# ANÁLISIS ESTRUCTURADO POR ETAPAS PARA LA CLASIFICACIÓN DE BANCOS DE GERMOPLASMA VEGETAL

**Osmany Molina Concepción[1] Raisa L. García Rodríguez[1] Marilys Milián Jiménez[2] Carmen C. Pons Pérez[1] Ricardo Grau Abalo[3]**

*1. Departamento de Bioinformática, Instituto de Investigaciones de Viandas Tropicales (INIVIT), Cuba.*
*2. Departamento de Genética, Instituto de Investigaciones de Viandas Tropicales (INIVIT), Cuba.*
*3. Departamento de Bioinformática, Universidad Central "Marta Abreu" de Las Villas (UCLV), Cuba.*
taxonumeric@inivit.cu

**Resumen**
El presente trabajo tiene como objetivo diseñar un procedimiento de análisis estructurado por etapas para la clasificación de Bancos de Germoplasma Vegetal que permite combinar las potencialidades de los métodos aglomerativos jerárquicos y definir en cada momento del análisis, la mejor variante a aplicar, así como la validación de conglomerados para determinar la consistencia de la clasificación de bancos de germoplasma a partir de los rasgos que los caracterizan. El desempeño de los métodos de aglomeración fue evaluado con el coeficiente de correlación cofenético. Se validó la calidad del agrupamiento de las accesiones que conforman los bancos de germoplasma con el índice de Dunn. Para los diferentes análisis se utilizaron funciones implementadas sobre la base del lenguaje de programación R. En está investigación se demostró la fortaleza del nuevo modelo que soluciona algunas de las debilidades que se manifiestan al aplicar los métodos taxonómicos aglomerativos clásicos y permite discernir mejor las características que comparten las accesiones en cada uno de los grupos que se forman.

**Palabras clave**: conglomerados, clasificación, combinación de agrupamientos, bancos de germoplasma.

## Introducción

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 and Hidalgo 2003). En este tipo de clasificación son extremadamente empleados los métodos de aglomeración jerárquica.

Los métodos de aglomeraciones jerárquicos intentan obtener particiones lo más naturales posibles, pero en la práctica se forman agrupaciones globales que no explican todas las relaciones reales entre las accesiones. En muchos casos, estos métodos no son capaces de reubicar accesiones que han sido ubicadas en un grupo al cual no pertenecen.

La selección del modelo más adecuado no es trivial, debido a la cantidad de variantes, por lo que se puede analizar la conveniencia de combinarlos para obtener resultados con niveles de exactitud y precisión superior al desempeño de ellos por separados, elementos fundamentales a alcanzar en problemas de análisis taxonómicos (García 2012).

Por lo anterior, se diseñó un procedimiento de análisis estructurado por etapas que soluciona algunas de las debilidades que se manifiestan al aplicar los modelos clásicos de análisis no supervisados de agrupamiento, y permite discernir mejor las características que comparten las accesiones en cada uno de los grupos que se van formando, lo cual permite obtener una mejor clasificación taxonómica de las accesiones presentes en los bancos de germoplasma.

**Materiales y Métodos**

Se usaron datos procedentes de un estudio de accesiones de ñame (*Dioscorea* spp.) del Banco de Germoplasma, que se conserva en el Instituto de Investigaciones de Viandas Tropicales (INIVIT).

La colección analizada de *Dioscorea* spp. contiene 86 accesiones donde se evaluaron 43 variables cualitativas y nueve variables cuantitativas incluidas en el Sistema de Descriptores Mínimos (Sánchez *et al*. 1995).

A continuación se describen un grupo de técnicas necesarias para llevar acabo el procedimiento de análisis estructurado por etapas para la clasificación de Bancos de Germoplasma Vegetal.

Primeramente se propone trabajar con los descriptores mínimos establecidos internacionalmente para este banco de germoplasma en estudio, que son los mejores caracteres taxonómicos, y como método de selección de variables, solo eliminar aquellos rasgos uniformes.

La estandarización se realiza con la métrica de Gower (1971), la cual estandariza cada variable dentro de un rango de [0,1] (Podani and Schmera 2006), y además permite el análisis de variables mixtas. Esta métrica esta implementada en la función *gowdis*(), descrita en el paquete *FD*.

Existen dos tipos básicos de agrupamientos que se usan para descubrir estructuras de clasificación natural en los datos (Naresh et al. 1986), estos se distinguen por ser de naturaleza jerárquica o no jerárquica (Johnson 2000).

En esta investigación se usan los métodos de aglomeración jerárquicos de Ward (Ward 1963), UPGMA *(Unweighted Pair-Group Method using Arithmetic Averages*) (Sneath and Sokal 1973), *Single Linkage Agglomerative Clustering* (Gower 1967), *Complete Linkage Agglomerative Clustering* (Sorensen 1948) con la función *hclust*() en el paquete *stats* que forma parte de la librería básica de R que se instala por defecto.

En la búsqueda de mejores algoritmos de clasificación aparece una tendencia a combinar (*Clustering Ensemble*) (Vega and Ruiz 2010) varios algoritmos de agrupamiento en el mismo problema. La base de estos algoritmos está en utilizar el criterio de varios expertos y combinarlos en aras de lograr un mejor rendimiento.

El paquete *clue* (Hornik 2011) permite crear y analizar combinación de agrupamientos, y obtener una estructura consenso. Para aglutinar los resultados de los diferentes algoritmos de aglomeración se usó la función *cl_ensemble*() de este paquete.

Una vez obtenido el resultado del método de aglomeración, en el análisis taxonómico es importante determinar si las estructuras obtenidas por los métodos jerárquicos son aceptables o si se introducen distorsiones inaceptables en las relaciones originales. Para verificar este hecho se puede obtener una significación estadística aplicando el *test* de Mantel (Mantel 1967), implementado en la función *mantel*() que calcula el estadístico de Mantel como una correlación entre dos matrices de disimilaridad del paquete *vegan* (Oksanen et al. 2011).

Uno de los pasos fundamentales cuando se emplean técnicas de aglomeración jerárquicas es determinar un subconjunto de particiones obtenidas a partir de la jerarquía completa. Las particiones se obtienen cortando el dendrograma o seleccionando una de las soluciones en la sucesión de conglomerados que comprende la jerarquía. Las soluciones propuestas han

sido diversos índices de validación donde algunos de los más utilizados son: Índice de Calinski-Harabasz (Calinski and Harabasz 1974), el índice de Davies-Bouldin (Davies and Bouldin 1979); el Ancho de la Silueta (*silhouette*) (Kaufman and Rousseuw 1990), todos implementados en el paquete *clusterSim* y el índice de Dunn (Dunn 1974), implementado en el paquete *clValid*.

Las herramientas antes descritas son el centro del algoritmo para el análisis estructurado por etapas para la clasificación de Bancos de Germoplasma Vegetal, el cual se describe a continuación de forma detallada:

1. Se eliminan las variables que tengan rasgos uniformes de los datos de entrada, formados por las variables discretas (nominales y ordinales) y continúas.
2. Se calcula la matriz de distancia a través de la métrica de Gower.
3. Se aplican los métodos jerárquicos aglomerativos.
4. Se aplica el *test* de Mantel que permite determinar la correlación entre la matriz ultramétrica de cada árbol jerárquico con la matriz de distancia original de los datos.
5. Se seleccionan los métodos cuya estadística del *test* de Mantel resulta mayor que el cuantil del 50%. Aunque este criterio puede ser variable según el razonamiento del especialista.
6. Los resultados de los métodos seleccionados se combinan y a partir de esto se obtiene un árbol consenso, cuyo principio se basa en clasificar cada individuo en el conglomerado dónde la mayoría de los métodos lo ubicó con anterioridad.
7. El dendrograma del árbol consenso se particiona con algunos de los índices de validación descritos anteriormente, para determinar el mejor corte.
8. Se selecciona el número de particiones para el cuál se haya obtenido el mayor valor del índice.
9. Mostrar las agrupaciones obtenidas e identificar las relaciones lógicas entre las accesiones por parte del taxónomo de acuerdo a su experiencia, y reubicar, si así lo considera, accesiones que han sido ubicadas en un grupo al cual no pertenecen según su criterio.
10. Para cada grupo obtenido repetir paso del 1 al 8 hasta que se obtengan grupos con número máximo de individuos igual a 5, criterio que puede ser variable de acuerdo al interés del especialista. O sea, se repite todo el procedimiento anterior para cada partición ya conformada, cuyo número de individuos sea mayor que cinco.

Para procesar la información se utilizó el 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.

**Resultados y Discusión**

Al aplicar la métrica de Gower para variables mixtas a la matriz de datos de variables cualitativas y cuantitativas de la colección de ñame (*Dioscorea* spp.) se obtuvo una matriz de distancia entre las accesiones, a la que se le aplicó el método de aglomeración UPGMA y se obtuvo el dendrograma que se muestra en la Figura 1.

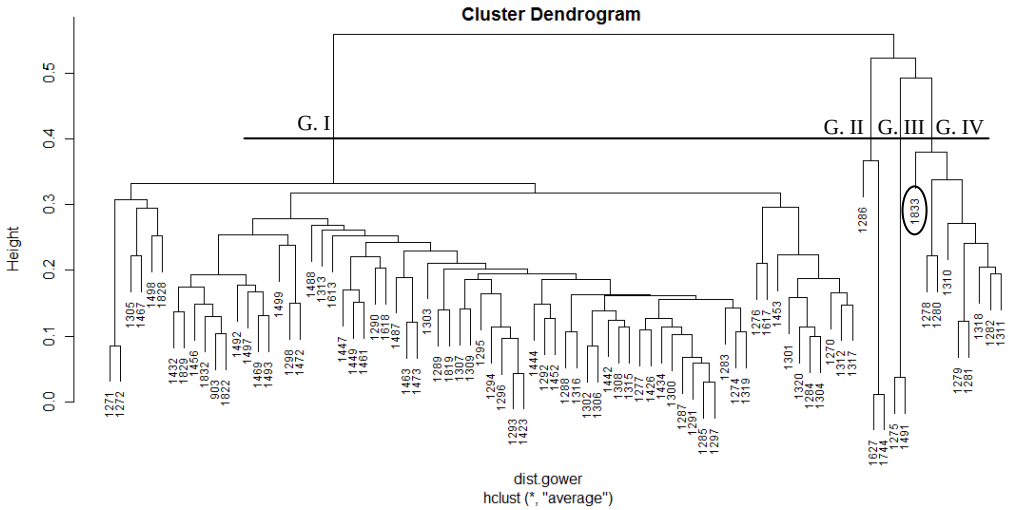


Figura 1. Dendrograma obtenido al aplicar método UPGMA con la métrica de Gower sobre la matriz de datos *Dioscorea* spp.

En la Figura 1, al trazar un corte en el dendrograma con un índice de Dunn de 0,549 teniendo en cuenta el mejor valor usado se formaron cuatros conglomerados. Como se puede observar en el grupo G.IV se encuentra la accesión 1833 que pertenece a la especie *D. alata*, ya que comparte caracteres que identifican a las accesiones de esta especie como el tallo y pecíolo alado o aristado, producción de bulbillos u tubérculos aéreos, entre otras, por lo que debe estar ubicada en el grupo G.I, formado por todas las accesiones de esta especie, por tanto su ubicación por el método UPGMA se puede considerar incorrecta, ya que en este conglomerado están sólo las especies *D. cayenensis y D. rotundata.* Mientras que al aplicar el algoritmo de análisis estructurado por etapas la ubicó correctamente en el conglomerado 1, compuesto por 73 accesiones de la especie *D. alata*, como se muestra en la Figura 2. Además, conformó tres más pequeños donde el segundo esta formado por la especie *D. bulbifera*, el tercero por las especies *D. esculenta,* el cuarto por las especie *D. cayenensis* y *D. rotundata* (Figura 3).

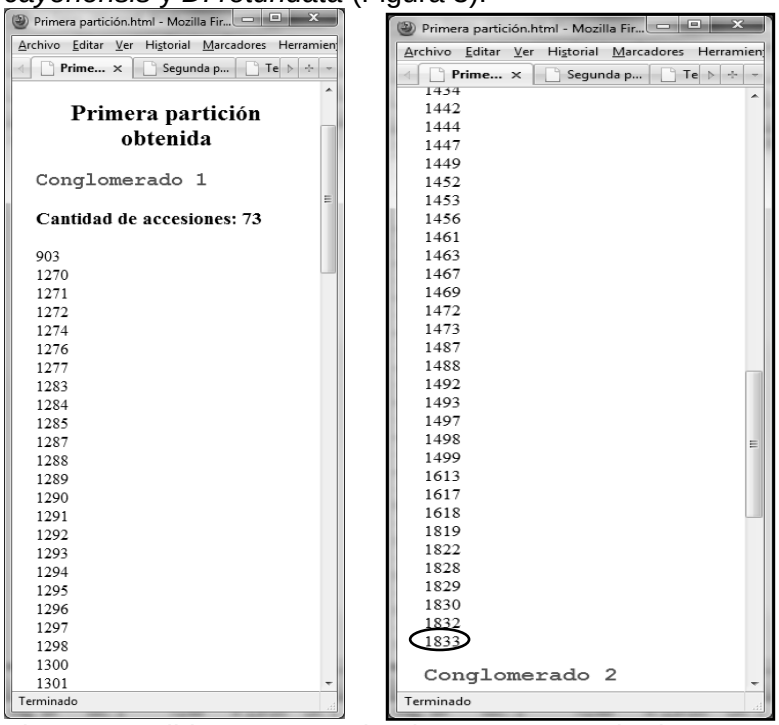


Figura 2. Salida en HTML donde se muestra el primer conglomerado de la primera partición obtenida por el algoritmo de análisis estructurado por etapas no supervisado.

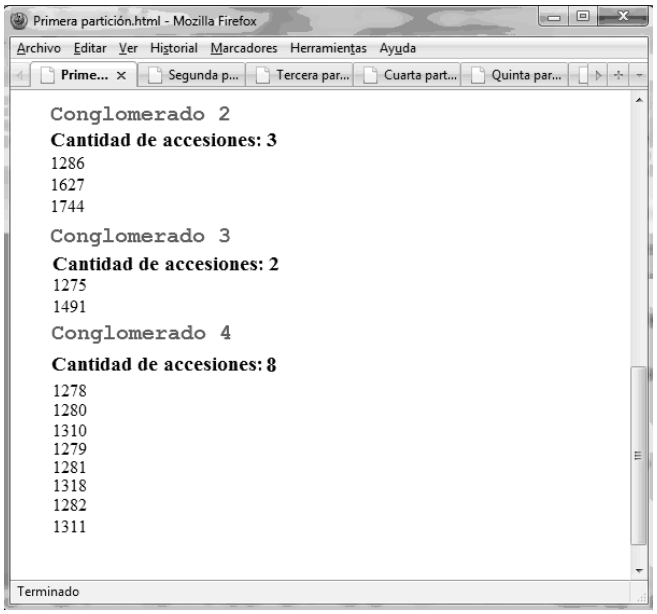

Figura 3. Salida en HTML donde se muestran los conglomerados 2, 3 y 4 de las primer partición obtenida por el algoritmo de análisis estructurado por etapas no supervisado.

Como se ha planteado, la clasificación es no supervisada, pero en el caso del ñame se puede tener una idea de la conformación de algunos grupos, debido a la existencia de estas especies.

Consecutivamente, para cada partición obtenida en cada rama del dendrograma, se busca el método de aglomeración que mejor comportamiento presenta y se obtiene el número de particiones más adecuado para establecer los subgrupos que conformarán las siguientes ramas, así sucesivamente hasta que se obtiene una estructura que permite discernir de una manera más natural la posición de cada accesión.

**Conclusiones**

Se demostró que con este procedimiento la calidad de la clasificación taxonómica obtenida es superior al desempeño individual de los métodos aglomerativos jerárquicos, ya que se establece una estructura de grupo más definida, lo que establece una conexión más natural entre las accesiones.

El algoritmo descrito permite discernir mejor las características que comparten las accesiones en cada uno de los grupos que se van formando.

**Bibliografía**

Calinski RB, Harabasz J (1974) A dendrite method for cluster analysis. Communs Statist 3:1-27

Davies DL, Bouldin DW (1979) A Cluster Separation Measure. IEEE Trans Pattern Analysis and Machine Intelligence 1:224-227

Dunn JC (1974) Well separated clusters and optimal fuzzy partitions. J of Cybernetics 4:95-104

Franco TL, Hidalgo R (eds) (2003) Análisis Estadístico de Datos de Caracterización Morfológica de Recursos Fitogenéticos. Boletín técnico IPGRI, vol 8. Instituto Internacional de Recursos Fitogenéticos (IPGRI), Cali, Colombia

García RR (2012) Análisis taxonómico númerico 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

Gower JC (1967) A comparison of some methods of cluster analysis. Biometrics 23:623-628

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857-871

Hornik K (2011) clue: Cluster ensembles. R package version 0.3-41. http://CRAN.R-project.org/package=clue.

Johnson DE (2000) Métodos Multivariados aplicados al análisis de datos. In: International Thomson Editores SA (ed)

Kaufman L, Rousseuw P (1990) Finding Groups in Data: an Introduction to Cluster Analysis. Wiley, New York

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Research 27 (2):209-220

Naresh CJ, Abhaya I, Lajpat R (1986) Monte Carlo comparison of six hierarchical clustering methods on random data. Pattern Recognition 19 (1):95-99

Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, H. Stevens MH, Wagner H (2011) vegan: Community Ecology Package. R package version 1.17-9. http://CRAN.R-project.org/package=vegan.

Podani J, Schmera D (2006) On dendrogram-based measures of functional diversity. Oikos 115:179-185

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

Sánchez I, Milián M, Rayas A, Rodríguez S (1995) Lista de descriptores y caracterización de la colección cubana de ñame (Discorea spp). Instituto de Investigaciones de Viandas Tropicales (INIVIT), Santo Domingo, Villa Clara, Cuba

Sneath PHA, Sokal RR (1973) Numerical taxonomy. The principles and practice of numerical classification. W. H. Freeman and Co, San Francisco, California, USA

Sorensen TA (1948) 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

Vega-Pons S, Ruiz-Shulcloper J (2010) A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence World Scientific Publishing Company

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Amer Statist Assoc 58:236-244