Rev. FCA UNCuyo. Tomo 42. N° 2. Año 2010. 183-200.

Ordenaciones de material genético a partir de información multidimensional

Ordinations of genetic data from multidimensional markers

Cecilia Bruno
Mónica Balzarini

Originales: Recepción: 02/08/2009 - Aceptación: 18/03/2010

RESUMEN

Nuevas biotecnologías permiten obtener información para caracterizar materiales genéticos a partir de múltiples marcadores, ya sean éstos moleculares y/o morfológicos. La ordenación del material genético a través de la exploración de patrones de variabilidad multidimensionales se aborda mediante diversas técnicas de análisis multivariado. Las técnicas multivariadas de reducción de dimensión (TRD) y la representación gráfica de las mismas cobran sustancial importancia en la visualización de datos multivariados en espacios de baja dimensión ya que facilitan la interpretación de interrelaciones entre las variables (marcadores) y entre los casos u observaciones bajo análisis. Tanto el Análisis de Componentes Principales, como el Análisis de Coordenadas Principales y el Análisis de Procrustes Generalizado son TRD aplicables a datos provenientes de marcadores moleculares y/o morfológicos. Los Árboles de Mínimo Recorrido y los biplots constituyen técnicas para lograr representaciones geométricas de resultados provenientes de TRD. En este trabajo se describen estas técnicas multivariadas y se ilustran sus aplicaciones sobre dos conjuntos de datos, moleculares y morfológicos, usados para caracterizar material genético fúngico.

ABSTRACT

New biotechnologies allow to obtain information for genetic characterization from multiple molecular and/or morphological markers. The ordination of genetic material through the exploration of variability patterns is addressed by multivariate methods. Dimension reduction techniques (DRT) and graphic representation are crucial to visualize multivariate data in low-dimensional spaces since it facilitate the understanding of relationships among individuals and/or markers. The principal component analysis, the multidimensional scaling and generalized procrustes analysis are dimension reduction techniques applicable to data from molecular and/or morphological markers. The minimum spanning trees and the biplots are techniques that allow geometrical representations of results from DRT. This paper describes such techniques and its applications are illustrated on a dataset with molecular and morphological markers characterizing fungus genetic material.

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Estadística y Biometría. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba. Av. Valparaíso s/n. Ciudad Universitaria. C. C. 509. (5000) Córdoba. Argentina. mbalzari@agro.unc.edu.ar

Palabras clave

análisis de componentes principales • escalamiento multidimensional • procrustes • árboles de recorrido mínimo • biplots

Keywords

principal components analysis • multidimensional scaling • generalized procrustes analysis • minimum spanning tree • biplots

INTRODUCCIÓN

El ordenamiento de taxones basado en la caracterización molecular y/o morfológica de material genético, realizado a partir de datos de múltiples marcadores, se optimiza cuando la descripción marcador a marcador se complementa con el estudio de relaciones o asociaciones entre marcadores y entre observaciones. Numerosas herramientas de la estadística multivariada permiten el análisis de observaciones multidimensionales, *i.e*. aquellas caracterizadas por múltiples marcadores (6).

Para la representación de datos mutidimensionales en un espacio de 2 ó 3 dimensiones, cobran especial importancia las técnicas de reducción de dimensión (TRD), las cuales permiten explorar las relaciones existentes entre el material genético mediante ordenaciones del mismo sobre planos que, bajo distintos criterios de representación, son "óptimos" para ordenar las observaciones y analizar interdependencias. Las TRD son útiles para:

1) resumir y graficar los datos multivariados,
2) explorar tendencias y relaciones entre observaciones, entre marcadores y entre observaciones y marcadores,
3) agrupar y clasificar observaciones y/o marcadores,
4) identificar relaciones importantes para etapas posteriores de modelación estadística.

En esencia, los métodos de ordenación extraen sucesivos componentes desde una matriz de similitudes (o distancias) entre las observaciones o casos en estudio calculada a partir de múltiples marcadores. Esos componentes son usados como ejes para la representación gráfica de los objetos. En la ordenación, cada individuo es ubicado sobre uno o más ejes tal que la posición geométrica relativa entre todos ellos refleja las similitudes y/o distancias correspondientes.

Las TRD usadas con fines exploratorios no requieren de supuestos distribucionales, por ejemplo datos que ajusten a una distribución normal. La característica de distribución libre las hace especialmente apropiadas para su utilización sobre información derivada de secuencias de fragmentos de nucleótidos, o marcadores de ADN, dado que éstos rara vez siguen una distribución normal debido al sesgo selectivo que se introduce cuando se usa un *primer* o un conjunto de enzimas de restricción particular sobre un genoma. En general, las ordenaciones basadas en matrices de distancias multivariadas son utilizadas cuando se desea relacionar material genético perteneciente a un único taxón, *i.e.*, donde la variación está más cercana a ser continua o semicontinua, sin amplios quiebres, como puede ocurrir cuando se involucran diferentes especies.

En este trabajo se describen TRD aplicables a la ordenación de material genético a partir de datos de marcadores moleculares y/o morfológicos. Se discute, a través de la ilustración sobre un conjunto de datos reales de marcadores, la pertinencia e interpretación de distintas estrategias de representación gráfica: biplots y árboles de mínimo recorrido sobre las ordenaciones producidas por los análisis de componentes principales, coordenadas principales y procrustes generalizados. Los ejemplos contrastan la aplicación de ordenaciones con datos de naturaleza cuantitativa y cualitativa.

Distancias y similitudes

Para ordenar las observaciones es necesario primero definir una métrica que represente similitudes y/o distancias entre pares de ellas. Para datos de marcadores moleculares codificados como presencia/ausencia (datos binarios) (2) suelen usarse medidas de distancia basadas en índices de similitud (tabla 1). Para marcadores morfológicos, de naturaleza continua, generalmente se usan distancias producidas por diferentes órdenes de la métrica de Minkowski, siendo la distancia Euclídea una de las principales.

Tabla 1. Medidas de distancia e índices de similitud.
Table 1. Distance metrics and similarity indexes.

Naturaleza del dato del marcador	Métrica	Expresión
Continua	Minkowski	$d_{ij} = \left[\sum_{k=1}^{m} \left| y_{ik} - y_{jk} \right|^{r} \right]^{1/r}$
	City Block o Manhattan (Minkowsi con r=1)	$d_{ij} = \sum_{k=1}^{m} \left| y_{ik} - y_{jk} \right|$
	Euclídea (Minkowski con r=2)	$d_{ij} = \left[\sum_{k=1}^{m} \left| y_{ik} - y_{jk} \right|^{2} \right]^{1/2}$
Binaria [1, 2]	Dice	$S_{ij} = 2a/(2a+b+c)$
	Jaccard	$S_{ij} = a/(a+b+c)$
	Emparejamiento Positivo	$S_{ij} = (a+d)/(a+b+c+d)$

[1] Los índices de similitud (S_{ij}) son llevados a medidas de distancia mediante transformaciones, usualmente $(1-S_{ij})^{1/2}$, donde S_{ij} representa la similitud entre el individuo *i* y *j*.
[2] a, b, c, y d indican las frecuencias absolutas de los eventos (1,1), (1,0), (0,1) y (0,0) respectivamente, que surgen al comparar el perfil de marcadores de dos individuos, donde 1 representa la presencia del marcador y 0 la ausencia.

[1] Similarity indices (Sij) are taken to distance measures by transformations, usually (1-Sij)1/2, where Sij is the similarity between individuals i and j.
[2] a, b, c, y d indicate the absolute frequency of events (1,1), (1,0), (0,1) and (0,0) respectively, which arise when comparing the profile of markers of two individuals , where 1 represents presence of marker and 0 absence.

Dado que las diferencias entre las observaciones multivariadas generan variación, el análisis de la variabilidad total contenida en la matriz de datos provee información útil para la ordenación. Esta variabilidad puede ser capturada por matrices $n \times n$, donde n es el número de observaciones, cuyos elementos representan las distancias entre cada par de observaciones, o bien, por matrices de varianzas $m \times m$, donde m es el número de marcadores. El siguiente esquema muestra dos estructuras diferentes de datos multivariados: una discreta binaria y otra continua.

Esquema. Representación de la codificación de un gel de amplificación molecular en una matriz de datos binarios y la obtención de una matriz de distancia a partir de ellos. La distancia fue calculada como $(1\text{-}S_{ij})^{1/2}$, donde S_{ij} representa la similitud de Dice entre el individuo *i* y *j* (arriba) y matriz de distancias Euclídeas obtenidas a partir de datos provenientes de marcadores morfológicos (abajo).

Scheme. Representation of the encoding of a gel molecular amplification in a binary data matrix and obtaining a distance matrix from them. The distance was calculated as $(1\text{-}S_{ij})^{1/2}$, where S_{ij} represents the Dice's similarity between individuals *i* and *j* (above) and Euclidean distance matrix based on data obtained from morphological marker (below).

TÉCNICAS DE REDUCCIÓN DE DIMENSIÓN PARA ORDENAMIENTO DE OBSERVACIONES MULTIVARIADAS

Análisis de componentes principales (ACP)

Tiene como objetivo la transformación de un conjunto de variables continuas, o al menos ordinales, que originalmente pueden estar correlacionadas, en un grupo de variables no correlacionadas denominadas componentes principales que serán usadas como ejes para generar un plano que permita representar las observaciones de manera tal que en ese plano se represente lo mejor posible la variabilidad de las variables originales en el espacio multidimensional. Las componentes se ordenan según los niveles de información (variabilidad que expresan los datos sobre éstas). Algebraicamente el ACP busca una base ortogonal de los datos de manera tal que el primer eje se encuentre en la dirección de mayor variación y los ejes subsecuentes maximicen la explicación de la varianza total remanente condicionados a que sean ortogonales a sus ejes previos (es decir, cada eje aporta nueva información sobre la variabilidad total). El método opera sobre una matriz de varianzas-covarianzas (**S**) entre variables ($m \times m$) preservando las distancias Euclídeas entre observaciones. Los datos pueden o no ser estandarizados. Si se estandariza, el ACP opera sobre la matriz de correlación (**R**) entre variables. La técnica de estandarización se recomienda para situaciones donde las variables (marcadores) no son conmensurables (se encuentran en distintas escalas).

El ACP podría aplicarse, además, sobre una matriz $n \times n$ de covarianzas (o correlaciones) entre observaciones. El ACP sobre la matriz $m \times m$ provee un ordenamiento de las observaciones, mientras que el ACP sobre la matriz $n \times n$ otorga un ordenamiento de las variables.

Luego, si **X** es una matriz $n \times m$, de datos de n observaciones (material genético) sobre los cuales se registran m variables (marcadores), la solución del ACP obtenida mediante la descomposición espectral de la matriz **X'X** ($m \times m$), que contiene la información de la matriz de varianzas y covarianzas de los marcadores, es:

$$\mathbf{X'X} = \sum_{j=1}^{m} \lambda_j \mathbf{e_j} \mathbf{e_j}' = \mathbf{E} \mathbf{D}_\lambda \mathbf{E'}$$

donde:

$\mathbf{e}_j$ es el j-ésimo autovector

λ_j es el j-ésimo autovalor de la descomposición espectral de **X' X**

E es la matriz que contiene todos los autovectores

$\mathbf{D}_\lambda$ es una matriz diagonal cuyos elementos no nulos son los autovalores

La otra solución puede obtenerse de manera análoga, a partir de la descomposición espectral de la matriz **XX'** ($n \times n$). Las componentes principales se construyen a partir de los autovectores de estas descomposiciones de la siguiente forma:

$$CP_j = \mathbf{e'}_j \mathbf{X} = \mathbf{e}_{1j} \mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{e}_{mj} \mathbf{X}_m$$

Es decir, cualquier componente principal es una combinación lineal de las *m* variables originales ponderadas por los autovectores. La varianza de la componente es $Var\,(CP_j) = \lambda_j$.

Análisis de coordenadas principales (ACoorP)

El análisis de coordenadas principales es utilizado para mostrar las relaciones entre las observaciones multidimensionales de naturaleza discreta, continua, o ambas. Se definen distancias (o similitudes) en función de la naturaleza de la variable. Por ejemplo, si son continuas se suele seleccionar la distancia Euclídea; si son binarias la distancia de Emparejamiento Simple o Dice y si existen ambos tipos de variables, se suele utilizar la distancia de Gower (4):

$$S_{ij} = \frac{\sum_{m=1}^{M} w_{ijm} s_{ijm}}{\sum_{m=1}^{M} w_{ijm}}$$

donde:

S_{ij} es la similitud entre la observación *i* y *j*

M es el número total de variables, ya sean marcadores moleculares o marcadores morfológicos

W_{ijm} es la ponderación de cada variable ente las observaciones *i* y *j*

Si la variable es del tipo binaria (proveniente de información molecular), la similitud S_{ijm} vale 0 si $i \neq j$ y 1 si $i = j$; si la variable es cuantitativa (información morfológica) esta similaridad estará dada por:

$$s_{ijm} = 1 - \frac{| x_{im} - x_{jm} |}{r_m}$$

donde r_m es el rango del carácter *m*

El objetivo es ordenar las observaciones en un espacio de baja dimensión tal que las distancias o similitudes sean preservadas tanto como sea posible respetando las existentes en el espacio multidimensional. El ACoorP es una forma de escalamiento multidimensional métrico o clásico. Esta técnica opera sobre la matriz **Q** derivada de un doble proceso de centrado de la matriz de similitudes (o distancias) **A**, tal que el elemento *ij*-ésimo es

$$Q_{ij} = A_{ij} - \bar{A}_{i.} - \bar{A}_{.j} + \bar{A}_{..}$$

donde:

A_{ij} es la similitud entre las observaciones *i* y j

$\bar{A}_{i.}$ es la media de las similitudes para la fila *i*

$\bar{A}_{.j}$ es la media de las similitudes para la columna *j*

$\bar{A}_{..}$ es la media general de las similitudes en A

El criterio de optimalidad implica la extracción de un conjunto de ejes ortogonales desde la descomposición espectral de **Q**:

$$\mathbf{Q = ED}\lambda\ \mathbf{E'}$$

Los autovalores, elementos de la diagonal de $\mathbf{D}_\lambda$, expresan la variabilidad de los datos explicada por cada dimensión. Como los autovalores se ordenan en forma decreciente, los dos primeros ejes (coordenadas principales) explican la mayor cantidad de variación en **Q** que puede representarse en un espacio bidimensional. Las columnas de $\mathbf{Z = ED}^{1/2}_{\lambda}$ forman las coordenadas principales.

Procrustes como técnicas para consensuar ordenaciones

Cuando las observaciones son caracterizadas mediante $k \geq 2$ conjuntos de variables (marcadores) puede ser de interés obtener una ordenación para cada conjunto y luego evaluar el consenso de tales ordenamientos. Las variables en estos conjuntos pueden ser de igual o diferente naturaleza. Las ordenaciones obtenidas para cada tipo de marcador pueden ser usadas para lograr una única configuración de las observaciones.

La cuantificación del consenso mediante análisis procrustes generalizado (APG) provee información acerca de la armonización o adecuación de las configuraciones producidas por cada conjunto de variables (1). El APG se basa en rotaciones y escalamientos de las ordenaciones individuales para su representación en un mismo espacio (espacio de consenso). Gower (5) describe la configuración final como configuración de consenso y propone una técnica de cálculo que produce, en el formato de un análisis de la varianza, una medida para cuantificar el consenso.

Las sucesivas transformaciones que se realizan en un APG incluyen normalización, rotación, traslación y escalamiento de los datos bajo las premisas:

1) que se mantengan las distancias entre los individuos de las configuraciones individuales,

2) que se minimice la suma de cuadrados entre puntos análogos (provenientes de distintas configuraciones pero para el mismo individuo) y su centroide.

La configuración de consenso se obtiene como la media de las configuraciones individuales apropiadamente transformadas. Para ello, a partir de los *n* puntos $\{\mathbf{P}_1,\mathbf{P}_2,...,\mathbf{P}_n\}$ representados en los *k* sistemas de coordenadas $\{\mathbf{X}_1,\mathbf{X}_2, ...,\mathbf{X}_k\}$ m_k- dimensionales, se deben encontrar: las traslaciones $\{\mathbf{T}_1,\mathbf{T}_2,...,\mathbf{T}_k\}$, las rotaciones ortogonales $\{\mathbf{H}_1,\mathbf{H}_2,...,\mathbf{H}_k\}$ y los coeficientes de escala $\{\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_k\}$ tal que las configuraciones resultantes $\{\mathbf{X}^*_i = \mathbf{r}_i\,\mathbf{X}_i\mathbf{H}_i + \mathbf{T}_i$; i=1,...,k$\}$ sean lo más parecidas posible entre ellas. El promedio de estos nuevos sistemas de coordenadas produce el sistema de consenso. El nuevo sistema puede ser submitido, por ejemplo, a un ACP para obtener una ordenación de consenso en un espacio de menor dimensionalidad.

REPRESENTACIÓN GRÁFICA DE ORDENAMIENTOS MULTIDIMENSIONALES

Gráfico de dispersión

El diagrama de dispersión representa un conjunto de puntos ordenados en el plano con coordenadas X e Y. Los gráficos de dispersión de las dos primeras componentes principales o de las dos primeras coordenadas principales son comúnmente utilizados para representar ordenaciones.

Biplot

Los gráficos biplots propuestos por Gabriel (3) permiten representar las observaciones y las variables en un mismo plano. Para encontrar ejes óptimos para la graficación de observaciones y variables en un espacio común se utiliza la idea de que cualquier matriz de datos $n \times m$ puede ser representada aproximadamente en d-dimensiones como el producto de dos matrices: $\mathbf{X = AB'}$, con $\mathbf{A}$ $(n \times d)$ y $\mathbf{B}$ $(m \times d)$, siendo d el rango de $\mathbf{X}$. A partir de la descomposición por valor singular de $\mathbf{X}$, DVS($\mathbf{X}$)=$\mathbf{UDV'}$, las matrices $\mathbf{A}$ y $\mathbf{B}$ pueden ser obtenidas como: $\mathbf{A = UD}^{\alpha}$ y $\mathbf{B = VD}^{1-\alpha}$ donde α es un escalar que toma usualmente los valores 0, ½ ó 1. Cuando α es igual a ½ el biplot se denomina simétrico. Debido a que $\mathbf{A}$ y $\mathbf{B}$ tienen una base común de d vectores, es posible mostrar las filas y las columnas de la matriz $\mathbf{X}$ sobre el mismo gráfico. Las filas de $\mathbf{A}$ representan las observaciones en el espacio de menor dimensión (puntos filas) y las columnas de $\mathbf{B'}$ representan las variables (puntos columnas) en ese mismo espacio. Luego, el biplot puede obtenerse como un gráfico de dispersión de los $n+m$ vectores de $\mathbf{A}$ y $\mathbf{B}$ en el espacio d-dimensional. Cuando d es mayor a dos, es común utilizar la aproximación bidimensional de $\mathbf{X}$.

Árbol de recorrido mínimo (ARM)

Un árbol de recorrido se construye como una colección de segmentos de línea recta que conectan puntos de un ordenamiento sin formar circuitos cerrados. Cada punto está conectado con el resto de manera directa o indirecta a través del conjunto de segmentos. El árbol de recorrido mínimo es generado conectando los puntos de manera tal que la suma de las longitudes de los segmentos entre puntos sea mínima.

Un ARM puede calcularse a partir de la matriz de distancia de las observaciones multivariadas en el espacio m-dimensional en el que viven o a partir de las matrices de distancia en espacios de menor dimensión. Cuando puntos m-dimensionales, con $m>2$, son conectados en el plano en función de su distancia en el espacio original, el ARM puede proveer información sobre similitudes de las observaciones en dimensiones no directamente representadas en el plano.

Por ejemplo: algunos puntos que se encuentran muy cerca en el espacio bidimensional podrían estar, en su espacio original, más lejos de lo que aparentan en el plano. Los ARM conceptualmente se ligan al algoritmo de agrupamiento conocido como encadenamiento simple y en ese sentido son usados no sólo para representación gráfica de las interdistancias entre puntos, sino también para formar conglomerados de éstos.

ILUSTRACIÓN DE TRD Y VISUALIZACIÓN GRÁFICA DE ORDENAMIENTOS

Se utiliza un conjunto de datos de marcadores moleculares y morfológicos para caracterizar material fúngico, con la finalidad de ilustrar varios aspectos de la interpretación de las ordenaciones obtenidas por las TRD discutidas anteriormente, complementadas con gráficos biplot y árboles de mínimo recorrido.

Los datos originales fueron recolectados en un estudio sobre descripción del origen, biogeografía y variación molecular de *Moniliophthora roreri* (Cif.) Evans *et al.*, causante de moniliasis en cápsulas del cacao. En dicho estudio, sobre un total de 84 aislamientos se realizó una clasificación de aislamientos del hongo en cinco grupos genéticos: Bolivar, Co-West, Co-East, Co-Central y Gileri (7). En este trabajo se calculó un perfil modal para cada uno de los grupos genéticos identificados por Phillips (7) para un conjunto de 4 marcadores morfológicos: Rint20 (cantidad de anillos a los 20 días), Prod (producción de esporas por caja de Petri), Ge24h (germinación de esporas a las 24 horas) y Glo (porcentajes de esporas globosas) y para 4 marcadores moleculares: W5, W8, X15, Y18 (tablas 2 y 3). Si bien se disponía de información sobre numerosos marcadores morfológicos y moleculares, se seleccionaron sólo aquellos que en estudios previos mostraron mayor poder de discriminación entre grupos y se trabajó con los perfiles modales y no con los aislamientos individuales para aplicar las técnicas de análisis sobre una matriz de datos que se pueda visualizar con facilidad.

Tabla 2. Perfil modal de cuatro marcadores morfológicos para cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.*

Table 2. Modal profile of four morphological markers for five genetic groups *Moniliophthora roreri* (Cif.) Evans *et al.*

Grupo Genético	Rint2O[1]	Prod[2]	Ge24h[3]	Glo[4]
Bolivar	2,56	261,35	2,33	72,12
Co-Central	3,93	176,25	6,40	56,65
Co-East	2,13	275,37	2,93	57,58
Co-West	2,16	128,77	7,15	59,31
Gileri	0,75	221,60	2,00	40,55

[1]Rint20: cantidad de anillos a los 20 días, [2]Prod: producción de esporas por caja de Petri, [3]Ge24h: germinación de esporas a las 24 horas, [4]Glo: porcentaje de esporas globosas.

[1]Rint20: number of rings after 20 days, [2]Prod: production of spores per petri box, [3]Ge24h: germination of spores after 24 hours, [4]Glo: percentage of spores globose.

Tabla 3. Perfil modal de cuatro marcadores moleculares binarios para cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.*

Table 3. Modal profile of four binary molecular markers for five genetic groups *Moniliophthora roreri* (Cif.) Evans *et al.*

Grupo Genético	W5	W8	X15	Y18
Bolivar	1	1	1	0
Co-Central	1	1	0	1
Co-East	0	1	0	0
Co-West	1	1	1	1
Gileri	1	0	0	1

Análisis de componentes principales y biplot

Se aplicó el análisis de componentes principales (ACP) sobre los datos de la tabla 2 (pág. 191), previa estandarización de las variables. Se usaron datos estandarizados (matriz **R**) dado que las unidades de medida de estos marcadores morfológicos son inconmensurables. En la tabla 4 se puede visualizar que las correlaciones entre variables y sus varianzas difieren notablemente. Si se realizara el ACP sobre **S** en lugar de **R,** *i.e.* sin previa estandarización, las variables con mayor influencia en las dos primeras componentes serán las de mayor varianza (Prod y Ge24h). En casos como éste, no es deseable que el peso de las variables en las componentes venga dado por la varianza que a su vez depende de la escala de medida; por ello, **R** produce componentes que mejor reflejan las correlaciones de las variables.

Tabla 4. Matriz de covarianzas (**S**) y matriz de correlación (**R**). Datos de marcadores morfológicos.

Table 4. Covariance matrix (**S**) and correlation matrix (**R**). Morphological data markers.

S	**Rint2O**	**Prod**	**Ge24h**	**Glo**
Rint2O	1,29			
Prod	-14,87	3686,60		
Ge24h	1,58	-129,48	5,87	
Glo	7,11	111,11	3,32	126,18
R	**Rint2O**	**Prod**	**Ge24h**	**Glo**
Rint2O	1,000			
Prod	-0,215	1,000		
Ge24h	0,573	-0,880	1,000	
Glo	0,556	0,163	0,122	1,000

Si se comparan los coeficientes de los autovectores obtenidos desde **R** y desde **S** y las correlaciones de cada variable con los primeros dos componentes principales, se observará que para datos estandarizados, las correlaciones ordenan las variables de la misma manera que los autovectores, mientras que para datos no estandarizados existen diferencias, en la primera componente, entre el orden dado por las correlaciones y el orden dado por las contribuciones a las componentes. Estas diferencias se deben a que las correlaciones de las variables con las componentes proveen sólo información univariada sobre cómo opera cada variable por sí misma ignorando la presencia de las otras variables. Por otro lado, ya que las componentes principales son ortogonales (provienen de autovectores de una matriz simétrica) es posible expresar el coeficiente de correlación múltiple de las dos primeras componentes con la variable X_i mediante la siguiente partición $r^2_{xi,CP1} + r^2_{xi,CP2} = r^2_{xi\,|CP1,CP2}$. Por ejemplo, para la variable Prod, el cuadrado del coeficiente de correlación múltiple es **$(-0,786)^2 + (-0,576)^2 = 0,950$**. Los valores de r^2 del ejemplo muestran que las variables de mayor contribución para separar los aislamientos en el plano son Ge24h (0,980) y Prod (0,950) (tabla 5, pág. 193). El cálculo de correlaciones entre variables y componentes puede ayudar a mejorar la interpretación de los datos.

Tabla 5. Valores de coeficiente de correlación múltiple al cuadrado de los datos estandarizados y sus correspondientes factores que dieron lugar al coeficiente de correlación múltiple.

Table 5. Values of multiple correlation coefficient square of the standardized data and associated factors that led to the multiple correlation coefficient.

Variable	Primera componente principal (CP1)	Segunda componente principal (CP2)	r^2
Ge24h	0,960	-0,241	0,980
Prod	-0,786	0,576	0,950
Rint20	0,748	0,528	0,838
Glo	0,335	0,864	0,859

En el ejemplo de ilustración, el análisis de los resultados se realizó a partir de los datos estandarizados y se consideraron sólo los dos primeros componentes principales: ellos explican un 90,6% de la varianza total. El primer autovector tiene elementos positivos a excepción del correspondiente a la variable Prod (-0,786); las variables de calidad de esporas (Rint20 y Ge24h) presentan coeficientes de magnitudes similares a la variable relacionada a la cantidad de espora (Prod), pero de signos opuestos. Luego la primera componente separa los aislamientos con mayor producción (CP1 de menor valor) de aquellos con menor producción de esporas pero con mayor germinación de las mismas a las 24 horas y mayor cantidad de anillos (CP1 de mayor valor). En la segunda componente se oponen los aislamientos con mayor porcentaje de esporas globosas (CP2 de mayor valor) de aquellos con menor porcentaje, principalmente dentro de los aislamientos de mayor producción de esporas. Estas relaciones se pueden interpretar mejor a partir de su representación gráfica en un biplot (figura 1).

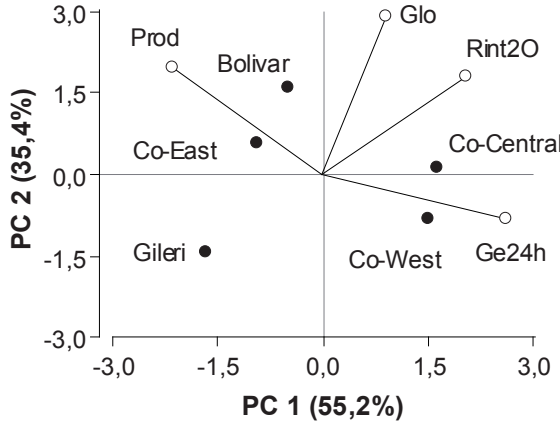

Figura 1. Gráfico biplot obtenido a partir de ACP. Ordenamiento de cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.* a partir de cuatro marcadores morfológicos.

Figure 1. Biplot obtained from PCA. Ordination from five genetic groups of *Moniliophthora roreri* (Cif.) Evans et al. from four morphological markers.

Los coeficientes de mayor valor absoluto corresponden a las variables que mayor peso tienen para caracterizar las observaciones (grupos genéticos). En este ejemplo, las variables Ge24h y Prod tienen una fuerte influencia para la caracterización morfológica de los grupos genéticos. A nivel del Eje 2, las variables que más se separan son Glo y Ge24h, oponiéndose entre sí con coeficientes de autovectores de 0,864 y -0,241 respectivamente. Las variables Rint20 (cantidad de anillos a los 20 días) y Glo (porcentaje de esporas globosas) presentan vectores con menor ángulo entre ellas, indicando una asociación positiva entre las mismas.

Puede observarse a nivel de la CP1 que la variable Ge24h está correlacionada positivamente con los grupos Co-Central y Co-West, mientras que Prod correlaciona positivamente con los aislamientos del grupo Co-East, Bolivar y Gileri. La CP2 sugiere que los aislamientos del grupo Gileri son los de grupo de menor número de anillos a los 20 días y de menor porcentaje de esporas globosas.

Biológicamente, el bajo número de anillos a los 20 días estaría evidenciando una alternancia del periodo de crecimiento y esporulación del hongo, procesos normalmente regidos por ciclos biológicos que dependen de la luz, *i.e*., un hongo con baja producción de anillos estaría indicando una falta de adaptación al ritmo biológico regido por la cantidad de luz recibida, afectando su nivel de esporulación y consecuentemente su sobrevivencia.

Análisis de coordenadas principales y ARM

Con los cuatro marcadores continuos se calculó una matriz de distancias Euclídeas (tabla 6) previa estandarización de los datos; sobre dicha matriz se extrajeron las coordenadas principales. Estas coordenadas son iguales a las que se obtendrían realizando un análisis de componentes principales sobre la matriz **R** de dimensión 4×4. Esta equivalencia entre el ACP y el ACoorP se logra sólo en el caso de trabajar sobre una matriz de distancias Euclídeas para datos cuantitativos continuos y estandarizados. En la figura 2 (pág. 195) se muestra el ordenamiento de los aislamientos para el ACoorP.

Tabla 6. Matriz de distancias Euclídeas para cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.* obtenidos a partir de cuatro marcadores morfológicos de naturaleza continua previa estandarización.

Table 6. Euclidean distance matrix for five genetic groups of *Moniliophthora roreri* (Cif.) Evans *et al.* obtained from four morphological markers of continuous nature and previous standarization.

	Bolivar	Co-Central	Co-East	Co-West	Gileri
Bolivar	0,000				
Co-Central	2,852	0,000			
Co-East	1,390	2,688	0,000		
Co-West	3,186	1,785	2,981	0,000	
Gileri	3,298	3,705	2,168	3,344	0,000

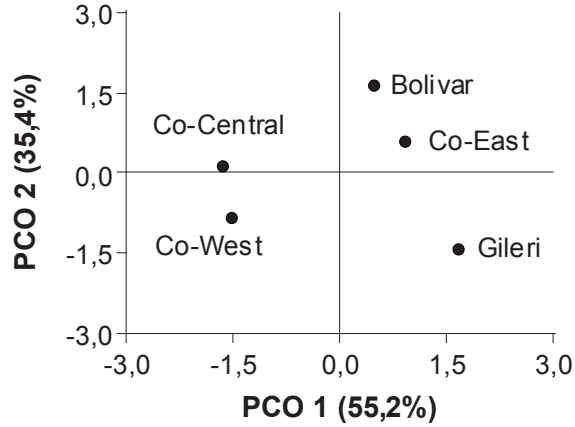

Figura 2. Diagramas de dispersión a partir de las coordendas principales (PCO) obtenidas utilizando distancias Euclídeas para cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.* a partir de cuatro marcadores morfológicos.

Figure 2. Scatter plots from the principal coordinates (PCO) obtained from Euclidean distances for five genetic groups of *Moniliophthora roreri* (Cif.) Evans *et al.* involving four morphological markers.

Esta equivalencia no se produce cuando los aislamientos deben ser ordenados a partir de datos de naturaleza discreta. Para ilustrar esta situación se usaron los datos binarios provenientes de los marcadores moleculares y se realizó un ACoorP sobre la matriz de distancia obtenida a partir de la transformación $(1 - S_{ij})^{1/2}$ del índice de similitud de Dice (tabla 7). Los autovalores asociados a cada coordenada principal y el valor de las mismas se muestran en la tablas 8 y 9, respectivamente (pág. 196).

Tabla 7. Matriz de distancias para cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.* usando datos de cuatro marcadores moleculares.

Table 7. Distance matrix for five genetic groups of *Moniliophthora roreri* (Cif.) Evans *et al.* using genetic data from four molecular markers.

	Bolivar	Co-Central	Co-East	Co-West	Gileri
Bolivar	0,000				
Co-Central	0,577	0,000			
Co-East	0,707	0,707	0,000		
Co-West	0,378	0,378	0,775	0,000	
Gileri	0,775	0,447	1,000	0,577	0,000

Distancia: $(1 - S_{ij})^{1/2}$ donde S_{ij} es el índice de similitud de Dice

Tabla 8. Autovalores obtenidos como resultado del ACoorP sobre la matriz de distancias de la tabla 7.

Table 8. Eingenvalues obtained as result of PCO on the distance matrix of table 7.

Lambda	Lamda (valor)	Proporción explicada	Proporción acumulada
1	0,537	0,617	0,617
2	0,245	0,281	0,898
3	0,063	0,072	0,970
4	0,026	0,030	1,000

Tabla 9. Coordenadas principales (PCO) obtenidas como resultado del ACoorP sobre la matriz de distancias de la tabla 7.

Table 9. Principal coordinates obtained as result of PCO on the distance matrix of table 7.

PCO(1)	PCO(2)	PCO(3)	PCO(4)
0,50	1,60	-0,41	-0,17
-1,62	0,12	0,78	-0,09
0,95	0,56	0,23	0,31
-1,50	-0,84	-0,71	0,08
1,68	-1,43	0,11	-0,13

En la figura 3 se presentan los diagramas de dispersión (sobre los que se superpusieron los ARM) de los aislamientos en función de: (A) las primeras dos coordenadas principales de la tabla 9 y (B) las dos primeras componentes principales del mismo conjunto de datos.

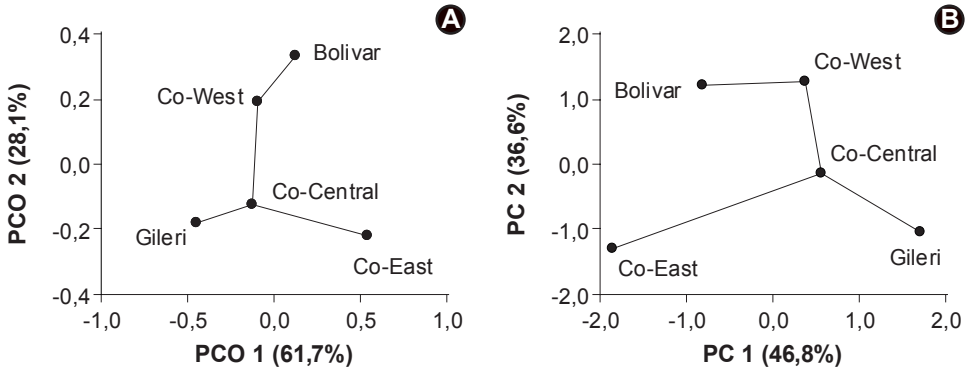

Figura 3. ARM sobre el ordenamiento de cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.* en el plano conformado por las dos primeras coordendas principales calculadas sobre la matriz de distancias genéticas obtenidas por la transformación $(1 - S_{ij})^{1/2}$ del índice de similitud de Dice (A). ARM sobre el ACP para datos de marcadores (B).

Figure 3. MST on the ordination of five genetic groups of *Moniliophthora roreri* (Cif.) Evans *et al.* in the plane formed by the first two principal coordinates calculated from genetic distancesobtained as $(1 - S_{ij})^{1/2}$ where S is Dice's similarity index (A). MST on PCA for molecular markers data (B).

Si bien no es apropiado realizar un ACP sobre datos binarios, en la figura 3 (pág. 196) se presentan ambos resultados ya que es muy frecuente su uso en numerosas bibliografías. En ella se observa que las figuras A y B muestran la equivalencia que se observó con los marcadores continuos ya que en A se ordenaron los aislamientos a partir de una distancia especialmente recomendada para este tipo de marcadores y en B al descomponer espectralmente la matriz de varianzas y covarianzas se preservan las distancias Euclídeas entre aislamientos (tabla 10), que más allá del cambio de escala no hacen diferencias entre casos donde la similitud de aislamientos se produce por la co-presencia de marcadores o por la co-ausencia.

Tabla 10. Matriz de distancias Euclídeas entre cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.* usando datos de cuatro marcadores moleculares.
Table 10. Euclidean distance matrix among five genetic group of *Moniliophthora roreri* (Cif.) Evans *et al.* using genetic data of four molecular markers.

	Bolivar	Co-Central	Co-East	Co-West	Gileri
Bolivar	0,000				
Co-Central	2,582	0,000			
Co-East	2,887	2,887	0,000		
Co-West	1,826	1,826	3,416	0,000	
Gileri	3,416	2,236	3,651	2,887	0,000

La distancia entre los perfiles modales del grupo Gileri respecto de Co-East es relativamente mayor si se trabaja con el índice de similitud de Dice respecto de las distancias Euclídeas, ya que el único parecido entre ambos perfiles se da por la ausencia simultánea del marcador X15. Las diferencias entre los perfiles de los grupos Bolivar y Co-West son relativamente menores, a nivel del primer Eje, para la distancia basada en el índice de Dice que para las distancias Euclídeas ya que el índice de Dice pondera con mayor peso el parecido entre ambos perfiles que provienen de la co-presencia de 3 de los 4 marcadores involucrados. Si bien en ambos gráficos la CP1 separa Gileri de Co-East, usando la distancia de Dice explica un 61,7% de la variabilidad total, mientras que con la distancia Euclídea la variabilidad sobre este eje más importante de análisis representa el 46,8% de la variabilidad total.

Análisis de procrustes generalizado (APG) y ARM de consenso

Para ordenar en un único espacio las observaciones según la información brindada tanto por marcadores morfológicos como moleculares, se realizó un análisis de procrustes generalizado (APG). Dado que las variables morfológicas son de diferente naturaleza que las variables moleculares, es conveniente, previo a realizar el APG, extraer ejes apropiados para cada conjunto de datos. En la tabla 11 (pág. 198) se muestran los resultados obtenidos del APG previa obtención de las componentes principales (CP) de los datos morfológicos estandarizados y las coordenadas principales (PCO) de los datos moleculares derivados del ACoorP sobre la matriz de distancia obtenida con la transformación $(1 - S_{ij})^{1/2}$ donde S_{ij} es el índice de similitud de Dice. Los autovalores indican que la variabilidad explicada a través del Eje 1 de la descomposición de la matriz de consenso es 47,4%. Con los dos primeros ejes se explica el 80,6% de la variabilidad contenida en el total de los marcadores. En el cuadro de Análisis de la Varianza

se presenta la suma de cuadrados dentro por caso (grupos genéticos) y la suma de cuadrados dentro por grupo de marcadores. El total de la suma de cuadrados (total de las suma de cuadrados de consenso dentro de cada grupo de marcador) es 2. Si se calcula el cociente entre el consenso y la suma de cuadrados total (1.726/2), se concluye que existe un 86,3% de consenso entre el ordenamiento producido por los marcadores moleculares y el producido por marcadores morfológicos para el ordenamiento de estos 5 grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.*

Tabla 11. Resultados del análisis de procrustes generalizado (APG) combinando información proveniente de cuatro marcadores morfológicos y cuatro marcadores moleculares de cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.*

Table 11. Output from generalized procrustes analysis (GPA) combining information from four morphological markers and four molecular markers for five genetic groups *Moniliophthora roreri* (Cif.) Evans *et al.*

Autovalores	Lamda (valor)	Proporción de la variabilidad total explicada por el eje	Proporción acumulada
1	0,409	0,474	0,474
2	0,287	0,332	0,806
3	0,133	0,154	0,960
4	0,035	0,040	1,000

Cuadro de Análisis de la Varianza

Sumas de cuadrados dentro por caso

	Consenso	Residual	Total
Bolivar	0,311	0,039	0,351
Co-Central	0,217	0,057	0,274
Co-East	0,388	0,094	0,482
Co-West	0,270	0,028	0,298
Gileri	0,540	0,056	0,596
Total	1,726	0,274	2,000

Sumas de cuadrados dentro por grupo de marcadores

	Consenso	Residual	Total
Grupo1 (Moleculares)	0,863	0,137	1,000
Grupo2 (Morfológicos)	0,863	0,137	1,000
Total	1,726	0,274	2,000

En la figura 4 (pág. 199) se puede observar el consenso de las ordenaciones dado por las componentes principales de los marcadores morfológicos y las coordenadas principales de los marcadores moleculares. Sobre la figura de la izquierda se construyó un ARM para las configuraciones individuales dadas por cada tipo de marcador y para la configuración de consenso. En el diagrama de dispersión de la derecha se observan las ordenaciones individuales y de consenso sin el ARM pero se identifican los casos para poder interpretar el consenso por aislamiento. En la figura 4 A se puede destacar el parecido del grupo Co-West y Co-Central, cuando se consideran simultáneamente ambos tipos de marcadores. El parecido morfológico de estos aislamientos es alto. Los grupos Gileri y Co-West se diferencian más a nivel molecular que morfológico.

El grupo Bolivar está más cercano a Co-East que cualquier otro grupo si se consideran ambos tipos de marcadores. En la figura 4 B se observa que la suma de cuadrados (SC), función de las diferencias entre las ordenaciones individuales y la de consenso dentro de cada caso, es bastante similar ya que la ordenación de consenso se da para todos los casos en el punto medio de las distancias entre las configuraciones individuales y estas distancias son similares para todos los casos.

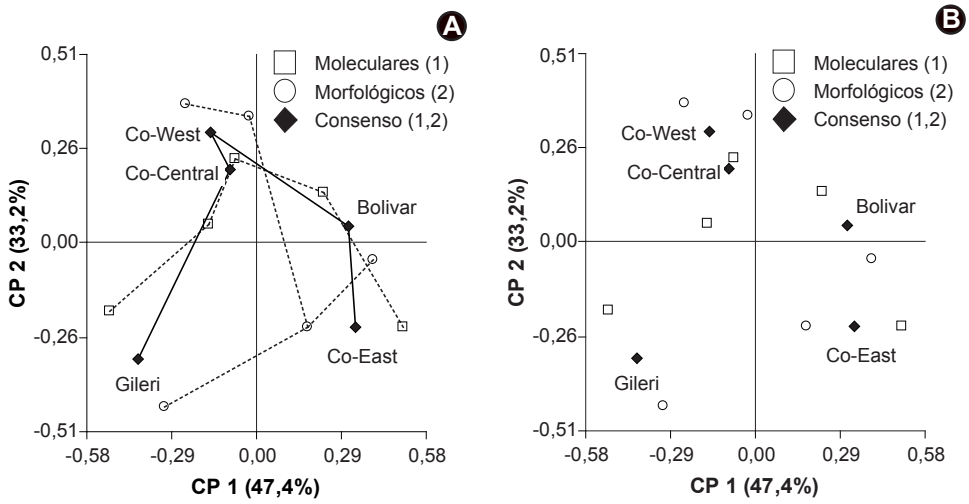

Figura 4. Ordenamiento de cinco grupos genéticos de *Moniliophthora roreri* (Cif.) Evans *et al.* en el plano conformado por los dos primeros ejes de un APG con ARM (A) y sin ARM (B). Se calculó la distancia genética del índice de similitud de Dice mediante la transformación $(1 - S_{ij})^{1/2}$ con cuatro marcadores moleculares para obtener las coordenadas principales, y la distancia Euclídea para los marcadores morfológicos.

Figure 4. Ordination of five genetic groups *Moniliophthora roreri* (Cif.) Evans *et al.* in the plane formed by the first two axes of a GPA with MST (A) and without MST (B). Genetic distance was estimated from Dice's similarity index by $(1 - S_{ij})^{1/2}$ transformation for four molecular markers to obtain the principal coordinates and the Euclidean distance for morphological markers.

A modo de ilustración sobre las alternativas de uso de APG, se realizó un APG directamente desde la matriz de datos moleculares y morfológicos previa estandarización de estos últimos. Este proceder es equivalente a realizar un APG partiendo de las coordenadas principales (PCO) de un ACoorP sobre una matriz de distancias Euclídeas obtenidas desde los marcadores moleculares y de las componentes principales (CP) de un ACP sobre la matriz de marcadores morfológicos. En la figura 4 se observa la dispersión de los ordenamientos individuales y de consenso, sobre las que se construyó un ARM. Los grupos Co-West y Co-Central se encuentran a una distancia muy pequeña al igual que antes, pero el grupo Gileri no se asocia de manera directa a estos perfiles, sino que lo hace a través del grupo Co-East y el grupo Bolivar. Observando los valores

de la tabla 2 (pág. 191) se advierte que las asociaciones que se presentan en este consenso hacen referencia a los parecidos de los perfiles morfológicos, pero éstos no se condicen con los parecidos a nivel molecular, de hecho, entre el perfil modal molecular del grupo Gileri y el perfil molecular del grupo Co-East hay sólo una coincidencia y ésta es por la ausencia simultánea de amplificación, mientras que con los grupos Co-Central y Co-West, el grupo Gileri presenta más coincidencias, y éstas son debido a las presencia simultánea de los marcadores, característica explotada en el análisis anterior donde se realizó previo al APG un ACoorP sobre la matriz de distancias del índice de similitud de Dice para los datos moleculares.

CONCLUSIONES

La interpretación de la información provista por múltiples marcadores mejora sustancialmente al poder visualizar en un espacio de baja dimensión las observaciones. El uso de las diferentes técnicas de análisis para obtener una reducción de la dimensión depende del tipo de datos con el que se trabaja y el objetivo del análisis. Para datos de naturaleza binaria es conveniente utilizar un ACoorP y una medida de distancia acorde a datos discretos como aquellas basadas en índices de similitud, mientras que cuando se trabaja con datos de naturaleza continua, como los obtenidos a partir de variables morfológicos se recomienda trabajar con ACP de datos estandarizados, sobre todos si las variables no son conmensurables o su variabilidades son muy distintas. El APG es una técnica útil cuando se desean estudiar las relaciones entre materiales a partir de datos de marcadores de diferente naturaleza. Este análisis permite estimar la magnitud de la correlación entre los distintos grupos de variables. Es importante tener en cuenta la naturaleza de los grupos de variables que se desean consensuar y en función de ello elegir la medida de distancia y el método de ordenación más apropiado según el tipo de marcador antes de consensuar las ordenaciones.

BIBLIOGRAFÍA CITADA

1. Bramardi, S. J.; Bernet, G. P.; Asíns, M. J.; Carbonell, E. A. 2005. Simultaneous agronomic and molecular characterization of genotypes via the generalized procrustes analysis: an application to cucumber. Crop Sci 45(4): 1603-1609.
2. Bruno, C.; Balzarini, M.; Di Rienzo, J. 2003. Comparación de medidas de distancais entre perfiles RAPD. Journal of Basic & Applied Genetics. 15: 69-78.
3. Gabriel, K. R. 1971. Biplot display of multivariate matrices with application to principal components analysis. Biometrika. 58: 453-467.
4. Gower, J. C. 1971. A general coefficient of similarity and some properties. Biometrics. 27: 857-872.
5. _____. 1975. Generalizad procrustes analysis. Psychometrika. 40: 33-51.
6. Johnson, R. A.; Wichern, D. W. 1998. Applied multivariate statistical analysis. 4th ed. Prentice Hall. Upper Saddle River. N. J. 816 p.
7. Phillips, W. 2003. Origin, biogeography, genetic diversity and taxonomic affinities of the cacao (*Theobroma cacao* L.) fungus *Moniliophthora roreri* (Cif.) Evans *et al*. as determined using molecular, phytopathological and morpho-physiological evidence. Tesis *Ph.D*., University of Reading. UK, 349 p.