW56 (g)	0.000006	0.0144

LR - laying rate from 40 to 60 weeks, EW40 - egg weight at 40 weeks of age (wk), EW44 - egg weight at 44 wk, EW52 - egg weight at 52 wk, EW60 - egg weight at 60 wk, BW - body weight at 40 wk, BW48 - body weight at 48 wk, BW56 - body weight at 56 wk.

The data matrix should be re-examined to evaluate if the results obtained by clustering methods are adequately partitioned. In such cases, it is common to use the discriminant analysis. According to Cruz & Carneiro (2003), the consistency of the clustering obtained by the optimization method may be evaluated by the discriminant analysis of the obtained data, generally based on the methodology of Anderson (1958). Therefore, after the cluster partition in the present study was evaluated using the Anderson discriminant analysis, the existence of two clusters was confirmed: cluster I comprising lines 1, 3 and 5, and cluster II comprising lines 2 and 4.

CONCLUSION

There was genetic divergence between the evaluated laying hens lines, so that two lines showed higher laying rate, egg weight and body weight.

The nearest-neighbour hierarchical method and the Tocher's optimization method did not agree in the evaluation of the genetic divergence among the laying hen lines.

The trait that contributed mostly to the genetic divergence was the body weight at 48 weeks of age.

REFERENCES

Anderson TW. An introduction to multivariate statistical analysis. New York (NY): John Wiley & Sons; 1958.

Braccini Neto J, Dionello NJL, Silveira Júnior P, Bongalhardo DC, Estivalet Júnior CNO. Estimativa de parâmetros genéticos e fenotípicos da curva de crescimento de galinhas poedeiras. Revista da Sociedade Brasileira de Zootecnia 1997; 26(5):894-904.

Cruz CD. Aplicação de algumas técnicas multivariadas no melhoramento de plantas. [Dissertação]. Piracicaba (SP): Escola Superior de Agricultura Luiz de Queiros / Universidade de São Paulo; 1990.

Cruz CD, Carneiro PCS. Modelos biométricos aplicados ao melhoramento genético. Volume 2. 1 ed. Viçosa (MG): UFV, Imprensa universitária; 2003.

Cruz CD, Regazzi AJ, Carneiro PCS. Modelos biométricos aplicados ao melhoramento genético. 3 ed., Viçosa (MG): UFV, Imprensa universitária; 2004.

Johnson RA, Wichern DW. Applied multivariate statistical analysis. 4. ed., Englewood Cliffs, Prentice Hall; 1998.

Khattree R, Naik DN. Multivariate data reduction and discrimination with SAS software, BBU Press and John Wiley Sons Inc.; 2000.

Mahalanobis, PC. On the generalized distance in statistics. Proceedings of Natural Institute of Sciences 1936; 2: 49-55.

Mardia KV, Kent JT, Bibby JM. Multivariate analysis. 6. ed., London, Academic press, 1997.

Piassi MA. Avaliação do desempenho de linhagens de postura mantidas na Universidade Federal de Viçosa, em competição com marcas comerciais. [Dissertação]. Viçosa (MG): Universidade Federal de Viçosa; 1994.

Piassi M, Silva MA, Regazzi AJ, Torres RA, Soares PR, Carneiro AM. Avaliação de diferentes grupos genéticos de aves de postura, usando-se análise de variância multivariada. Revista da Sociedade Brasileira de Zootecnia 1995; 24(3):453-460.

Piassi M, Silva MA, Regazzi AJ, Torres RA, Soares PR, Torres Junior AA. Estudo da divergência genética entre oito grupos de aves de postura, por meio de técnicas de análise multivariada. Revista da Sociedade Brasileira de Zootecnia 1995; 24(5):715-727.

Rao CR. Advanced statistical methods in biometric research. New York (NY):John Wiley & Sons; 1952.

Regazzi AJ. Análise multivariada. Viçosa: Universidade Federal de Viçosa, 2002. (INF-766) (notas de aula).

Rohlf FJ. Adaptative hierarquical clustering schemes. Systematic Zoology 1970; 19(1):58-82.

Sakaguti ES, Silva MA, Regazzi AJ, Torres RA, Martins LN. Análise de divergência genética entre nove grupos genéticos de coelhos. Revista da Sociedade Brasileira de Zootecnia 1996; 25(3):647-660.

Silveira Neto S. Análise fenética. In: ALVES, S.B. (coord.) Controle microbiano de insetos; 1986; São Paulo, Manole, p.384-407.

Singh D. The relative importance of characters affecting genetic divergence. Indian Jornal of Genetics 1981; 41(2):237-245.

Sokal RA, Rohlf FJ. The comparison of dendograms by objective methods. Taxonomy; 1962:11-33-40.