

COMBINACIÓN DE CONGLOMERADOS PARTICIONALES PARA LA CLASIFICACIÓN DEL GERMOPLASMA DE MALANGA (*Xanthosoma* spp.)

Osmany Molina Concepción*, Carmen C. Pons Pérez, Yaselis Guillen López
Instituto de Investigaciones de Viandas Tropicales (INIVIT), Cuba. INIVIT, Apartado 6, Santo Domingo,
CP: 53 000, Villa Clara, Cuba.

*Autor para correspondencia: taxonumeric@inivit.cu, <https://orcid.org/0000-0003-4199-6652>

Recibido: 7 de mayo de 2021; Aceptado: 16 de junio de 2021

RESUMEN

En el presente trabajo se evalúan diferentes métodos de agrupamiento particionales implementados en la función *kmeans* en el lenguaje de programación R, con el objetivo de identificar grupos naturales inherentes al germoplasma de malanga (*Xanthosoma* spp.) que se conserva en el Inivit. La matriz de datos está conformada por variables mixtas donde se integran las variables cualitativas (nominales y ordinales) y las cuantitativas, según sus características morfológicas. A partir de la matriz de distancia se combinan las soluciones para obtener una estructura consenso, en aras de lograr un mejor rendimiento y calidad de los resultados alcanzados por los algoritmos individuales. Para los análisis se utilizaron funciones implementadas sobre la base del lenguaje de programación R. En esta investigación se demostró que los algoritmos particionales estudiados, así como la combinación de agrupamientos para lograr una partición consenso son técnicas efectivas en la conformación de estructuras consistentes del genofondo de malanga. Este análisis puede ser aplicado a otros estudios de clasificación en germoplasma vegetal.

Palabras clave: algoritmo K-medias; conglomerado particional; consenso; germoplasma

COMBINATION OF PARTITIONAL CLUSTERS FOR THE CLASIFICATION OF COCOYAM (*XANTHOSOMA* SPP) GERMPLASM

ABSTRAT

In this work different partitional grouping methods are evaluated and implemented in the *kmeans* function in programming language R. This is done with the objective of identifying natural groups inherent to cocoyam germplasm conserved in Inivit. The matrix of data includes mixed variables, where qualitative variables (nominal and ordinal) and quantitative are integrated according to their morphological characteristics. The solutions are combined departing from the distance matrix, so as to obtain a consensus structure, for a better efficiency and quality of results of the individual algorithms. For the analysis functions were implemented in programming language R. In this research it was demonstrated that the partitional algorithms studied, as well as the combination of groupings for a consensus partition, are effective techniques for the formation of consistent structures of cocoyam germplasm. This analysis can be applied to other studies of classification in plant germplasm.

Keywords: K-means algorithm; partitional clusters; consensus; germplasm

INTRODUCCIÓN

Los bancos de germoplasma son esenciales para los programas de mejoramiento genético, si se desea desarrollar este último sobre bases sólidas y obtener un avance consistente en todos los lineamientos de mejoramiento que se persigan (Castillo, 1991). A lo largo de la historia, los recursos fitogenéticos han contribuido a la estabilidad de los agroecosistemas y han proporcionado la materia prima fundamental para el surgimiento del fitomejoramiento científico moderno.

Cuando se hace una colección de material germoplásmico, el paso obvio y necesario es hacer una descripción morfológica cualitativa y cuantitativa para su identificación y una evaluación adecuada del material genético (Enríquez, 1991).

Los métodos para el estudio morfológico se basan en el empleo de descriptores o caracteres cualitativos que se pueden observar a simple vista y cuantitativos que se pueden medir y se expresan en casi todos los ambientes. Estos métodos son relativamente económicos y constituyen la base de la caracterización de las plantas en los bancos de germoplasma (Silva, 1997; FAO, 1996). Una vez que se reúne la información se procede a su análisis estadístico mediante los métodos multivariados de descripción, ordenación o clasificación adecuados.

La clasificación puede considerarse una de las tareas fundamentales de la ciencia, es una acción propia del hombre dado que los objetos no están ordenados por sí mismos; toda clasificación se considera artificial ya que el hombre, como resultado de su capacidad de razonar e interpretar el mundo que le rodea, es el que asigna el orden a las entidades según sus ideas, necesidades y criterios (Kragh, 1989).

Los algoritmos de agrupación varían entre sí por el mayor o menor grado de reglas heurísticas que utilizan e inversamente, por el nivel de procedimientos formales involucrados. Todos ellos se basan en el empleo sistemático de las “distancias” entre los vectores (objetos a agrupar), así como entre los grupos que se van haciendo y deshaciendo a lo largo del proceso correspondiente.

En el Instituto de Investigaciones de Viandas Tropicales (Inivit) mantiene desde 1967, una colección de accesiones cultivadas de malanga del género *Xanthosoma* procedente de colectas, de introducciones y del programa de mejoramiento genético del cultivo. La experiencia adquirida en Cuba sobre el cultivo, la riqueza genética conservada en la isla, constituyen motivaciones importantes para profundizar en el conocimiento de tan importante género de plantas hacia el objetivo de garantizar el manejo sostenible del patrimonio genético existente, lo que incluye un mayor acercamiento a la adecuada ubicación taxonómica de todos los cultivares y a un mejor conocimiento de sus características y potencial genético, que permita lograr su empleo efectivo en los programas de mejoramiento; así como el perfeccionamiento de las estrategias de conservación de los recursos disponibles.

Además, dadas las circunstancias actuales de escasez de recursos financieros y materiales se hace necesario conservar y estudiar solo aquellos materiales que sean únicos, es decir, que no se encuentren duplicados y de esta forma facilitar que se profundice en los estudios morfológicos, bioquímicos y moleculares.

Para el mejoramiento genético de plantas, en cualquiera de sus formas, es indispensable partir de la mejor clasificación taxonómica posible ya que este conocimiento, optimiza los procesos de cruzamientos y los hace más eficientes.

En el Inivit, se han realizado estudios de técnicas estadísticas no supervisada de clasificación jerárquica y de reducción de dimensionalidad sobre el germoplasma de

malanga (*Xanthosoma* spp.) (Milián, 2008; Molina, 2013a; 2013b; 2015; 2017; 2018; Milián, 2018), a partir de los cuales se pretende desarrollar el presente trabajo con el objetivo de contribuir a una mejor clasificación específica del germoplasma de malanga *Xanthosoma* spp. aún ambigua a nivel internacional. El objetivo, es ampliar el estudio con la aplicación de técnicas estadísticas no supervisada de clasificación con métodos particionales, que integren las diferentes variables que componen las accesiones, para obtener una mayor comprensión sobre la distribución de las accesiones dentro del germoplasma del género *Xanthosoma* que se conservan en Cuba.

MATERIALES Y MÉTODOS

Para la investigación se utilizó la colección de trabajo del germoplasma de malanga (*Xanthosoma* spp.) que cuenta con 70 accesiones donde se evaluaron 20 variables cualitativas y 16 variables cuantitativas (Milián, 2008; 2018) del banco de germoplasma, que se conserva en el Instituto de Investigaciones de Viandas Tropicales (Inivit). Las 70 accesiones están distribuidas en 57 accesiones de la especie *Xanthosoma sagittifolium* (L.) Schott (SAG) y 10 accesiones que conforman un grupo que no tienen una especie definida, o cuya clasificación específica no está definida de *Xanthosoma* sp. (SP). Además, está presente la especie *Xanthosoma brasiliense* Engl. (BRA), cuyo único representante es el clon 'Belembe'; la especie *Xanthosoma atrovirens* Koch & Bouché (ATR), constituida por el clon 'Amarilla Trinidad'; mientras que la accesión 'Jardín' representa a la especie *X. nigrum* (Vell.) Mansf. (NIG) (Milián, 2008; 2018).

En un primer paso se integraron las variables cualitativas transformadas a través del análisis de correspondencia múltiple (Tenenhaus, 1985) con la función *MCA()* descrita en el paquete *FactoMineR*, con las cuantitativas normalizadas para lo cual se usó la función *data.Normalization(n1)* del paquete *clusterSim* (Walesiak, 2017).

Como medidas para evaluar las diferencias y similitudes entre accesiones se usó la distancia Euclidiana implementada en la función *dist()* descrita en el paquete *stats*.

A pesar que la investigación está centrada en una clasificación no supervisada, para estimar el número de agrupaciones, se usó la identificación ya establecida de algunas especies basada en investigaciones de los curadores del germoplasma en estudio de malanga (Milián, 2014).

Se usaron los métodos de aglomeración particionales *k-means* descrito en el paquete *stats* en la función *kmeans()* con los cuatro algoritmos: Forgy (Forgy, 1965), Hartigan-Wong (Hartigan, 1979), Lloyd (Lloyd, 1982) y MacQueen (MacQueen, 1967); que forma parte de la librería básica de R que se instala por defecto. Y los métodos *Pam* y *Clara* implementados en el paquete *cluster* (Maechler et al., 2019) en las funciones *pam()* y *clara()*.

Una vez obtenidos los resultados e interpretados es conveniente establecer un criterio de validación de la coherencia dentro de los conglomerados mediante los coeficientes devueltos con el Ancho de la Silueta (Rousseeuw, 1987; Kaufman y Rousseeuw, 1990) a través de la función *silhouette()* implementada en el paquete *cluster*. El índice proporciona una representación gráfica que da una medida de la similitud de una accesión con su propio conglomerado (cohesión) en comparación con otros conglomerados (separación). El valor de la silueta varía de -1 a 1, donde un valor alto muestra que el objeto está bien adaptado a su propio conglomerado y un valor bajo indica que está mal adaptado a los conglomerados vecinos.

Con el objetivo de lograr una mejor calidad de los resultados alcanzados por los algoritmos individuales y compensar posibles errores cometidos por un algoritmo individual se combinaron los resultados de los diferentes algoritmos de aglomeración particionales con la función *cl_ensemble()* y finalmente se obtiene la partición consenso, la cual es capaz de combinar toda la información existente de los diferentes resultados individuales, se obtuvo con la función *cl_consensus()* ambas descritas en el paquete *clue* utilizando siete de los métodos disponibles en esta función ("GV1", "SM", "SE", "GV3", "soft/symdiff" y "DWH", "HE") para calcular las particiones.

Para medir el desempeño de la clasificación de cada uno de los métodos particionales y el consensos, se usaron las matrices de confusión (Kohavi y Provost, 1998), tratadas posteriormente por Provost y Fawcett, 2001; Fawcett, 2003; Weiss y Provost, 2003; Maimon y Rokach, 2005 a través de la función *confusion.matrix()* implementadas en el paquete *clv* y el área bajo la curva ROC (AUC) (Swets, 1988), tratados más profundamente *a posteriori* por el propio Swets y otros autores (Swets *et al.*, 2000a; 2000b) con la *multiclass.roc()* descrita en el paquete *pROC*.

Las matrices de confusión contienen información acerca de los valores reales y las clasificaciones predichas hechas. El desempeño de un tal sistema es usualmente evaluado usando los datos en dicha matriz. Las curvas ROC (en inglés *Receiver Operating Characteristics*) conjuntamente con las matrices de confusión, constituyen otra manera de examinar el desempeño de una clasificación. El área bajo la curva ROC (AUC) se puede usar como una medida de la exactitud en muchas aplicaciones. Entre mayor área bajo la curva mejor clasificación.

El software utilizado para realizar los experimentos presentes en esta investigación es R versión 3.6.1 (R Development Core Team, 2019).

RESULTADOS Y DISCUSIÓN

Atendiendo a la naturaleza mixta de los datos las variables cualitativas (nominales y ordinales) se transformaron a través del análisis de correspondencia múltiple, se toman 31 dimensiones con un porcentaje de varianza mayor que uno, que representan el 91,176 % de la variabilidad total. Por otro lado, las 16 variables cuantitativas se estandarizaron y, posteriormente, se unieron ambos resultados en una sola matriz de datos continuos a la cual se le aplicó la distancia Euclidiana. Sorense (1948) y Gatna (2004) han demostrado que al menos el 70 % de la inercia total explicada es el valor de corte más adecuado.

A partir de la ejecución del método de k-medias con los cuatro algoritmos que emplea para realizar las agrupaciones y, los métodos *Pam* y *Clara* con la distancia Euclidiana, en un análisis global se puede observar en la tabla 1 la distribución de las accesiones en los cinco conglomerados (los cinco grupos que corresponden con el número de especies presente en la base de datos). En especial las 10 accesiones de *Xanthosoma* sp. en las cuales la clasificación específica no está definida (Milián, 2018), ocho (1232-SP, 1238-SP, 1243-SP, 1255-SP, 1261-SP, 1464-SP, 1475-SP y 1480-SP) siempre comparten el mismo conglomerado independientemente del método de aglomeración, o sea nunca se separan, adicionalmente siempre están acompañadas de dos accesiones de la especie *X. sagittifolium* (906-SAG, 1251-SAG) y de la única representación de la especie *X. brasiliense* (1245-BRA), lo cual muestra una medida de similitud entre ellas. Las dos accesiones (1443-SP y 1466-SP) nunca se unen al resto de la especie *Xanthosoma* sp., por lo cual se puede afirmar que no tienen una afinidad morfológica

demostrada a través de estos métodos de clasificación; están más relacionadas a las especies *X. nigrum*, *X. atrovirens* y *X. sagittifolium*.

Tabla 1. Salida en consola de accesiones por conglomerados de los algoritmos de aglomeración particionales.

Algoritmos	Accesiones por conglomerados
<i>Hartigan</i>	1: {905-SAG, 1229-ATR, 1230-SAG, ..., 1594-SAG} 2: {1233-SAG, 1242-SAG, 1244-SAG, 1262-SAG, 1443-SP , 1446-SAG, 1466-SP } 3: {906-SAG, 1232-SP , 1238-SP , 1239-SAG, 1241-SAG, 1243-SP , 1245-BRA, 1251-SAG, 1255-SP , 1259-SAG, 1261-SP , 1267-SAG, 1464-SP , 1475-SP , 1480-SP , 1486-SAG} 4: {1235-SAG, ..., 1490-SAG} 5: {904-SAG, 1231-SAG, 1236-NIG, 1247-SAG, ..., 1610-SAG}
<i>Lloyd</i>	1: {1233-SAG, 1235-SAG, 1237-SAG, 1242-SAG, 1244-SAG, 1246-SAG, 1252-SAG, 1262-SAG, 1265-SAG, 1421-SAG, 1443-SP , 1445-SAG, 1446-SAG, 1460-SAG, 1466-SP , 1471-SAG, 1478-SAG, 1481-SAG, 1490-SAG} 2: {1229-ATR, 1234-SAG, ..., 1436-SAG} 3: {906-SAG, 1232-SP , 1238-SP , 1239-SAG, 1241-SAG, 1243-SP , 1245-BRA, 1251-SAG, 1255-SP , 1259-SAG, 1261-SP , 1267-SAG, 1464-SP , 1475-SP , 1480-SP , 1486-SAG} 4: {905-SAG, ..., 1594-SAG} 5: {904-SAG, 1231-SAG, 1236-NIG, 1247-SAG, ..., 1610-SAG}
<i>Forgy</i>	1: {905-SAG, 1229-ATR, 1230-SAG, ..., 1594-SAG} 2: {1239-SAG, ..., 1486-SAG} 3: {906-SAG, 1232-SP , 1238-SP , 1243-SP , 1245-BRA, 1251-SAG, 1255-SP , 1261-SP , 1464-SP , 1475-SP , 1480-SP } 4: {1233-SAG, 1235-SAG, 1237-SAG, 1242-SAG, 1246-SAG, 1252-SAG, 1262-SAG, 1265-SAG, 1421-SAG, 1443-SP , 1445-SAG, 1446-SAG, 1460-SAG, 1466-SP , 1471-SAG, 1478-SAG, 1481-SAG, 1490-SAG} 5: {904-SAG, 1231-SAG, 1236-NIG, 1247-SAG, ..., 1610-SAG}
<i>MacQueen</i>	1: {1233-SAG, 1235-SAG, 1237-SAG, 1242-SAG, 1244-SAG, 1246-SAG, 1252-SAG, 1262-SAG, 1265-SAG, 1421-SAG, 1443-SP , 1445-SAG, 1446-SAG, 1460-SAG, 1466-SP , 1471-SAG, 1478-SAG, 1481-SAG, 1490-SAG} 2: {1231-SAG, ..., 1485-SAG} 3: {905-SAG, 1229-ATR, 1230-SAG, ..., 1594-SAG} 4: {904-SAG, 1236-NIG, 1248-SAG, ..., 1610-SAG} 5: {906-SAG, 1232-SP , 1238-SP , 1239-SAG, 1241-SAG, 1243-SP , 1245-BRA, 1251-SAG, 1255-SP , 1259-SAG, 1261-SP , 1267-SAG, 1464-SP , 1475-SP , 1480-SP , 1486-SAG}
<i>Pam</i>	1: {904-SAG, 1231-SAG, 1236-NIG, 1247-SAG, ..., 1610-SAG} 2: {905-SAG, 1229-ATR, 1230-SAG, ..., 1594-SAG} 3: {906-SAG, 1232-SP , 1238-SP , 1243-SP , 1245-BRA, 1251-SAG, 1255-SP , 1261-SP , 1464-SP , 1475-SP , 1480-SP } 4: {1233-SAG, 1235-SAG, 1237-SAG, 1242-SAG, 1246-SAG, 1262-SAG, 1265-SAG, 1421-SAG, 1443-SP , 1445-SAG, 1446-SAG, 1460-SAG, 1466-SP , 1471-SAG, 1478-SAG, 1481-SAG, 1490-SAG} 5: {1239-SAG, ..., 1486-SAG}
<i>Clara</i>	1: {904-SAG, 1231-SAG, 1236-NIG, 1247-SAG, ..., 1610-SAG} 2: {905-SAG, 1229-ATR, 1230-SAG, ..., 1594-SAG} 3: {906-SAG, 1232-SP , 1238-SP , 1243-SP , 1245-BRA, 1251-SAG, 1255-SP , 1261-SP , 1464-SP , 1475-SP , 1480-SP } 4: {1233-SAG, 1235-SAG, 1237-SAG, 1242-SAG, 1246-SAG, 1262-SAG, 1265-SAG, 1421-SAG, 1443-SP , 1445-SAG, 1446-SAG, 1460-SAG, 1466-SP , 1471-SAG, 1478-SAG, 1481-SAG, 1490-SAG} 5: {1239-SAG, ..., 1486-SAG}

Además, se puede observar en esta misma tabla que los métodos *Pam* y *Clara* coinciden en las accesiones por conglomerados.

El objetivo general perseguido por las técnicas de clustering consiste en identificar grupos o clusters compactos, pero también se hace necesario la evaluación para medir la eficiencia de las técnicas de clustering aplicado. Para el caso de evaluación de clusters en este trabajo se utiliza el concepto de silueta de un clúster y la entrega de sus coeficientes de medición, donde un valor de silueta del individuo cercano a 1 (eje Y), quiere decir que el miembro del grupo tiene mayor afinidad al clúster asignado. Por el contrario, si la silueta del miembro del clúster está cercana a 0 (eje Y), quiere decir que tiene afinidad morfológica con otro grupo. Por otra parte, si la silueta del individuo es menor a 0 (eje Y), la probabilidad de que el miembro del clúster pertenezca a otro grupo es muy alta.

Teniendo en cuenta los resultados obtenidos por los métodos de agrupamiento, es conveniente aglutinarlos en aras de lograr una mejor calidad de los resultados alcanzados por los algoritmos individuales, hacer un conjunto de particiones diversas es recomendable, ya que mientras más variado es el conjunto de particiones, mayor es la información disponible en la etapa de combinación.

La combinación de agrupamientos para lograr una partición consenso se realizó con el algoritmo “GV3” de los siete disponibles para calcular los conglomerados consenso en los algoritmos particionales, ya que fue el que mayor área bajo la curva (*Multi-class area under the curve*: 0.8711). En la tabla 2 se puede observar la distribución de las accesiones por clústeres.

Tabla 2. Salida en consola de las accesiones por conglomerados, método consenso con el algoritmo “SV3”.

Algoritmo consenso	Accesiones por conglomerados
SV3	1: {1239-SAG, 1241-SAG, 1244-SAG, 1259-SAG, 1267-SAG, 1486-SAG} 2: {1233-SAG, 1235-SAG, 1237-SAG, 1242-SAG, 1246-SAG, 1252-SAG, 1262-SAG, 1265-SAG, 1421-SAG, 1443-SP , 1445-SAG, 1446-SAG, 1460-SAG, 1466-SP , 1471-SAG, 1478-SAG, 1481-SAG, 1490-SAG} 3: {904-SAG, 1231-SAG, 1236-NIG, 1247-SAG, 1248-SAG, 1254-SAG, 1256-SAG, 1258-SAG, 1264-SAG, 1266-SAG, 1422-SAG, 1433-SAG, 1450-SAG, 1457-SAG, 1470-SAG, 1474-SAG, 1485-SAG, 1593-SAG, 1610-SAG} 4: {905-SAG, 1229-ATR, 1230-SAG, 1234-SAG, 1240-SAG, 1249-SAG, 1250-SAG, 1253-SAG, 1257-SAG, 1263-SAG, 1427-SAG, 1428-SAG, 1435-SAG, 1436-SAG, 1482-SAG, 1594-SAG} 5: {906-SAG, 1232-SP , 1238-SP , 1243-SP , 1245-BRA, 1251-SAG, 1255-SP , 1261-SP , 1464-SP , 1475-SP , 1480-SP }

En la figura 1 se puede observar que el ancho de silueta para el consenso es de 0.29. En el clúster cinco con 0.28 de ancho de silueta están las accesiones de la especie SP (1232-SP, 1238-SP, 1243-SP, 1255-SP, 1261-SP, 1464-SP, 1475-SP y 1480-SP), haciendo una evaluación de la información sobre los coeficientes de silueta individuales (Tabla 3) se puede afirmar que son accesiones que están bien clasificadas con afinidad morfológica entre ellas. Las accesiones 1443-SP y 1466-SP pertenecen al grupo dos, pero el ancho de silueta de esta primera es de -0.037, lo que indica que esta observación está mal clasificada y tienen una alta probabilidad de pertenecer al grupo

uno. En la figura 2 se muestran los miembros que conforman cada uno de los cinco conglomerados según similitudes morfológicas.

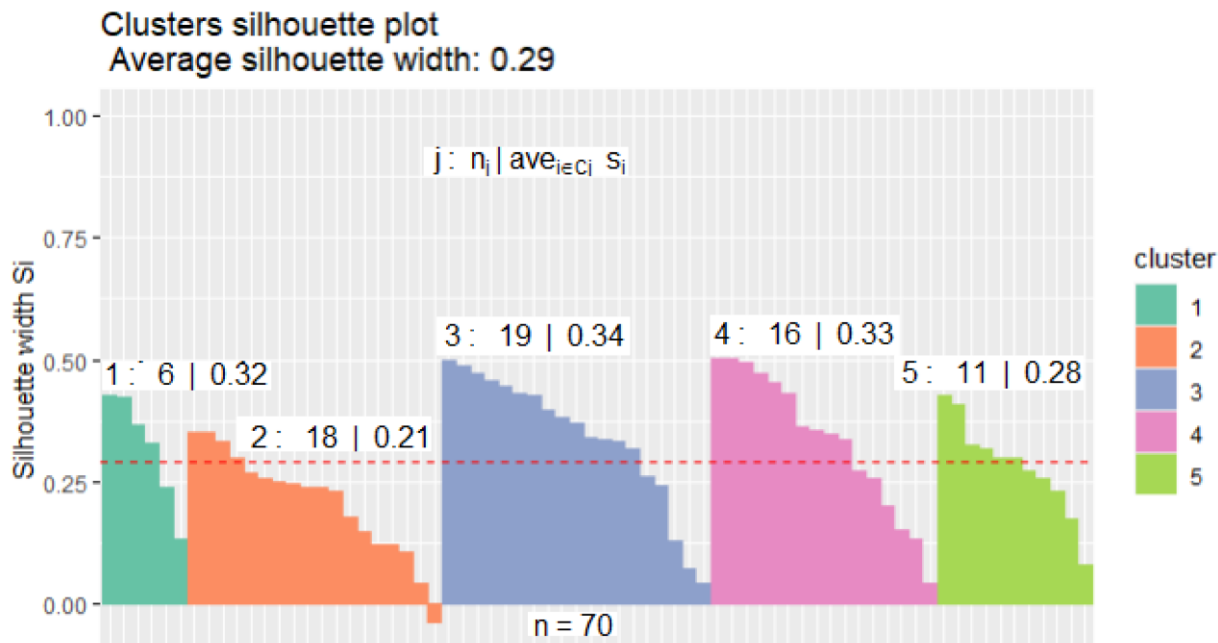


Figura 1. Ancho de silueta de miembros de cada clúster. Método consenso con el algoritmo “SV3”

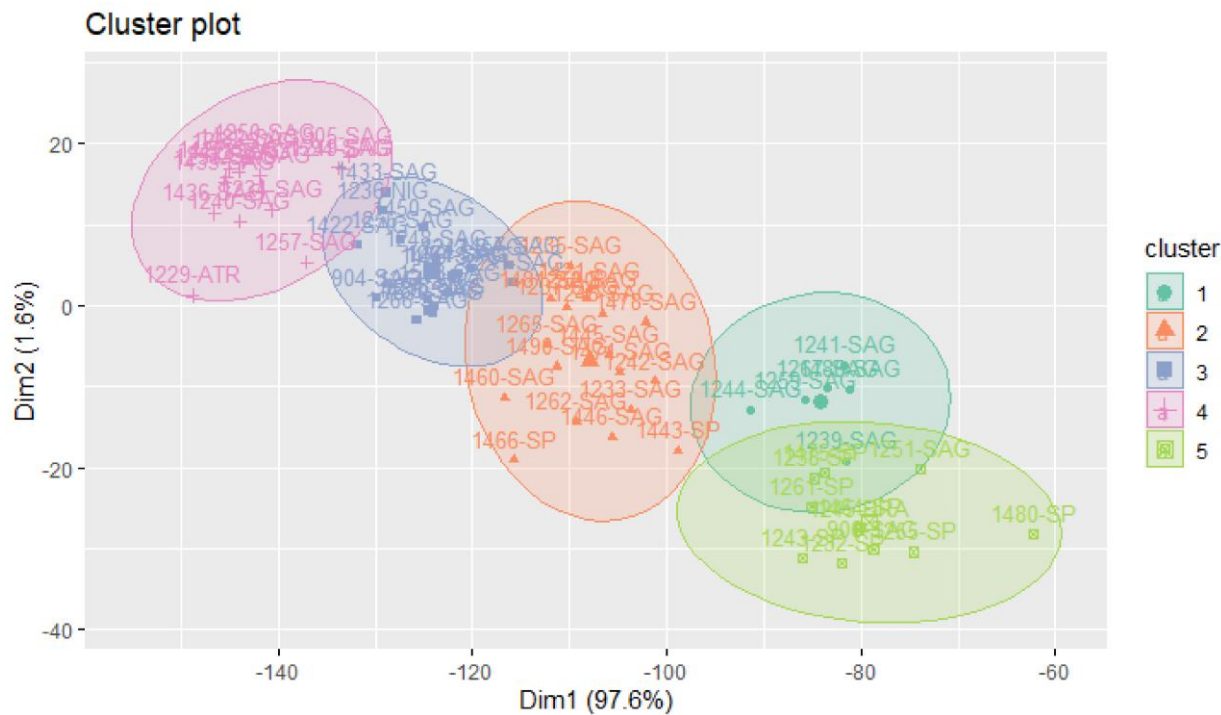


Figura 2. Visualización para cinco clústeres identificando accesiones. Método consenso con el algoritmo “SV3”

Tabla 3. Ancho de silueta individuales clúster dos y cinco. Método consenso con el algoritmo “SV3”

ID.cultivares	cluster	neighbor	sil_width
1233-SAG	2	1	0,267
1235-SAG	2	3	0,120
1237-SAG	2	3	0,107
1242-SAG	2	1	0,179
1246-SAG	2	3	0,229
1252-SAG	2	3	0,043
1262-SAG	2	1	0,349
1265-SAG	2	3	0,248
1421-SAG	2	3	0,259
1443-SP	2	1	-0,037
1445-SAG	2	1	0,332
1446-SAG	2	1	0,236
1460-SAG	2	3	0,246
1466-SP	2	3	0,120
1471-SAG	2	1	0,349
1478-SAG	2	3	0,237
1481-SAG	2	3	0,147
1490-SAG	2	3	0,296
906-SAG	5	1	0,081
1232-SP	5	1	0,427
1238-SP	5	1	0,231
1243-SP	5	1	0,300
1245-BRA	5	1	0,408
1251-SAG	5	1	0,174
1255-SP	5	1	0,318
1261-SP	5	1	0,324
1464-SP	5	1	0,296
1475-SP	5	1	0,256
1480-SP	5	1	0,271

En todos los análisis, las accesiones de *Xanthosoma* sp. (SP) (1232-SP, 1238-SP, 1243-SP, 1255-SP, 1261-SP, 1464-SP, 1475-SP y 1480-SP) se mantienen unidas y siempre su conglomerado cercano más probable es el mismo para las ocho y el ancho de silueta es consistente, por lo cual se puede deducir que todas pertenecen a la misma especie. Las accesiones 1443-SP y 1466-SP siempre son ubicadas en el mismo conglomerado, sin embargo, el clúster más probable es sólo igual en el algoritmo Hartigan y los índices de siluetas individuales son muy pobres especialmente para la accesión 1443-SP.

CONCLUSIONES

1. La combinación de agrupamientos para lograr una partición consenso es una técnica efectiva en aras de lograr una mejor calidad de los resultados alcanzados por los algoritmos individuales sobre la clasificación del germoplasma de malanga (*Xanthosoma* spp.) y se debe tener en cuenta como una posible solución para mejorar la clasificación de especies vegetales.
2. El uso del ancho de silueta para evaluar la clasificación, tanto individual como por conglomerados, es efectiva en la identificación de la afinidad morfológica entre los miembros de un mismo conglomerado y entre éstos.
3. Se amplía el conocimiento sobre la combinación técnicas de agrupamiento particionales en la clasificación de bancos de germoplasma vegetal.

RECOMENDACIONES

Teniendo en cuenta estos resultados sería favorable realizar análisis moleculares sobre accesiones de la especie SP en futuros estudios de caracterización para determinar si son una nueva especie dentro de este género.

BIBLIOGRAFÍA

- CASTILLO, R. 1991. Nuevos Departamentos de Recursos Fitogenéticos en Ecuador. *Diversity*, 7(1-2): p.37-39.
- ENRÍQUEZ, G. 1991. Relación de los recursos fitogenéticos con otras ciencias. En: Castillo, R.; J. Estrella y C. Tapia (eds). Técnicas para el manejo y uso de los recursos fitogenéticos. Departamento de Recursos Fitogenéticos. Instituto Nacional de Investigaciones Agropecuarias, Quito, Ecuador. p. 314.
- FAO.1996. Informe sobre el Estado de los Recursos Fitogenéticos en el Mundo. Preparado para la Conf. Técn. Internac. sobre los Recursos Fitogenéticos. Leipzig, Alemania. 75p.
- FORGY, E. 1965. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*. 21:768-769.
- FAWCETT, T. 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Hewlett-Packard Company, 28p.
- GATNAR, E. and M. WALESIK (eds.) 2004. Metody statystycznej analizy wielowymiarowej w badaniach marketingowych [Multivariate statistical analysis methods in marketing research]: Wrocław.
- HARTIGAN, J. and M. WONG. 1979. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28: p.100-108.
- KRAGH, H. 1989. Introducción a la Historia de la Ciencia. Crítica, Barcelona, 281p.
- KAUFMAN, L. and P. ROUSSEEUW. 1990. Finding Groups in Data: An introduction to cluster analysis. Wiley, New York.
- KOHAVER, R. and F. PROVOST. 1998. Glossary of Terms. Editorial for the Special Issue on Applications Machine Learning and the Knowledge Discovery Process, 30(2-3).
- LLOYD, S. 1982. Least squares quantization in PCM. Technical Note, Bell Laboratories. In *IEEE Transactions on Information Theory*, 28:128-137.
- MILIÁN, M. 2008. 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, 123p.

- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ; C. PONS y R. GRAU. 2013a. Clasificación y determinación del número óptimo de conglomerados en bancos de germoplasma. *Revista Centro Agrícola*, 40(4):19-24.
- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ; C. PONS y R. GRAU. 2013b. Análisis taxonómico con variables mixtas en accesiones de malanga (*Xanthosoma* spp.) y plátano (*Musa* spp.). *Revista Centro Agrícola*, 40(4):7-10.
- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ y C. PONS. 2015. Métodos de estandarización en colecciones de germoplasma vegetal. *Revista Agricultura Tropical*, 1(2):67-73.
- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ y C. PONS. 2017. Taxobanger v.1.0: Aplicación informática en R para el análisis taxonómico en bancos de germoplasma vegetal. *Revista Cuba Ciencias Informáticas*, 11(3):159-173.
- MOLINA, O.; R. GARCÍA; M. MILIÁN; L. GONZÁLEZ y C. PONS. 2018. Evaluación de técnicas para la clasificación del genofondo de raíces, rizomas y tubérculos tropicales, plátanos y bananos en Cuba. Informe del Proyecto PAH P131LH001027. Instituto de Investigaciones de Viandas Tropicales (Inivit), Minag, Cuba. 146p.
- MILIÁN, M.; O. MOLINA and Y. FIGUEROA. Integrated Characterization of Cuban Germplasm of Cocoyam (*Xanthosoma Sagittifolium* (L.) Schott). *Journal of Plant Genetics and Crop Research*. 2018;1(1):1-18.
- MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, editors. LE Cam LM, Neyman J. Berkeley, CA: University of California Press. 281-297.
- MAECHLER, M.; P. ROUSSEEUW; A. STRUYF; M. HUBERT and K. HORNIK. 2019. Cluster: cluster Analysis Basics and Extensions. R package version 2.1.0. Available from: <https://cran.r-project.org/web/packages/cluster/index.html>.
- MAIMON, O. and ROKACH, L. 2005. Data Mining and Knowledge Discovery Handbook. Springer US.
- ROUSSEEUW, P. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*, 20:53-65.
- PROVOST, F. and T. FAWCETT. 2001. Robust classification for imprecise environments. *Machine Learning*, 42:203-231.
- SILVA, S.; A. PIRES DE MATOS; E. ALVES y K. SHEPHERD. 1997. Mejoramiento de bananos diploides (AA) en EMBRAPA/CNPMPF. *Infomusa*, 6(2):4-6.
- SORENSEN, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of vegetation on Danish commons. *Biologiske Skrifter*, 5, 1-34.
- SWETS, J. 1988. Measuring the accuracy of diagnostic systems. *Science*, 240:1285-1293.
- SWETS, J.; R. DAWES and J. MONAHAN. Better decisions through science. 2000a; 82-87. Consultado: 24 de octubre de 2020. Disponible en: <http://www.psychologicalscience.org/newsresearch/publications/journals/%siam.pdf>.
- SWETS, J.A.; R. DAWES and J. MONAHAN. 2000b. Psychological science can improve diagnostic decisions. *Psy-Chological Science in the Public Interest*, 1(1):1-26.
- R Development Core Team. 2019. R: A language and environment for statistical computing (version 3.6.1). R Foundation for Statistical Computing. Vienna, Austria. Available from: <http://www.r-project.org/>.

- TENENHAUS, M. and F. YOUNG. 1985. An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data. *Psychometrika*, 50(91).
- WALESIK, M. and A. DUDEK. 2017. ClusterSim: Searching for optimal clustering procedure for a data set. Consultado: 12 de diciembre de 2017. Disponible en: <http://cran.fhcrc.org/web/packages/clusterSim/clusterSim.pdf>.
- WEISS, G.M. and F. PROVOST. 2003. Learning when training data are costly: The effect of class distribution on tree induction. *JAIR*, 19:315-354.