Centro Agrícola, 40(4): 19-24; octubre-diciembre, 2013
ISSN papel: 0253-5785 ISSN on line: 2072-2001
CE: 40,13 CF: cag054131941

40 Aniversario Centro Agrícola

EDITORIAL Feijóo

ARTICULO GENERAL

Clasificación y determinación del número óptimo de conglomerados en bancos de germoplasma

Classification and determination of the optimal cluster number in Plant Germplasm Banks

Osmany Molina Concepción[1], Raisa García Rodríguez[1], Marilys Milián Jimenez[1], Lianet González Díaz[1], Carmen C. Pons Pérez[1] y Ricardo Grau Abalo[2]

[1] Instituto de Investigaciones de Viandas Tropicales (INIVIT). Apdo #6. Santo Domingo, Villa Clara, Cuba.
[2] Universidad Central "Marta Abreu" de Las Villas. Carretera a Camajuaní km 6½. Santa Clara, Villa Clara, Cuba.

E-mail: osmany@inivit.cu

RESUMEN. El presente trabajo está encaminado a determinar el método de combinación de agrupamiento que mejor responda a la clasificación no supervisada de los bancos de germoplasma en estudio y el índice de validación que mejor determine el número óptimo de conglomerados jerárquicos aglomerativos. En la investigación se utilizó la métrica de Gower para variables mixtas y se combinó las soluciones provenientes de los diferentes métodos de agrupamiento para obtener una topología consenso. La fortaleza de los agrupamientos obtenidos del conjunto de datos por los métodos de aglomeración fue evaluada con el coeficiente aglomerativo y los conglomerados consenso, con el índice RV. El número de conglomerado óptimo se determinó utilizando cinco índices de validación que no brindaron un resultado homogéneo. La mejor estructura consenso para las bases de datos en estudio se obtuvo con el método *Manhattan*. Se utilizaron funciones implementadas en el lenguaje de programación R. Esta estrategia de análisis fue aplicada por primera vez para la clasificación taxonómica de las colecciones cubanas de germoplasma de malanga (*Xanthosoma* spp.) y plátanos (*Musa* spp.).

Palabras clave: conglomerados, clasificación no supervisada, bancos de germoplasma, consenso.

ABSTRACT. The present work aims at determining the clustering combination method that better fits the non-supervised classification of germplasm banks, as well as the validation index which more efficiently determines the optimal number of agglomerative hierarchical clusters. Mixed variables Gower metric was used in the research and the solutions coming out of the diverse clustering methods were combined so as to obtain a consensus structure. The strength of the clusters obtained by the clustering methods out of the data was evaluated using the agglomerative coefficient and the consensus conglomerate with the RV index. The optimal conglomerate number is obtained using five validation indexes which didn't bring about a homogeneous result. The better consensus structure for data base was obtained using the Manhattan method. Functions implemented in R programming language were used. This analysis strategy was used for the first time for the taxonomic classification of Cuban collections of new cocoyam (*Xanthosoma* spp.) and banana (*Musa* spp.) germplasm.

Key words: clusters, non-supervised classification, germplasm banks, consensus.

INTRODUCCIÓN

Desde su creación en 1967, el banco de germoplasma del Instituto de Investigaciones de Viandas Tropicales (INIVIT), ha permitido trabajar en la caracterización, evaluación, documentación y estudio de la variabilidad conservada en Cuba de malanga (*Xanthosoma* spp.) (Milián, 2000; 2008) y plátano (*Musa* spp.) (González, 2005), la cual representan unas de las más amplias y diversas a nivel mundial y particularmente de América latina.

Los recursos fitogenéticos se han convertido en una prioridad científica, sobre todo aquellos con poco estudio y potencial comercial, lo cual hace importante el análisis de esta diversidad mediante métodos cuantitativos que ayuden a agrupar poblaciones de un mismo género o especie. Los métodos que se utilizan generalmente en el estudio de divergencias entre individuos siguen una aproximación fenética o numérica (Franco y Hidalgo, 2003).

La taxonomía numérica intenta construir clasificaciones "naturales", basadas en las semejanzas fenotípicas de los individuos que se valoran partiendo de una adecuada elección de un coeficiente de similaridad (Cuadras, 1981). En los últimos años se han desarrollado diversos algoritmos de clasificación no supervisada capaces de realizar tal tarea: jerárquicos y no jerárquicos, los cuales dan solución a un número considerable de problemas en las ciencias biológicas, a pesar de no producir muchas vece, clasificaciones objetivas ni estables, pues se pueden obtener diferentes agrupamientos a partir de la misma matriz de datos si se utilizan distintos procesamientos.

Para resolver en parte, esta problemática se ha sugerido la idea de la combinación de agrupamientos (Clustering Ensemble) (Vega-Pons *et al.* 2010a). Esta responde a la idea intuitiva de que si no se conoce la calidad de ciertos resultados individuales, la opción de combinarlos puede ser superior a seleccionar algún resultado simple.

El presente trabajo tiene como objetivo analizar el desempeño de la combinación de agrupamientos con cuatro métodos de aglomeración jerárquicos y una estructura de datos, así como la determinación del número óptimo de conglomerados en diferentes topologías en la clasificación de un grupo de genotipos de malanga (*Xanthosoma* spp.) y plátanos (*Musa* spp.).

MATERIALES Y MÉTODOS

El estudio se realizó con dos conjuntos de datos compuesto de accesiones de malanga (*Xanthosoma* spp.) que contuvo 71 accesiones, donde se evaluaron 20 variables cualitativas y 16 cuantitativas (Milián, 2008). El de plátano (*Musa* spp.) incluyó 131 accesiones, con 20 variables cualitativas y siete cuantitativas, incluidas en el Sistema de Descriptores Mínimos (INIBAP-IPGRI-CIRAD, 1996), ambas colecciones se conservan en el INIVIT.

Al ser las medidas de distancia sensibles a las diferencias de escalas o de magnitudes hechas entre las variables cuantitativas (Milligan y Cooper, 1988), éstas fueron estandarizadas con la propia métrica de Gower (Gower, 1971). Después se calculó la matriz de distancia con la métrica de Gower para variables mixtas implementada en la función *daisy* (Paquete "cluster"). Finalmente se realizó un análisis de conglomerados mediante cuatro métodos de agrupamiento para ambas matrices. Los métodos de aglomeración usados fueron: Ward (Ward, 1963), Promedio ó UPGMA (Sneath y Sokal, 1973), agrupación de enlace simple (Gower, 1967) y agrupación de enlace completo (Sorensen, 1948), con la función *hclust* (paquete "stats").

Para combinar los resultados de los diferentes algoritmos de aglomeración se utilizó la función *cl_ensemble* del paquete "clue". El árbol consenso, que es capaz de combinar toda la información existente de los diferentes resultados en un árbol final, se obtuvo con la función *cl_consensus* (paquete "clue"), para ello se utilizó los tres métodos disponibles en esta función y calculó los conglomerados consenso en los algoritmos jerárquicos (*Euclidean, Manhattan y Majority*).

En el análisis taxonómico, una vez obtenido el resultado del método de aglomeración y su correspondiente dendrograma, se determinó si el conjunto de datos muestra una tendencia a formar grupos, lo cual fue a través del Coeficiente Aglomerativo (CA), que describe la fortaleza de la estructura obtenida y sirve para comparar la calidad de los diferentes conglomerados formados en el caso que se utilice el mismo algoritmo de agrupamiento (Rousseeuw, 1986).

No existe todavía un algoritmo de clasificación por excelencia, para un problema determinado es difícil seleccionar el método de aglomeración que logra encontrar una mejor estructura para separar las accesiones; debido a lo cual, en la búsqueda de mejores algoritmos de clasificación, aparece una tendencia a combinar varios algoritmos de agrupamiento en el mismo problema (Vega-Pons y Ruiz-Shulcloper, 2010b). en la verificación de la estructura del árbol consenso se usó el coeficiente RV (Josse, 2007) implementado en el paquete 'FactoMineR', función *coeff RV.*

Cuando se emplean técnicas de aglomeración

jerárquicas, el investigador no siempre está interesado en la jerarquía completa sino en un subconjunto de particiones obtenidas a partir de ella. Las particiones se obtienen cortando el dendrograma por lo que surgen diversos índices de validación, para determinar el mejor nivel en el que se debe cortar; o sea, validar la partición de los datos obtenida con un algoritmo de agrupamiento. Los índices utilizados en esta investigación fueron: El índice de Calinski-Harabasz (*indice.G1*) considerado el de mejor desempeño en un estudio realizado por Milligan y Cooper (1985); el índice de Baker & Hubert, una adaptación del estadístico Gamma de Goodman & Kruskal's (*indice.G2*); el índice de Hubert & Levine (*indice.*G3); el Ancho de la Silueta, implementados en el paquete "clusterSim"; y el índice de Dunn, implementado en el paquete "clValid".

En el procesamiento de la información se utilizó un lenguaje de programación orientado a objetos denominado R (R Development Core Team, 2011); el cual es un conjunto de programas integrados para análisis estadísticos y gráficos. R es un *software* libre, por lo que la implementación de cualquier técnica en este lenguaje le da mayor potencialidad e independencia (*indice.G2*), índice de Hubert & Levine (*indice*.G3) y Ancho de la Silueta implementados en el paquete 'clusterSim', y el índices de Dunn implementado en el paquete 'clValid'.

Para procesar la información se utilizó un lenguaje de programación, orientado a objetos denominado R (R Development Core Team, 2011); el cual es un conjunto de programas integrados para análisis estadísticos y gráficos. R es un *software* libre, por lo cual la implementación de cualquier técnica en este lenguaje le dará mayor potencialidad e independencia.

RESULTADOS Y DISCUSIÓN

Al aplicar la métrica de Gower para variables mixtas a la matriz de datos de variables cualitativas y cuantitativas se obtuvo una matriz de distancia entre las accesiones de *Xanthosoma* spp. y *Musa* spp., a la que se le aplicó los cuatro métodos de aglomeración seleccionados. En la Tabla 1 se observa que el método Ward es el que mayor CA tiene con relación a los demás, en las dos bases de datos, por lo cual se considera que se obtuvo una estructura fuerte.

Tabla 1. Resumen del Coeficiente Aglomerativo con las matrices de datos en estudio

Bases de Datos	Métodos	CA	
Xanthosoma spp.	Ward	1	0.884
	Promedio	3	0.554
	Simple	4	0.464
	completo	2	0.625
Musa spp.	Ward	1	0.992
	Promedio	3	0.813
	Simple	4	0.784
	completo	2	0.858

Los dendrogramas obtenidos por los métodos son diferentes tanto para la base de datos de *Xanthosoma* spp. (Figura 1) como para la de *Musa* spp., por lo que estos métodos no logran encontrar estructuras parecidas. En la tabla 2 el mayor coeficiente RV fue con *Manhattan* para ambas bases de datos y el *p-value* asociado al *test* es significativo.

Tabla 2. Resumen consenso con el índice RV para las dos matrices de datos en estudio

Bases Datos	Coeficiente RV/p.value		
	Euclidean	*Manhattan*	*Majority*
Xanthosoma spp.	0.666	0.780	0.368
	1,73E-12	5,51E-14	1,03E-07
Musa spp.	0.824	0.883	0.858
	7,27E-24	1,85E-32	1,14E-34

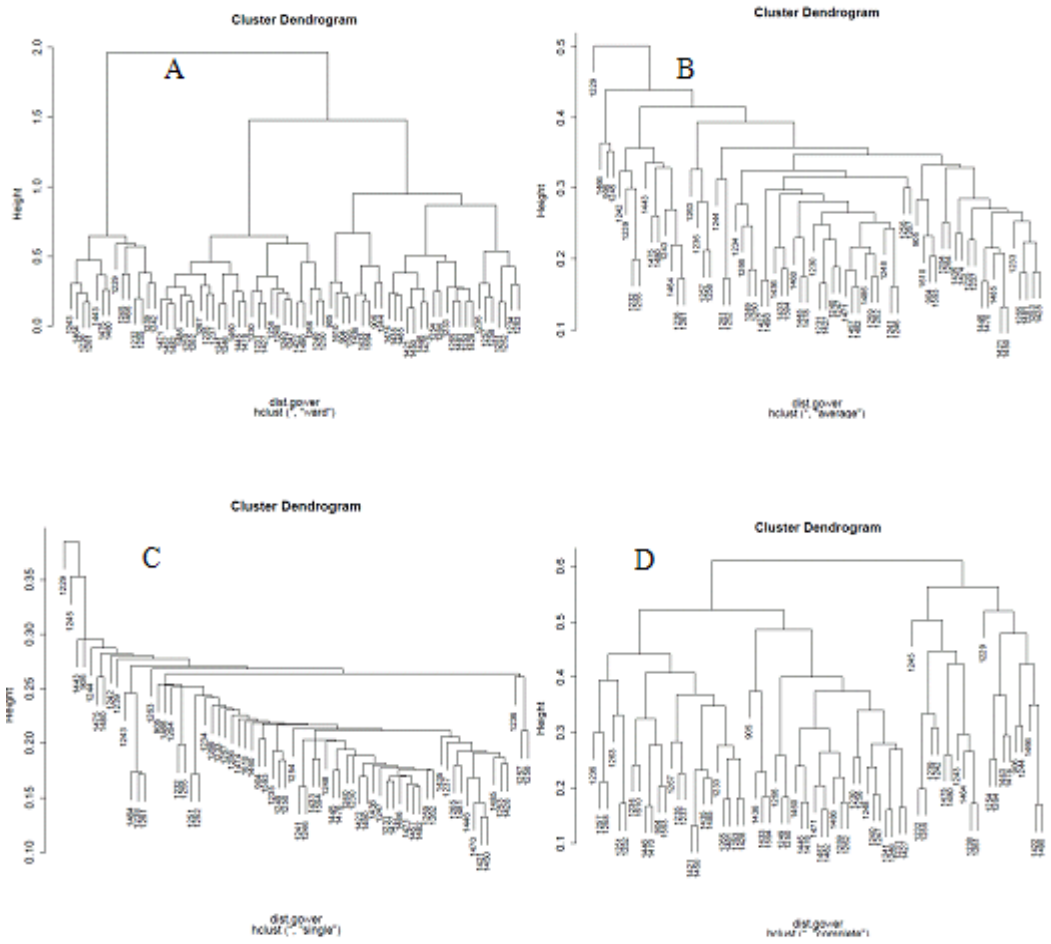

Figura 1. Dendrogramas (A, B, C y D) con la métrica de Gower. Base de datos de malanga (*Xanthosoma* spp.), para los cuatros métodos en estudio

Al obtener el árbol consenso con las tres metodologías disponibles para calcular los conglomerados consenso en los algoritmos jerárquicos, con el método de la distancia euclidiana, se logra una mejor interpretación de los resultados que con *Manhattan* y *Majority* (Figura 2).

Los cinco índices de validación probados para determinar la mejor partición de los datos (Tabla 3) muestran una tendencia a elegir la partición dos en *Xanthosoma* y *Musa* con una relación 11/20 y 7/20 respectivamente en las bases de datos, lo cual hace difícil tomar una decisión sobre cuál es la mejor partición sin tener una clase *a priori* (tabla 3).

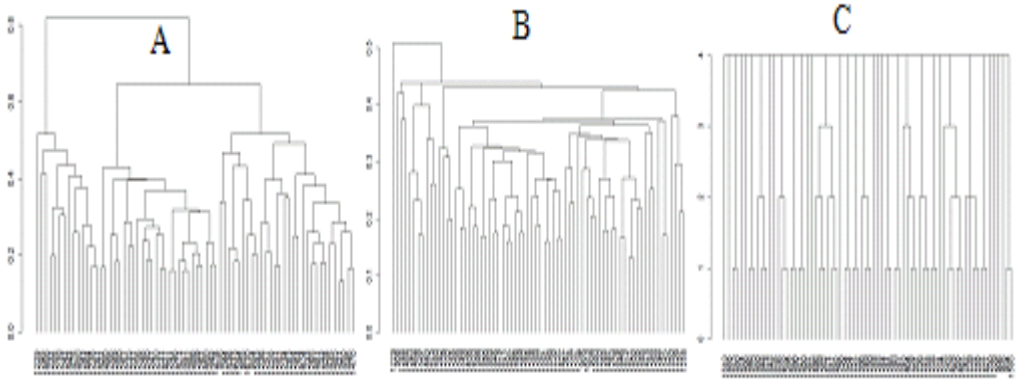

Figura 2. Dendrograma consenso con los métodos *Euclidean* (A), *Manhattan* (B) y *Majority*

Estos índices de validación con los cuatro métodos de aglomeración, no son homogéneos lo cual provoca la duda a la hora de escoger el verdadero número de clústeres ya que puede aparecer más de una posible solución (Rodríguez, 2012); por ello se impone un análisis de la estructura del

dendrograma por especialistas de los cultivares junto a los índices estudiados para determinar aquellos conglomerados que se ajustan mejor a las características de las bases de datos en estudio según el cultivo evaluado.

Tabla 3. Resultados de cinco índices de validación aplicando cuatro métodos jerárquicos aglomerativos sobre dos matrices de datos

Bases Datos	Métodos	Índices de validación				
		ID	S	G1	G2	G3
Xanthosoma spp.	Ward	18/0.453	2/0.193	2/10.680	20/0.862	2/0.477
	Promedio	2/0.631	2/0.281	4/5.504	2/0.883	2/0.265
	Simple	2/ 0.631	2/0.281	3/2.185	3/ 0.906	2/0.265
	completo	12/ 0.44	20/0.172	2/7.628	20/0.862	3/0.508
Musa spp.	Ward	2/ 0.430	11/0.458	2/43.934	11/0.965	2/0.507
	Promedio	18/0.518	15/0.505	3/39.742	13/0.989	2/0.536
	Simple	2/ 0.544	18/0.454	5/14.827	18/0.988	2/0.467
	completo	5/ 0.482	11/0.492	3/39.742	11/0.984	2/0.536

CONCLUSIONES

1. La selección del método de clasificación más adecuado no es trivial debido a la cantidad de variantes, por lo que al combinarlos se obtienen resultados con niveles de exactitud y precisión superior al desempeño de ellos por separados, elementos fundamentales a alcanzar en problemas de análisis taxonómicos.

2. La selección del índice de validación para determinar el número óptimo de clúster no es homogéneo, se impone acompañar a este de la experiencia del especialista del cultivar.

3. Por su flexibilidad, la concepción de estos análisis pueden ser aplicadas a otros estudios de clasificación en bancos de germoplasma vegetal.

BIBLIOGRAFÍA

1. Cuadras, C. M.: Métodos de Análisis Multivariante. *Eunibar* DL XII: 642, 1981.

2. Franco, T. L.; R. E. Hidalgo (eds): Análisis Estadístico de Datos de Caracterización Morfológica de Recursos Fitogenéticos. Instituto Internacional de Recursos Fitogenéticos (IPGRI), Cali, Colombia, *Boletín técnico* vol. 8, 2003.

3. González, Lianet: Caracterización de la variabilidad genética en genotipos viandas del género *Musa*. Tesis Maestría, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Villa Clara, Cuba, 2005.

4. Gower, J. C.: A comparison of some methods of cluster analysis. *Biometrics* (23):623-628. 1967.

5. Gower, J. C.: A general coefficient of similarity and some of its properties. *Biometrics* (27):857-871. 1971.

6. INIBAP-IPGRI-CIRAD (ed.): Descriptors for Banana (*Musa* spp.). 1996.

7. Josse, J.: Testing the significance of the RV coefficient. Aveiro, Portugal, 2007.

8. Milián, J. Marilys.: Caracterización de la variabilidad del género *Xanthosoma* en Cuba. Tesis para obtener el Título de Maestro en Ciencias Biológicas, Universidad de la Habana, La Habana, Cuba, 2000, 103 p.

9. Milián, J. Marilys.: Caracterización de la variabilidad de los cultivares de la colección cubana de germoplasma del género *Xanthosoma* (Araceae). Tesis para aspirar al grado de Doctor en Ciencias Biológicas, Ciudad de la Habana, La Habana, Cuba, 2008, 123 p.

10. Milligan, G. W.; M. C. Cooper: An examination of procedures for determining the number of clusters in data set. *Psychometrika* 50 (2):159-179; 1985.

11. Milligan, G. W.; M. C. Cooper: A study of standardization of variables in cluster analysis. *Journal of Classification* (5):181-204; 1988.

12. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. R Foundation for Statistical Computing. 2011. En sitio web: http://www.r-project.org/. 2011.

13. Rousseeuw, P. J.: A visual display for hierarchical classification. *Data Analysis and Informatics* (4):743-748; 1986.

14. Rodríguez García, Raisa: Análisis taxonómico numérico de Bancos de Germoplasma. Tesis presentada en opción al título académico de Master en Ciencia de la Computación, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Villa Clara, Cuba, 2012.

15. Sneath, P. H. A.; R. R. Sokal: Numerical taxonomy. The principles and practice of numerical classification. W. H. Freeman and Co, San Francisco, USA, 1973.

16. Sorensen, T.A.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to alalysies of vegetation on Danish commons. *Biologiske Skrifter* (5):1-34; 1948.

17. Vega-Pons, S. [*et al*.]: Weighted partition consensus via kernels. Pattern Recognition 43 (8):2712–2724, 2010a.

18. Vega-Pons, S.; J. Ruiz-Shulcloper: Combinación de agrupamiento: un estado del arte. ISSN 2072-6287. *CENATAV*, La Habana, Cuba, 2010b.

19. Ward, J. H.: Hierarchical grouping to optimize an objective function. *J Amer Statist Assoc* (58):236-244; 1963.

Recibido: *20/05/2013*
Aceptado: *30/08/2013*