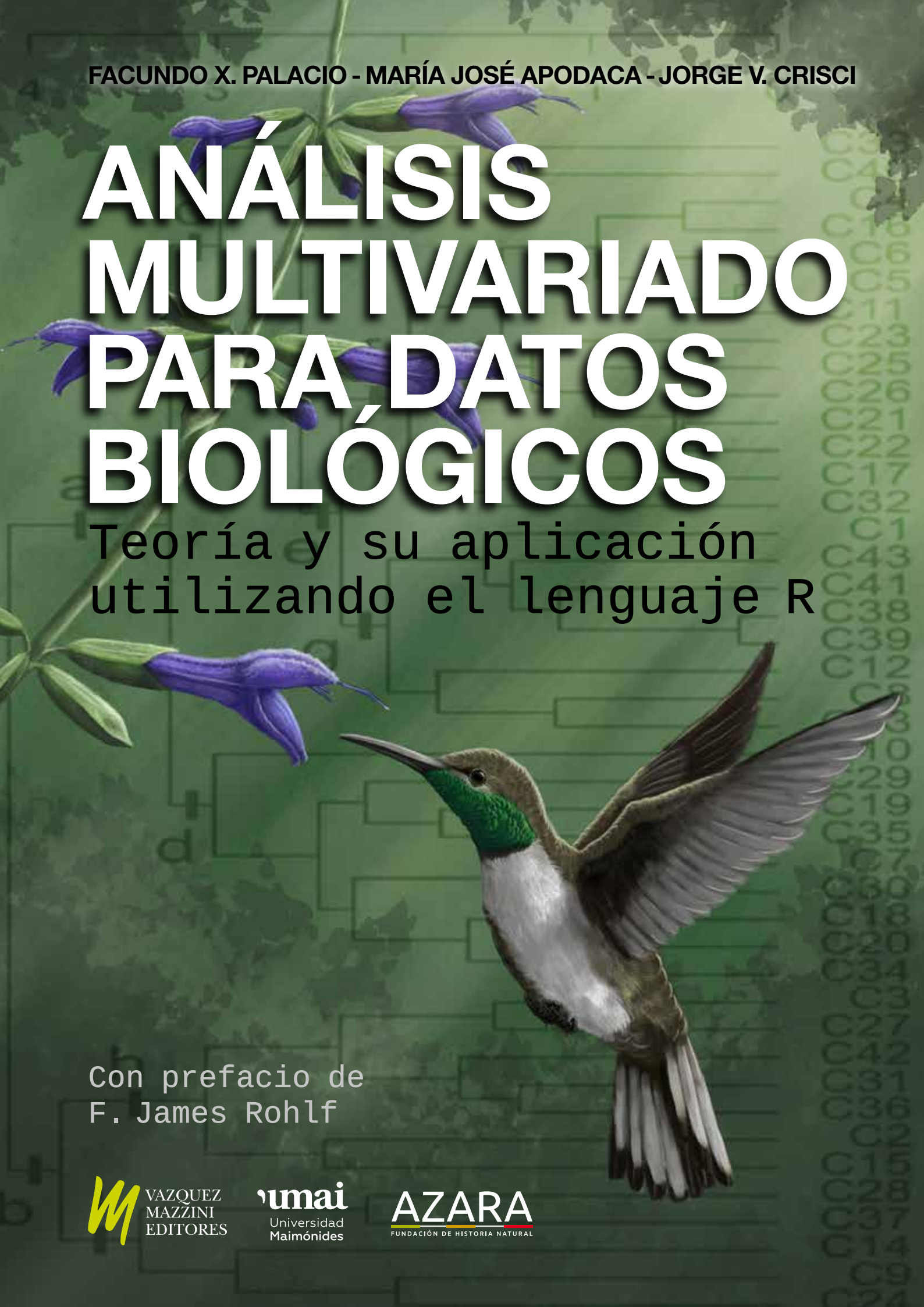
FACUNDO X. PALACIO - MARÍA JOSÉ APODACA - JORGE V. CRISCI

ANÁLISIS MULTIVARIADO PARA DATOS BIOLÓGICOS

Teoría y su aplicación utilizando el lenguaje R



Con prefacio de
F. James Rohlf

VAZQUEZ MAZZINI EDITORES
umai Universidad Maimónides
AZARA FUNDACIÓN DE HISTORIA NATURAL

ANÁLISIS MULTIVARIADO PARA DATOS BIOLÓGICOS

FACUNDO X. PALACIO - MARÍA JOSÉ APODACA - JORGE V. CRISCI

ANÁLISIS MULTIVARIADO PARA DATOS BIOLÓGICOS

Teoría y su aplicación utilizando el lenguaje R

VAZQUEZ MAZZINI EDITORES

Universidad Maimónides

AZARA
FUNDACIÓN DE HISTORIA NATURAL

Fundación de Historia Natural Félix de Azara
Departamento de Ciencias Naturales y Antropológicas
CEBBAD - Instituto Superior de Investigaciones
Universidad Maimónides
Hidalgo 775 - 7° piso (1405BDB) Ciudad Autónoma de Buenos Aires - República Argentina
Teléfonos: 011-4905-1100 (int. 1228)
E-mail: secretaria@fundacionazara.org.ar
Página web: www.fundacionazara.org.ar

Dibujo de tapa: ejemplar macho de picaflor andino (*Oreotrochilus leucopleurus*) visitando una flor de salvia azul (*Salvia guaranitica*), con un dendrograma de fondo.
Autor: Martín Colombo.

Primera Edición: 2020

Impreso en la Argentina.
Se terminó de imprimir en el mes de Mayo 2020, en la Ciudad de Buenos Aires.

VAZQUEZ MAZZINI EDITORES
Tel. (54-11) 4905-1232
info@vmeditores.com.ar
www.vmeditores.com.ar

Palacio, Facundo Xavier

Análisis multivariado para datos biológicos : teoría y su aplicación utilizando el lenguaje R / Facundo Xavier Palacio ; María José Apodaca ; Jorge Víctor Crisci. - 1a ed . - Ciudad Autónoma de Buenos Aires : Fundación de Historia Natural Félix de Azara, 2020.

268 p. ; 30 x 21 cm.

ISBN 978-987-3781-49-0

1. Biología. 2. Bioestadísticas. I. Apodaca, María José. II. Crisci, Jorge Víctor. III. Título.
CDD 570.285

“La imposibilidad de penetrar el esquema divino del universo no puede, sin embargo, disuadirnos de planear esquemas humanos, aunque nos conste que éstos son provisorios”.

Jorge Luis Borges

El idioma analítico de John Wilkins (1952)

Índice

PRESENTACIÓN DE F. JAMES ROHLF

I was very pleased to be asked to write a foreword to this book by my longtime friend Jorge Crisci. We first met at the 12th Numerical Taxonomy Conference at Stony Brook University back in 1978. This book on multivariate statistical methods and their applications to problems in biology (including phylogenetic estimation) can be seen as a result of his long interest in multivariate methods and their applications to taxonomy and evolutionary inference.

An understanding of multivariate methods and their proper use is very important in many biological disciplines. Biological relationships are often complex and variation in many variables must be considered. It would be difficult to understand the patterns of variation among organisms if one only performed analyses using just a single variable at a time. Thinking in terms of multivariate variation and thus thinking in terms of multidimensional spaces does take more effort but it is a skill that is now essential in many areas of biology.

An important contribution of the book is that it includes practical information about how to perform various multivariate analyses using the R language that has become increasingly popular among biologists. There is a chapter with an introduction to the use of R and each chapter presenting a type of multivariate method has its own section giving practical examples of how to use R to perform the analyses presented in that chapter. I am sure many will find the book to be an essential practical guide.

F. James Rohlf
Stony Brook, USA
Research Prof. Dept. of Anthropology
Distinguished Prof. Emeritus Dept. Ecology and Evolution
Stony Brook University, USA
5 August 2019

TRADUCCIÓN DE LA PRESENTACIÓN DE F. JAMES ROHLF

"Me llenó de satisfacción que mi amigo desde hace mucho tiempo, Jorge Crisci, me pidiera escribir una presentación de este libro. Con Jorge nos conocimos en la *12th Numerical Taxonomy Conference* en *Stony Brook University* en 1978. Este libro sobre métodos estadísticos multivariados y sus aplicaciones a problemas en la Biología (incluyendo estimación filogenética), puede verse como el resultado de su continuo interés en los métodos multivariados y sus aplicaciones a la taxonomía y a la inferencia evolutiva.

La comprensión de los métodos multivariados y el uso apropiado de los mismos, es muy importante en numerosas disciplinas biológicas. Las relaciones biológicas son a menudo complejas y en estos casos deben considerarse muchas variables. Sería difícil comprender los patrones de variación de los organismos si se realizaran análisis empleando una sola variable a la vez. Pensar en términos de variación multivariada y, por lo tanto, pensar en términos de espacios multidimensionales, requiere más esfuerzo, pero actualmente es una destreza esencial en muchas áreas de la Biología.

Una contribución importante del libro es la inclusión de información práctica sobre cómo realizar un análisis multivariado utilizando el lenguaje R, lenguaje cada vez más utilizado por los biólogos. Hay un capítulo con una introducción al uso de R, y a su vez, cada uno de los otros capítulos tiene su propia sección con ejemplos prácticos de cómo usar R para realizar los análisis presentados en él. Estoy seguro de que muchos encontrarán en este libro una guía práctica esencial."

UNAS PALABRAS ACERCA DE F. JAMES ROHLF

F. James Rohlf es uno de los líderes mundiales en bioestadística, cuya labor tuvo y tiene una enorme influencia en los últimos 60 años del análisis multivariado. Es imposible pensar el análisis multivariado sin recurrir a los aportes fundamentales que realizó y realiza en esta disciplina.

Los autores de este libro agradecen a F. James Rohlf la presentación de este libro, sus consejos, su permanente ayuda y la inspiración que su liderazgo, en los temas de análisis multivariado, ha suscitado en científicos de varias generaciones.

PRÓLOGO DE LOS AUTORES

Los métodos multivariados se aplican en la Biología desde principios del siglo XX, pero han tenido una enorme difusión en los últimos años, debido a la gran cantidad de información acumulada en las bases de datos y al enorme progreso de la tecnología computacional que comenzó en la década de 1960.

Un aspecto común a las aplicaciones del análisis multivariado es que todas consideran un conjunto de objetos, donde cada objeto es descripto por una serie de atributos o variables. Estos objetos pueden ser conjuntos de individuos, especímenes, taxones, comunidades o cuadrantes geográficos, entre otros. Las variables pueden ser características de individuos o de taxones, la presencia o ausencia de una especie en una comunidad, o de un espécimen en un cuadrante geográfico. La elección de los objetos y de las variables depende de las preguntas planteadas por el investigador.

El análisis multivariado intenta encontrar patrones de similitud entre objetos sobre la base de las variables utilizadas. Estos patrones permiten formar grupos cuyos objetos son más similares entre sí, que con los objetos integrantes de otros grupos. Asimismo, el análisis multivariado busca identificar aquellas variables que permiten discriminar dichos grupos de objetos. Los patrones resultantes del análisis multivariado permiten contrastar hipótesis sobre las relaciones entre los objetos y explicar la causalidad de los agrupamientos, como así también, predecir objetos y variables todavía no descubiertos.

A pesar de que la mayoría de los libros sobre el análisis multivariado no incluyen a los análisis filogenéticos, los hemos incorporado a esta obra, pues ambos métodos utilizan el mismo tipo de matriz de datos, objetos (individuos, poblaciones o taxones) × variables (caracteres en filogenia), diferenciándose en los algoritmos de análisis que utilizan. Por otro lado, la filogenia tiene actualmente un gran impacto sobre todas las subdisciplinas de la Biología, y muy especialmente en la Biogeografía, la Biología de la Conservación, la Ecología, la Etología y la Epidemiología.

El objetivo de este libro es explicar e ilustrar las técnicas más utilizadas del análisis multivariado aplicadas a datos biológicos, de manera de facilitar su comprensión y empleo por los investigadores. Todas las técnicas son, a su vez, presentadas dentro del contexto del programa R, para hacer posible su aplicación computacional. Los conjuntos de datos y las rutinas utilizados en este libro están disponibles en el siguiente enlace: https://fundacionazara.org.ar/analisis-multivariado-para-datos-biologicos/

Además de contribuir a la formulación de hipótesis sobre problemas particulares que requieran del análisis multivariado, esta obra intenta auxiliar en la toma de decisiones respecto a cuáles técnicas son las apropiadas de acuerdo con los datos relevados, y a interpretar los resultados obtenidos. De manera complementaria, se brindan las rutinas necesarias para la aplicación de estas técnicas mediante el programa libre y sin costo R (libre, en este caso, se refiere a la libertad de los usuarios para ejecutar, copiar, distribuir, estudiar, cambiar y mejorar el software). También se introducen de manera sumaria ejemplos empíricos, algunos de ellos obtenidos de nuestras propias investigaciones.

El libro está dividido en cinco grandes ejes temáticos distribuidos en ocho capítulos: (1) construcción de la matriz de datos (objetos × variables), (2) cálculo de medidas de similitud entre objetos, (3) análisis de agrupamientos, (4) técnicas de ordenación y (5) análisis filogenéticos.

Finalmente, el libro intenta responder dos preguntas que a menudo los investigadores se formulan: ¿qué es el análisis multivariado y qué puede hacer por mí?

F.X. Palacio, M.J. Apodaca y J.V. Crisci

AGRADECIMIENTOS

Expresamos nuestro más profundo agradecimiento a la Fundación de Historia Natural Félix de Azara y a su presidente, Adrián Giacchino, que hicieron posible la publicación de este libro.

Agradecemos también la lectura crítica de partes del manuscrito a Liliana Katinas (capítulos 1 y 2), Adrián Jauregui, Martín Colombo y Exequiel González (capítulo 3), Analía Lanteri (capítulo 4), Marta Fernández (capítulos 4, 5, 7 y 8), María Fernanda López Armengol, Mariana Grossi y Edgardo Ortiz-Jaureguizar (capítulo 5), Pablo Demetrio (capítulo 6), Martín Ramírez (capítulo 7) y Santiago Benitez-Vieyra (capítulo 8). Lucas Garbin nos asesoró en el uso de la base de datos del sitio web GenBank.

El consejo y permanente ayuda de Elián Guerrero, Liliana Katinas, Marta Fernández y Piero Marchionni han permitido que este libro fuera de mejor calidad de la que hubiera sido. Martín Colombo diseñó la tapa y el fondo de la contratapa, y realizó las ilustraciones de la Figura 2.2 que muestran especies del género *Bulnesia*. Omar Varela y Pedro Blendinger brindaron fotografías de *B. retama* y *B. sarmientoi*.

Agradecemos a la Agencia Nacional de Promoción Científica y Tecnológica por los PICT-2017-0965 (MJA y JVC) y PICT-2017-0081 (FXP), al Consejo Nacional de Investigaciones Científicas y Técnicas por los PIP 2013-2015 # 0446 y PIP 2017-2019 # 0421 (MJA y JVC) y a la Universidad Nacional de La Plata, por el apoyo financiero y el respaldo institucional brindados (FXP, MJA y JVC).

Estos agradecimientos no implican responsabilidad alguna para las instituciones y colegas de los posibles errores que aún permanezcan en nuestro libro.

CAPÍTULO 1
INTRODUCCIÓN: PASOS ELEMENTALES DE LAS TÉCNICAS MULTIVARIADAS

La asociación de conceptos biológicos con variables numéricas ha dado como resultado una inmensa cantidad y variedad de técnicas multivariadas. Estas técnicas han sido y son ampliamente utilizadas, por ejemplo, en Ecología, Sistemática, Biogeografía y Microbiología (Sneath y Sokal 1973, James y McCulloch 1990, Legendre 1990, Shi 1993, Kreft y Jetz 2010, Paliy y Shankar 2016).

A pesar de esta diversidad, es posible hallar una serie de pasos comunes a todas ellas. En este capítulo presentaremos esos pasos elementales y analizaremos en detalle dos de ellos: la elección de las unidades de estudio (UE) y la elección de las variables. Los restantes pasos serán abordados en el resto de los capítulos.

La elección del tipo de análisis dependerá esencialmente de la pregunta del investigador y el tipo de datos. Por ejemplo, si el objetivo del trabajo es buscar relaciones entre las UE, se pueden usar análisis de agrupamientos o análisis filogenéticos. Si lo que se quiere es generar un espacio de dimensiones reducidas, se deberá utilizar algún análisis de ordenación. Si el objetivo es realizar un análisis filogenético con una matriz de datos morfológicos, sólo se podrá usar el método de parsimonia, ya que los supuestos de los modelos evolutivos probabilísticos para este tipo de datos no suelen cumplirse (Goloboff *et al.* 2018). Otro aspecto que tiene que ver con la pregunta del investigador, es si los grupos están definidos *a priori* o *a posteriori* del análisis, denominados en la jerga de la inteligencia artificial, aprendizaje supervisado (*supervised learning*) y no supervisado (*unsupervised learning*), respectivamente (Tarca *et al*. 2007, James *et al*. 2013).

PASOS ELEMENTALES DEL ANÁLISIS MULTIVARIADO

Los pasos elementales comunes a casi todas las técnicas de análisis multivariado son los siguientes (Fig. 1.1):

1. **Elección de las unidades de estudio (UE).** Este paso depende del problema que el investigador quiere resolver, de la estrategia que pretende utilizar y, en gran medida, de la disciplina en la cual ese problema se enmarca: Sistemática, Ecología, Biogeografía, Palinología, Paleontología, etc.
2. **Elección de las variables.** Se eligen variables (morfológicas, climáticas, ecológicas, genéticas) que difieren entre las UE y se registra el valor (dato) de cada una ellas para las UE.
3. **Construcción de una matriz básica de datos (MBD).** Con la información obtenida en los pasos anteriores se construye una MBD de UE × variables (ver Cap. 2).
4. **Cálculo de un coeficiente de similitud.** A partir de la MBD se obtiene un coeficiente de similitud para cada par posible de las UE (ver Cap. 3).

5. **Construcción de una matriz de similitud (MS).** Con los valores de similitud calculados en el paso anterior se construye una MS de UE × UE (ver Cap. 4). En el caso de los análisis filogenéticos, éstos generalmente se aplican directamente sobre la MBD (ver Caps. 7 y 8).
6. **Aplicación de un análisis cuantitativo.** A partir de la MS, se pueden aplicar distintos métodos (análisis de agrupamientos, ordenación, y en algunos casos análisis filogenéticos; ver Caps. 5 a 8). En la Figura 1.2 se presentan los métodos y relaciones de los análisis abordados en el libro.
7. **Identificación de patrones**. Se obtienen patrones de relaciones entre las UE sobre la base de la MBD.
8. **Inferencias acerca de las UE.** Se formulan las generalizaciones acerca de las UE, tales como identificación de grupos de UE y elección de variables discriminatorias.

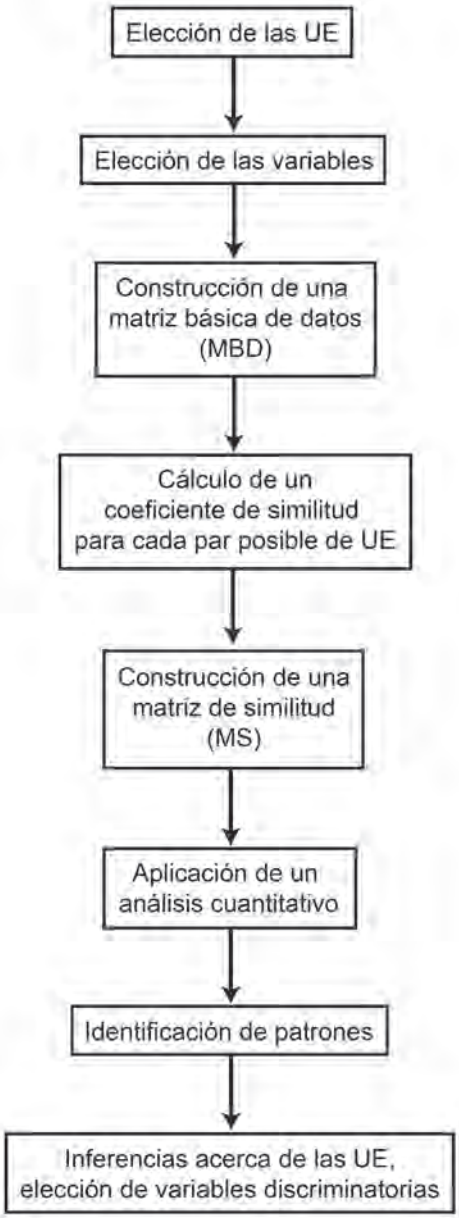


Fig. 1.1. Pasos elementales para la aplicación de las técnicas de análisis multivariado.

Es conveniente tener en cuenta los siguientes tres puntos:

1. Las inferencias acerca de las UE no pueden formularse antes de que los patrones biológicos de las UE hayan sido reconocidos.

2. Los patrones no pueden ser reconocidos antes de calcular el valor de similitud entre las UE.

3. Esa similitud no puede calcularse antes de que las UE y sus variables hayan sido descriptas.

En consecuencia, el orden de los pasos no puede ser alterado sin destruir la racionalidad del análisis multivariado (Crisci y López Armengol 1983).

ELECCIÓN DE LAS UNIDADES DE ESTUDIO

El primer paso en cualquier análisis multivariado consiste en elegir las UE. El investigador puede tener ante sí una gran variedad de entidades: individuos, poblaciones, especies, géneros, localidades, comunidades, sitios paleontológicos, secuencias génicas, etc. La elección de las UE entre esas entidades dependerá, en gran medida, de la estrategia y de los objetivos del estudio. Por ejemplo, si se trata establecer las relaciones entre las especies de género *Turdus* (Aves), las UE serán las especies. Si la finalidad del estudio es la variación geográfica en *Turdus rufiventris*, las unidades serán las poblaciones de esa especie. Si el objetivo fuera un estudio de comunidades de aves, las UE podrían ser especies, y las variables podrían ser caracteres morfológicos y de comportamiento. Si la finalidad es un estudio biogeográfico de áreas, las UE podrían ser cuadrículas y las variables serían las especies presentes en cada una de ellas. Por último, si el propósito es realizar un estudio filogenético molecular, las UE serán las secuencias génicas de un individuo como representantes de una especie y las variables serán los sitios de las bases nitrogenadas de ADN. Ante la imposibilidad de examinar todas las entidades posibles que componen las UE, las muestras (subconjunto de UE) son válidas.

VARIABLES

Todo análisis multivariado se basa en las diferencias existentes entre las UE a analizar. Una variable puede definirse como cualquier característica o propiedad que difiere entre las UE. Los datos son las medidas que se tomaron de las UE. Por ejemplo, si la especie A con hojas aserradas se distingue de la especie B que posee hojas enteras, la variable es “margen de la hoja”, mientras que “aserrado” y “entero” son los estados de esa variable. Podemos resumir algunos ejemplos de variables utilizadas en Biología:

1. Morfológicas

   a. Externas (presencia de tricomas, número de glándulas, longitud del pico).

   b. Internas (cóndilo hipertrofiado, presencia de células del mesófilo en empalizada).

2. Fisiológicas (concentración de hormonas, hematocrito, presencia de diapausa invernal).

3. Químicas (concentración de plomo u oxígeno, pH, salinidad).

4. Ecológicas

   a. Hábitat (uso y selección de hábitat).

   b. Interacciones bióticas (número y tipo de parásitos, polinizadores, herbívoros).

   c. Alimentación (dieta, tipo de alimento).

   d. Fenología (pico de floración, estatus migratorio).

   e. Especies (presencia-ausencia, abundancia, cobertura, biomasa).

   f. Comportamiento (mecanismo de cortejo y defensa, tamaño del territorio o “*home range*”, personalidad animal).

5. Geográficas
   a. Patrones de distribución (al azar, regular, agrupado).
   b. Relación entre poblaciones (simpatría, alopatría).
   c. Ambientales (temperatura, precipitación, humedad, altitud, tipo de suelo).
6. Genéticas
   a. Cromosomas (número y tamaño cromosómico).
   b. Secuencias de ADN/ARN (nucleares, mitocondriales, cloroplásticos, ribosómicos).
   c. Secuencias de aminoácidos (isoenzimas, aloenzimas, proteínas estructurales).

ELECCIÓN DE LAS VARIABLES

En un análisis multivariado es aconsejable elegir aquellas variables relevantes en el contexto del estudio que se intenta llevar a cabo. Solamente deben excluirse las siguientes variables:

a. Variables sin sentido biológico. Lo que se busca es evitar la utilización de aquellas variables que no tengan una relación potencialmente causal con lo que se observa en las UE. Un ejemplo es el número dado por el colector a un ejemplar de herbario.
b. Variables correlacionadas lógicamente o colineales. Se debe excluir toda propiedad que sea consecuencia lógica de otra propiedad ya utilizada. Hay que evitar la utilización por dos o más vías diferentes de la misma información. Sólo una de las variables correlacionadas es aceptada como válida. Un ejemplo de este tipo de variables es el diámetro y el radio del tallo.
c. Caracteres invariables o no informativos en las UE. Por definición, no son variables. Un ejemplo es considerar la "presencia de vértebras" como variable en un estudio de un grupo de vertebrados o la presencia de una especie cosmopolita en un estudio de regionalización en biogeografía histórica.

NÚMERO DE VARIABLES A UTILIZAR

En un análisis multivariado es imposible señalar el número mínimo de variables a utilizar. Sólo se recomienda utilizar el máximo posible dentro de las posibilidades que las UE ofrecen o presentan, y de las herramientas disponibles por el investigador. Por ejemplo, un estudio morfológico, no ofrece generalmente más de 100 variables, mientras que en un estudio de secuencias génicas puede llegar a miles. Un estudio biogeográfico de presencia-ausencia de especies de la familia Asteraceae (Angiospermae), en Patagonia ofrecería alrededor de 300 especies, mientras que un estudio mundial, cuanto menos, 20000 especies.

EL PROBLEMA DE LA IMPORTANCIA DE LAS VARIABLES

¿Existe algún tipo de variable que sea más importante que otro? Esta pregunta, *a priori*, no tiene respuesta. Todos los tipos de variables tienen la misma importancia al comenzar un análisis. Ciertas técnicas, como el análisis de componentes principales y el análisis discriminante, generan un valor *a posteriori* para cada variable en función de su poder discriminatorio. Los conceptos de datos, observaciones y variables son desarrollados en detalle en el capítulo siguiente.

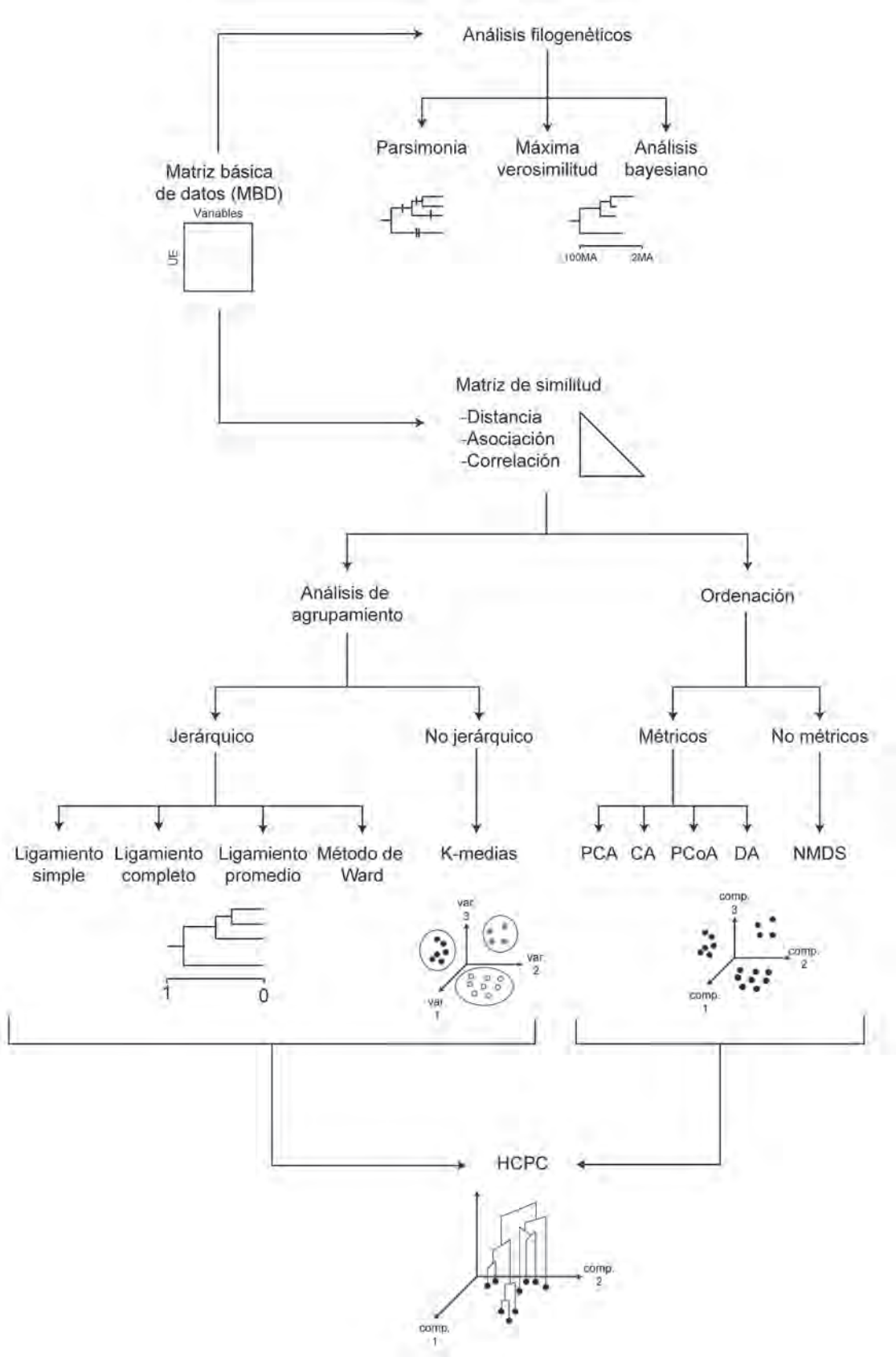


Fig. 1.2. Métodos y relaciones de los análisis multivariados abordados en el libro. PCA: análisis de componentes principales, CA: análisis de correspondencias, PCoA: análisis de coordenadas principales, DA: análisis discriminante, NMDS: escalamiento multidimensional no métrico, HCPC: agrupamiento jerárquico sobre componentes principales. El método de Ward también puede ser aplicado directamente a la MBD.

LA MALDICIÓN DE LA DIMENSIONALIDAD

Un problema general de todas las técnicas multivariadas es el conocido como "la maldición de la dimensionalidad" en la que la eficiencia y la precisión en la clasificación de las UE disminuye rápidamente a medida que aumenta el número de dimensiones o variables (Bellman 1957). Este problema se vio incrementado en las últimas décadas con las innovaciones tecnológicas que aumentaron la capacidad para obtener y procesar datos de forma masiva (Donoho 2000).

Un ejemplo en el que el número de variables excede ampliamente el número de UE son los datos de Biología Molecular, que pueden incluir cientos de marcadores genéticos obtenidos a partir de unos pocos individuos (Culhane *et al.* 2002, Cardini *et al.* 2019). Por ejemplo, una única muestra del microbioma humano puede producir cientos de millones de secuencias (variables). A principios del siglo XXI comenzó la denominada revolución "ómica" que incluye la genómica, la proteómica y la metabolómica (Evans 2000, Quackenbush 2007). Sin embargo, el análisis de estas grandes cantidades de datos plantea grandes desafíos estadísticos, computacionales (Li 2015) y epistemológicos (Mehta *et al.* 2004).

Otro ejemplo son los datos de morfometría geométrica, en los cuales los puntos anatómicos de un ejemplar (denominados *landmarks* y *semilandmarks*) tienen coordenadas en el espacio, dando como resultado dos o tres variables por punto (x, y, en un estudio en 2D o x, y, z, en un estudio en 3D; Blackith y Reyment 1971, Rohlf y Marcus 1993, Cardini *et al.* 2019). Por lo tanto, si se toman 100 *landmarks* para una especie, habrá 200 (2D) o 300 (3D) variables en la matriz. El caso de una MBD con numerosas dimensiones y pocas UE aún resulta un problema difícil de resolver, pero ha surgido una gran diversidad de técnicas para el tratamiento de este tipo de matrices (Kemsley 1996, Bouveyron *et al.* 2007a, b, Qiao *et al.* 2009).

CAPÍTULO 2
DATOS, OBSERVACIONES Y VARIABLES

Las variables o tipos de datos describen características o propiedades de las UE de un estudio (Legendre y Legendre 1998, Zar 1999). Las variables forman parte del universo denominado "datos científicos" y responden a las exigencias de éste. El científico observa hechos y los registra en forma de datos. Los hechos suceden o subsisten, son eventos y/o estados, mientras que los datos son representaciones simbólicas de estos últimos, y se obtienen por la observación (Kneller 1978). Los hechos tienen una estructura que refleja la realidad a la que llamaríamos intrínseca (Fig. 2.1). A esta estructura se le agrega una estructura extrínseca, que distorsiona la representación de la realidad en el paso que va de la observación a la representación simbólica. En ese paso el científico debe tratar de reducir la estructura extrínseca, para ello debe hacer una observación y una representación simbólica. Una observación científica debe ser sistemática, detallada y variada. Sistemática, pues debe ser controlada por una hipótesis o por una idea precisa del fenómeno estudiado (Fig. 2.1). Detallada, por el uso de instrumentos precisos y/o por concentrarse en una propiedad particular del fenómeno estudiado. Variada, ya que el fenómeno es registrado bajo diferentes condiciones o en forma experimental cuando se añade a la observación el control de ciertos factores.

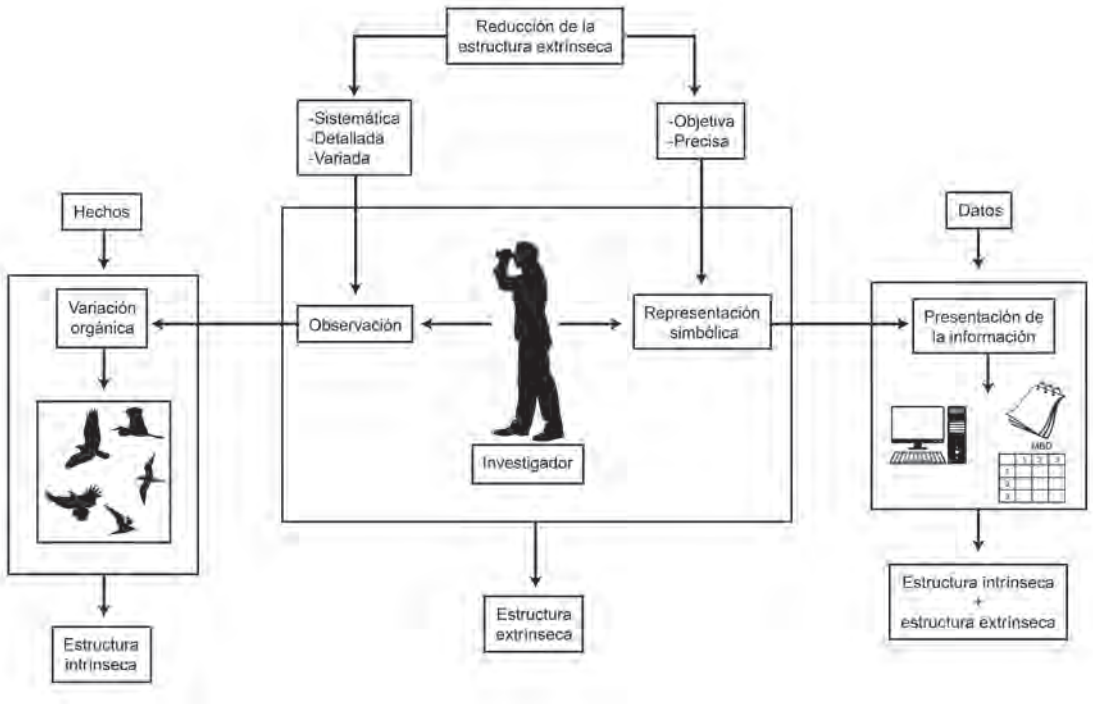


Fig. 2.1. Las variables como datos científicos.

Los datos obtenidos por la observación deben ser objetivos y precisos. Objetivos (a pesar de que no existe dato que no esté sesgado por nuestros preconceptos) en el sentido de que cualquier otro científico, capacitado para la observación y que lleve a cabo las mismas operaciones, logre reconocer los mismos hechos que fueron registrados y, por lo tanto, obtenga los mismos datos. Con este fin, los datos son expresados en un lenguaje de validez universal, más que en función de sensaciones propias del observador. Los datos son precisos cuando describen los hechos y los diferencian, en el mayor grado posible, de hechos similares. Los datos más objetivos y precisos son los expresados en forma cuantitativa. La estadística multivariada exige que todos los datos sean expresados en forma cuantitativa, de modo que sean computables; es decir, que con ellos se puedan realizar operaciones de cálculos mediante números. Sin embargo, no todos los datos miden relaciones cuantitativas en sentido estricto, de ahí que algunos deban ser sometidos a la codificación, para ser transformados en datos cuantitativos (ver en este capítulo *Transformaciones entre tipos de datos*).

TIPOS DE DATOS

Existen distintos tipos de datos o variables y se han propuesto numerosas clasificaciones para expresar esa variabilidad (Walker 1968, Bunge 1969, Cohen y Nagel 1971, Sneath y Sokal 1973, Clifford y Stephenson 1975, Zar 1999). La clasificación propuesta en este libro puede observarse en la Tabla 2.1, y es una combinación entre la clasificación de Legendre y Legendre (1998) y la de Whitlock y Schluter (2015). A continuación, y basándonos en esa clasificación, examinaremos los distintos tipos de datos y su posible codificación. Los datos cualitativos se pueden expresar numéricamente, lo que se denomina codificación. El número aquí desempeña una función de rótulo o marca de identificación que facilita el tratamiento cuantitativo, pero los datos también podrían expresarse utilizando los símbolos + (presencia) y – (ausencia), o cualquier otra forma convencional. Las distintas propiedades de las variables son también denominadas “estados”.

Tabla 2.1. Tipos de datos y ejemplos.

Tipos de datos		**Ejemplo**	**Estados**
1. Cualitativos o categóricos	1.1. Nominales	Sexo cromosómico	XX, XY, XXY, X0, XYY
		Presencia-ausencia de una especie en un área	Presencia, ausencia
	1.2. Ordinales	Pubescencia de la hoja	Glabra, pelos poco abundantes, pelos muy abundantes
		Grado de disturbio	Bajo, medio, alto
2. Cuantitativos o numéricos	2.1. Continuos	Longitud del abdomen	10 mm, 10,2 mm, 20,1 mm, …
		Temperatura del cuerpo	36,2 °C, 37,1 °C, …
	2.2. Discretos	Número de inflorescencias	1, 2, 3, 10, …
		Número de tipos de aminoácidos en una proteína	10, 15, 20, …

Datos cualitativos o categóricos

Los datos cualitativos establecen categorías (cualidades no mensurables) para caracterizar a las unidades de estudio (UE), cuyos estados pueden estar ordenados (siguen una secuencia) o no (Tabla 2.1).

Datos nominales

Establecen categorías cuyos estados no están ordenados.

Datos doble-estado, estados excluyentes

También llamados datos binarios o dicotómicos, tienen sólo dos estados, no ordenados. Se suelen expresar numéricamente como 0 y 1, otorgándole 0 ó 1 a cualquiera de los estados. Los estados son mutuamente excluyentes. El número aquí es utilizado en forma convencional y a manera de rótulo (Tabla 2.2). En la práctica, ambas distinciones generan los mismos resultados.

Tabla 2.2. Ejemplos de codificación para datos doble-estado.

Variable	Estados	Codificación 1	Codificación 2
Rayas en el abdomen	Presencia de rayas	0	1
	Ausencia de rayas	1	2
Tipo de fruto	Indehiscente	0	1
	Dehiscente	1	2

Datos multiestado

Son aquellos datos que poseen tres o más estados. Son datos cualitativos que no pueden ser ordenados en una secuencia de grados de la variable. Por ejemplo, la variable "tipo de pubescencia de la hoja" con los siguientes estados: escabrosa, estrigosa, híspida, hirsuta, serícea. Es el tipo de dato más difícil de codificar, ya que por no presentar una secuencia o grados es imposible representarlo satisfactoriamente con números. En el ejemplo podemos atribuir 1, 2, 3, 4 y 5 a los estados nombrados en el orden dado. Los números cumplen aquí funciones convencionales y de rótulo. Pero ¿por qué otorgar 5 al estado serícea y no 2? Es decir, carecemos de fundamentos para establecer un orden. En el caso de los datos doble-estado este inconveniente no existe, ya que son posibles dos secuencias equivalentes entre sí. El problema podría resolverse transformando cada uno de los estados en datos presencia-ausencia (Tabla 2.3). Esta solución tiene el inconveniente de que confiere mayor peso a la variable original (en este caso "tipo de pubescencia de la hoja"). Ésto se debe a que al transformarse en varios caracteres independientes, aumenta su valor en perjuicio de caracteres no codificados de esta forma. Los programas filogenéticos permiten asignar el mismo valor a cada cambio posible de este tipo de variable.

Tabla 2.3. Codificación para datos nominales multiestado.

Variable	Estado	Codificación
Superficie de la hoja escabrosa	Ausente	0
	Presente	1
Superficie de la hoja estrigosa	Ausente	0
	Presente	1

Superficie de la hoja híspida	Ausente	0
	Presente	1
Superficie de la hoja hirsuta	Ausente	0
	Presente	1
Superficie de la hoja serícea	Ausente	0
	Presente	1
Superficie de la hoja estrellada	Ausente	0
	Presente	1

Datos ordinales

Es el caso de variables cualitativas que pueden ser ordenadas en una secuencia lógica. Si se codifica el ejemplo anterior, observamos que existen distintas posibilidades, de las cuales damos dos ejemplos (Tabla 2.4). Aquí los números desempeñan una función ordinal, ya que indican la posición en una escala establecida (otra codificación válida consistiría en otorgar los números 1, 1,5 y 2 a los estados "escasa", "común" y "abundante"). Los valores ordinales correspondientes a cada estado de la variable no pueden someterse a operaciones aritméticas como la adición. Por eso, no es posible afirmar que el estado "abundante" tiene tres veces más pelos que el estado "escasa". Sin embargo, cuando se usa este sistema de rangos se asume que la distancia entre una categoría y la siguiente es la misma, independientemente de los valores medidos (por ejemplo, 1 pelo –escasa–, 10 pelos –común– y 50 pelos –abundante–), por lo tanto siempre hay pérdida de información.

Tabla 2.4. Codificación para datos ordinales.

Variable	Estado	Codificación 1	Codificación 2
Abundancia de pelos	Escasa	1	1
	Común	2	1,5
	Abundante	3	2

Datos cuantitativos o numéricos

Miden relaciones cuantitativas en sentido estricto. El campo de variabilidad de los datos es un conjunto de números. Los números indican relaciones cuantitativas entre cualidades y, por lo tanto, pueden ser sometidos a operaciones matemáticas (aunque a veces con restricciones, por ejemplo, pueden sumarse los números de estomas de dos hojas, pero no las densidades de estomas de esas dos hojas). En general, estos datos no necesitan codificación y pueden ser de dos categorías: continuos o discretos.

Datos continuos

Expresan dimensiones continuas, es decir, cualidades cuya variabilidad numérica se distribuye en una escala continua. La expresión de estos datos es un número real. Ejemplos de este tipo de datos son todas las expresiones de tamaño: "longitud de la hoja", "altura de la planta", "ancho del pétalo", "longitud del pico".

Datos discretos

Representan cualidades que son expresables sólo por números enteros; por ejemplo, la variable "número de pétalos", "número de glándulas".

CODIFICACIÓN DE DATOS CUANTITATIVOS

Los datos cuantitativos pueden ser codificados como datos presencia-ausencia si los cálculos así lo exigen, como veremos más adelante al tratar los coeficientes de asociación (Cap. 4). Un ejemplo podría ser la variable "número de pétalos" con variabilidad entre 2 y 10, que puede ser subdividida en intervalos que se transformarán en caracteres binarios (Tabla 2.5). Las variables con esta codificación se denominan ficticias o *dummies* (Hardy 1993).

Tabla 2.5. Transformación y codificación de una variable discreta a nominal.

Variable	Estado	Codificación
De 2 a 4 pétalos	Ausente	0
	Presente	1
De 5 a 7 pétalos	Ausente	0
	Presente	1
De 8 a 10 pétalos	Ausente	0
	Presente	1

Los datos continuos como, por ejemplo, la variable "longitud de la hoja" con variabilidad entre 2 y 10 cm pueden ser transformados como se muestra en la Tabla 2.6.

Tabla 2.6. Transformación y codificación de una variable continua a nominal.

Variable	Estados	Codificación
Longitud de la hoja entre 2 y 3,99 cm	Ausente	0
	Presente	1
Longitud de la hoja entre 4 y 5,99 cm	Ausente	0
	Presente	1
Longitud de la hoja entre 6 y 7,99 cm	Ausente	0
	Presente	1
Longitud de la hoja entre 8 y 10 cm	Ausente	0
	Presente	1

Esta transformación es a veces necesaria, por ejemplo para aplicar los coeficientes de asociación (Cap. 4). Sin embargo, tiene el inconveniente del peso que otorga a la variable (en este caso "longitud de la hoja") y de la relativa arbitrariedad al determinar los intervalos.

EL PROBLEMA DE LA VARIACIÓN INTRA-UNIDADES DE ESTUDIO

Si bien las UE deben ser lo más homogéneas posible, hay variación dentro de ellas (Hallgrímsson y Hall 2005, Herrera 2009, 2017, Palacio *et al.* 2017). Por lo tanto, la pregunta que surge es: ¿cuál es el tratamiento que debe recibir esa variación intra-UE? El problema admite, al menos, tres soluciones posibles; dos son aplicables a datos cuantitativos y una a datos cualitativos.

Una primera solución consiste en considerar que la variación intra-UE puede reducirse a una medida de posición estadística, por ejemplo, la media, la mediana o la moda (Box 2.1). La media es utilizada para datos cuantitativos continuos; sin embargo, no es recomendable para datos cualitativos ordinales y cuantitativos discretos, pues el valor obtenido podría no corresponderse con un valor real (por ejemplo, una media de 3,5 para la variable número de pétalos). En este caso, es más apropiada la mediana. La moda es el valor más frecuente del conjunto de datos. La media y la mediana no reflejan la dispersión de los valores, por lo que algunos autores (Sneath y Sokal 1973) complementan esta solución agregando como nueva variable una medida de dispersión, como por ejemplo el desvío estándar. La variable original se transforma, entonces, en dos variables para cada UE: la media (o moda) y el desvío estándar.

Una segunda solución al problema de la variación intra-UE, consiste en elegir al azar un organismo de los que componen la UE, para considerar que los estados presentes en ese organismo son los estados representativos de la UE. Con esta solución (denominada método del ejemplar) se corre el riesgo de elegir un organismo que presente estados atípicos para la UE en cuestión.

Los datos doble-estado (y los multiestados cualitativos sin secuencia lógica, transformados a doble-estado, ver en este capítulo *Datos multiestados y su codificación*) exigen una solución diferente a las anteriores. Esta solución se encuentra mediante la codificación, expresando la variación intra-UE como un estado independiente. Por ejemplo, en un estudio de taxonomía numérica de géneros de la subtribu Nassauviinae (Crisci 1974), se utilizó la variable "pubescencia en el receptáculo" que originalmente tiene dos estados posibles: presente (1) y ausente (0). Sin embargo, se encontró que algunos géneros incluían especies con pubescencia en el receptáculo y especies sin pubescencia. La codificación que tuvo en cuenta esa variación se muestra en la Tabla 2.7.

Tabla 2.7. Codificación de la variable "pubescencia en el receptáculo" en géneros de la subtribu Nassauviinae (Crisci 1974).

Variable	Estados	Codificación
Pubescencia en el receptáculo	Ausente	0
	Ausente y presente	1
	Presente	2

Una alternativa, sería utilizar como variable el porcentaje de especies de cada género que presentan pubescencia en el receptáculo. Por ejemplo, si un género tiene tres especies con pubescencia de un total de cinco especies, tomará el valor de 3/5. En este caso la variable sería discreta, ya que el cálculo resulta de dividir el número de especies de un género con pubescencia respecto del total de especies de ese género.

ESTUDIO DE CASO: ANÁLISIS MULTIVARIADO DEL GÉNERO *BULNESIA* (ZYGOPHYLLACEAE)

Analizaremos un ejemplo tomado de un estudio realizado en el género *Bulnesia* (Crisci *et al.* 1979). Las ocho especies del género *Bulnesia* (*B. arborea*, *B. bonariensis, B. carrapo*, *B. chilensis, B. foliosa, B. retama, B. sarmientoi* y *B. schickendantzii*) representan las UE (Fig. 2.2). Los datos consisten en 43 variables morfológicas registradas para las ocho UE. El conjunto de caracteres comprende 19 cualitativos, 19

cuantitativos continuos y cinco cuantitativos discretos (Tabla 2.8). Los valores que se volcaron en la matriz básica de datos (MBD) corresponden a la media en el caso de los caracteres cuantitativos continuos y a la moda en el caso de los cuantitativos discretos (con excepción de la variable "numero de óvulos", en la que se registró la media). Las variables cualitativas fueron codificadas (Tabla 2.8).

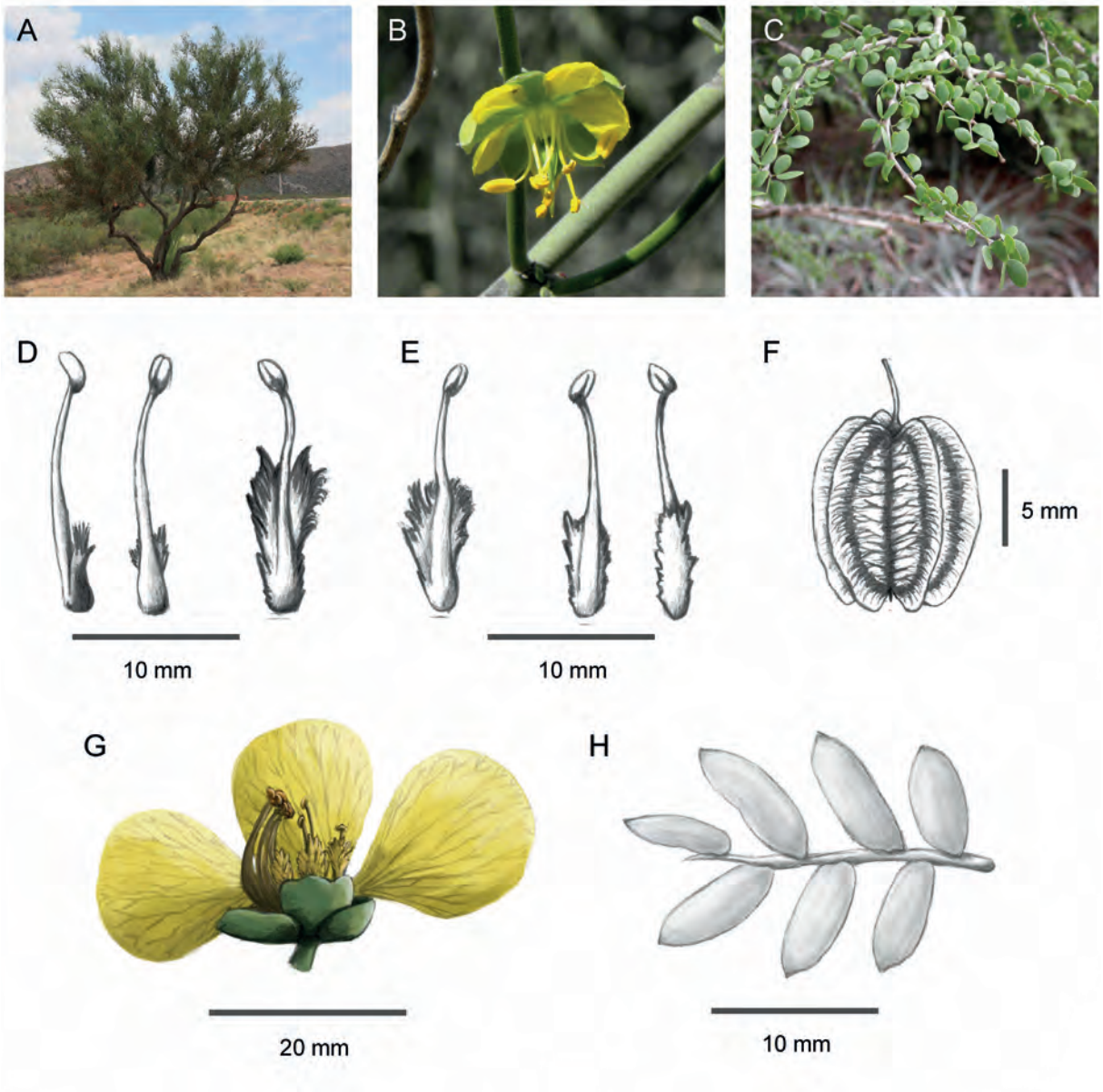


Fig. 2.2. (A) y (B) *Bulnesia retama*; (C) *B. sarmientoi*; (D) estambres de *B. arborea*; (E) estambres de *B. carrapo*; (F) fruto de *B. schickendantzii*; (G) flor de *B. carrapo* (se muestran sólo tres carpelos); (H) hojas de *B. retama*. Fotografías: Varela, O (A) y Blendinger, PG (B y C). Ilustraciones: Colombo, M (modificadas de Palacios y Hunziker 1984 y Novara 2012).

Tabla 2.8. Variables, estados y codificación de ocho especies del género *Bulnesia* (Crisci *et al.* 1979). Las variables que presentan guiones corresponden a variables cuantitativas que tienen múltiples estados y no requieren codificación.

Variable	Estados	Codificación
1. Hábito	Arbustos	0
	Arbustos y árboles	1
	Árboles	2
2. Longitud del internodio (cm)	-	-
3. Diámetro del internodio (cm)	-	-
4. Longitud de la hoja (cm)	-	-
5. Ancho de la hoja (cm)	-	-
6. Longitud del pecíolo (cm)	-	-
7. Número de folíolos	-	-
8. Presencia de peciólulos	Folíolos no sésiles	0
	Folíolos sésiles o no sésiles	1
	Folíolos sésiles	2
9. Disposición de los folíolos en el raquis	Folíolos alternos	0
	Folíolos subopuestos	1
	Folíolos opuestos	2
10. Pubescencia de la hoja	Ausente	0
	Ausente y presente	1
	Presente	2
11. Longitud del folíolo (mm)	-	-
12. Ancho del folíolo (mm)	-	-
13. Número de nervaduras primarias del folíolo	-	-
14. Posición de los folíolos terminales	Paralelos	0
	Paralelos y divergentes	1
	Divergentes	2
15. Presencia de mucrón en folíolos	Folíolos no mucronados	0
	Folíolos mucronados	1
16. Tipo de inflorescencia	Flores solitarias	1
	Inflorescencia en dicasio	2
17. Longitud del pedúnculo (mm)	-	-
18. Longitud del sépalo (mm)	-	-
19. Ancho del sépalo (mm)	-	-

Variable	Estados	Codificación
20. Color de los pétalos	Blanco	1
	Amarillo	2
21. Longitud del pétalo (mm)	-	-
22. Ancho del pétalo (mm)	-	-
23. Número de nervaduras del pétalo	-	-
24. Tipo de estambres	Heterogéneos	0
	Heterogéneos y homogéneos	1
	Homogéneos	2
25. Modificación de los estambres	No modificados	0
	Modificados y no modificados	1
	Modificados	2
26. Presencia de gran escama junto al estambre	Ausente	0
	Ausente y presente	1
	Presente	2
27. Presencia de pelos en la base del filamento estaminal	Sin pelos	0
	Con o sin pelos	1
	Con pelos	2
28. Presencia de una escama suplementaria junto al estambre	Ausente	0
	Presente	1
29. Agrupación de los estambres	No agrupados	0
	Agrupados o no agrupados	1
	Agrupados	2
30. Longitud del filamento (mm)	-	-
31. Longitud de la antera (mm)	-	-
32. Longitud de la escama (mm)	-	-
33. Presencia de ápice laciniado en la escama estaminal	Sin ápice laciniado	0
	Con o sin ápice laciniado	1
	Con ápice laciniado	2
34. Número de carpelos	En número de 3	0
	En número de 3 y 5	1
	En número de 5	2
35. Curvatura del estilo	Estilo no curvado	0
	Estilo curvado o no curvado	1
	Estilo curvado	2

Variable	Estados	Codificación
36. Número de óvulos por carpelo	-	-
37. Pubescencia del fruto	Glabro	0
	Pubescente	1
38. Longitud del fruto (mm)	-	-
39. Ancho del fruto (mm)	-	-
40. Desarrollo del carpóforo	Reducido	1
	Bien desarrollado	2
41. Longitud del carpóforo (mm)	-	-
42. Forma de la semilla	Semicircular o semielíptica	1
	Oblongo-reniforme	2
43. Longitud de la semilla (mm)	-	-

En algunas de las variables cualitativas, como el hábito, con dos estados (árbol y arbusto) es posible hallar especies cuyos individuos son todos árboles, especies cuyos individuos son todos arbustos y especies que presentan individuos árboles e individuos arbustos. Una codificación para esta variable es la que se muestra en la Tabla 2.9. Otra alternativa sería la que se muestra en la Tabla 2.10.

Tabla 2.9. Alternativa 1 para la codificación de la variable "hábito".

Variable	Estado	Codificación
Hábito	Arbusto	0
	Arbusto y árbol	1
	Árbol	2

Tabla 2.10. Alternativa 2 para la codificación de la variable "hábito".

Variable	Estado	Codificación
Arbusto	Ausente	0
	Presente	1
Árbol	Ausente	0
	Presente	1

Los datos obtenidos se presentan en forma de una matriz básica de datos (MBD), la cual representa la materia prima de todo análisis multivariado. Esta es una matriz de $n \times p$ (Fig. 2.3), donde las n filas representan las UE y las p columnas representan las variables (Quinn y Keough 2002). Cada celda de la matriz X_{ij} representa el valor de la UE i para la variable j. Cabe mencionar que la MBD puede contener datos faltantes (*missing data*), que pueden deberse a dos situaciones:

a. Variables que no pueden ser medidas debido a la naturaleza de lo que se intenta medir. Por ejemplo: la longitud de la hoja de una especie áfila en un estudio del género donde el resto de las especies tienen hojas.

b. Variables que pueden ser medidas pero que debido a situaciones aleatorias la medición no pudo concretarse. Por ejemplo: ejemplar incompleto, rotura de un aparato de medición o eventos climáticos en el momento de la medición.

Para estos datos faltantes se aplican distintos símbolos de acuerdo al software que se utilice, por ejemplo en R (ver Cap. 3) se utiliza NA (*not available*). Los distintos software utilizan distintas estrategias para el tratamiento de estos datos, la más común es ignorar la celda con ese dato durante los cálculos. La MBD para las especies de *Bulnesia* se muestra en la Tabla 2.11.

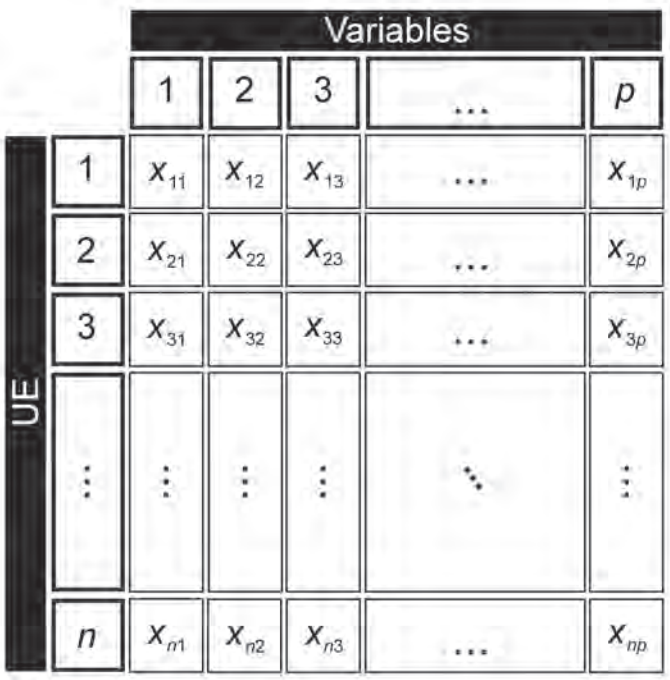


Fig. 2.3. MBD de UE (filas) × variables (columnas). Representa la materia prima para todo análisis multivariado.

Tabla 2.11. MBD de especies de *Bulnesia* × caracteres (Crisci *et al.* 1979). Ver la Tabla 2.8 para la nomenclatura de cada variable. NA: dato no disponible.

Especie	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21
B. arborea	2	35	2,1	85	57	7,7	13	2	0	2	30	8,6	6	0	1	2	17	7,1	3,4	2	22
B. carrapo	2	36	1,6	97	71	9	7	2	0	2	40	16	6	1	1	2	18	6,4	5,8	2	24
B. chilensis	0	24	2,6	14	8,9	1,8	8	2	1	0	5,2	2,4	NA	2	1	1	9	7,1	4,2	2	9,8
B. bonariensis	0	20	1,3	26	18	3,4	14	1	0	2	8,9	2	1	2	1	1	12	6,8	3,9	2	18
B. retama	1	40	2	13	11	3,1	5	1	1	2	6,6	2,6	2	1	2	1	10	7,4	4,4	2	7,7
B. foliosa	0	19	1,3	28	25	5,8	4	1	1	2	14	7,8	3	2	1	1	13	5,3	3	2	8,5
B. schickendantzii	0	10	1,9	20	12	2,7	10	0	0	2	5,7	1,9	1	2	1	1	10	5,9	3,1	2	9,3
B. sarmientoi	2	22	1,4	21	27	5,1	2	2	2	1	17	12	5	2	1	1	3,9	2,9	2,3	1	12

Especie	C22	C23	C24	C25	C26	C27	C28	C29	C30	C31	C32	C33	C34	C35	C36	C37	C38	C39	C40	C41	C42	C43
B. arborea	17	16	0	1	2	0	1	2	10	1,5	5,2	2	2	1	2	1	46	41	2	7,7	1	13
B. carrapo	19	12	0	2	2	0	1	2	9,9	1,4	53	2	2	2	2	1	56	52	2	5,3	1	12
B. chilensis	5,7	8	0	0	0	2	1	2	7	2	4,4	1	1	NA	7	1	13	12	2	0,6	2	2,7
B. bonariensis	10	10	0	1	1	0	1	1	11	1,6	4,4	0	2	2	1	1	36	33	2	4,8	1	11
B. retama	4,6	7	1	0	0	1	1	1	7,1	2,2	3,1	1	2	1	8	1	23	19	1	0,8	2	11
B. foliosa	2,7	6	2	0	0	0	1	0	6,2	1,6	3,9	1	2	1	4	2	16	13	1	0,7	2	4,9
B. schickendantzii	4,3	5	1	0	0	0	2	0	7,4	1,7	4,4	2	2	1	4	2	12	13	1	0,4	2	5,3
B. sarmientoi	7	6	2	0	0	0	1	0	4,1	1,1	2,9	2	0	0	2	1	52	48	2	5,2	1	14

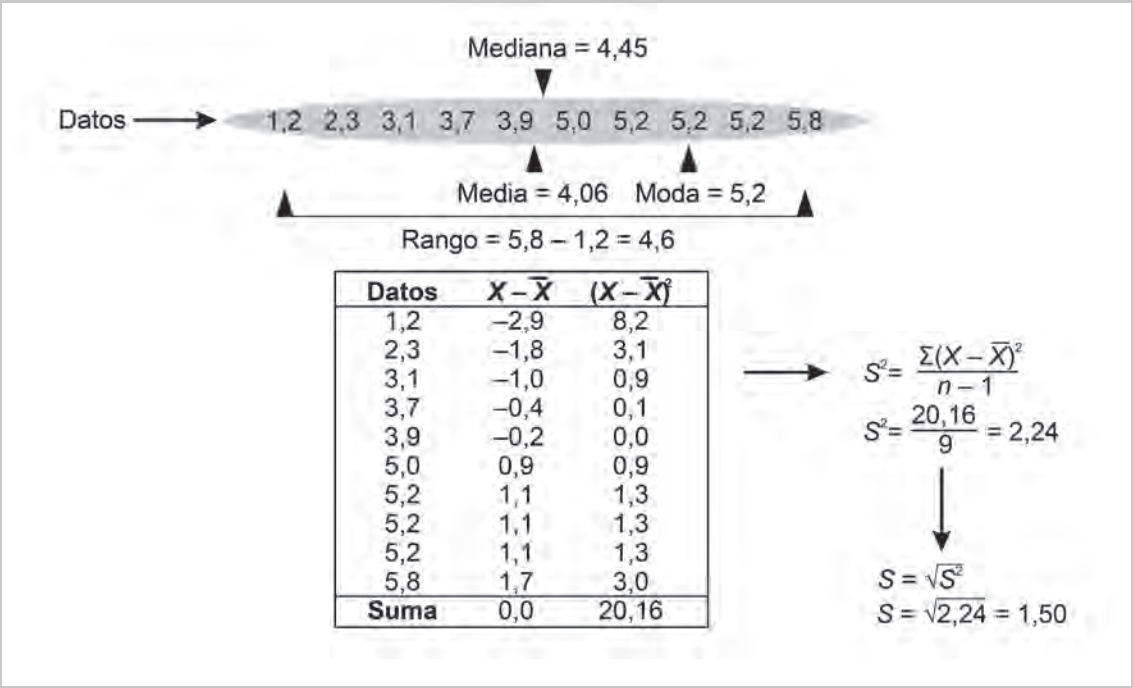
Box 2.1. Descriptores estadísticos de un conjunto de observaciones o datos

Medidas de tendencia central: describen el centro de la distribución del conjunto de datos (Whitlock y Schluter 2015), entre las que podemos mencionar por ejemplo:

- Media aritmética ($\bar{X}$): se obtiene a partir del cociente entre la suma de todos los valores y el número de observaciones.
- Mediana: los datos se ordenan de menor a mayor y se selecciona aquel valor que se encuentra en la posición central. Si hay un número impar de medidas la mediana es la observación central. Si hay un número par de valores, la mediana es la media de las dos observaciones centrales.
- Moda: es el valor más frecuente del conjunto de datos.

Medidas de dispersión: describen la variabilidad del conjunto de datos. Algunas de ellas son, por ejemplo:

- Varianza (S^2): expresa la variabilidad del conjunto de datos con respecto a la media aritmética.
- Desvío estándar (S): raíz cuadrada de la varianza. El valor expresado se encuentra en las mismas unidades que la variable original.
- Rango: diferencia entre el valor máximo y el mínimo del conjunto de datos.



Datos	$X - \bar{X}$	$(X - \bar{X})^2$
1,2	–2,9	8,2
2,3	–1,8	3,1
3,1	–1,0	0,9
3,7	–0,4	0,1
3,9	–0,2	0,0
5,0	0,9	0,9
5,2	1,1	1,3
5,2	1,1	1,3
5,2	1,1	1,3
5,8	1,7	3,0
Suma	0,0	20,16

$$S^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$

$$S^2 = \frac{20,16}{9} = 2,24$$

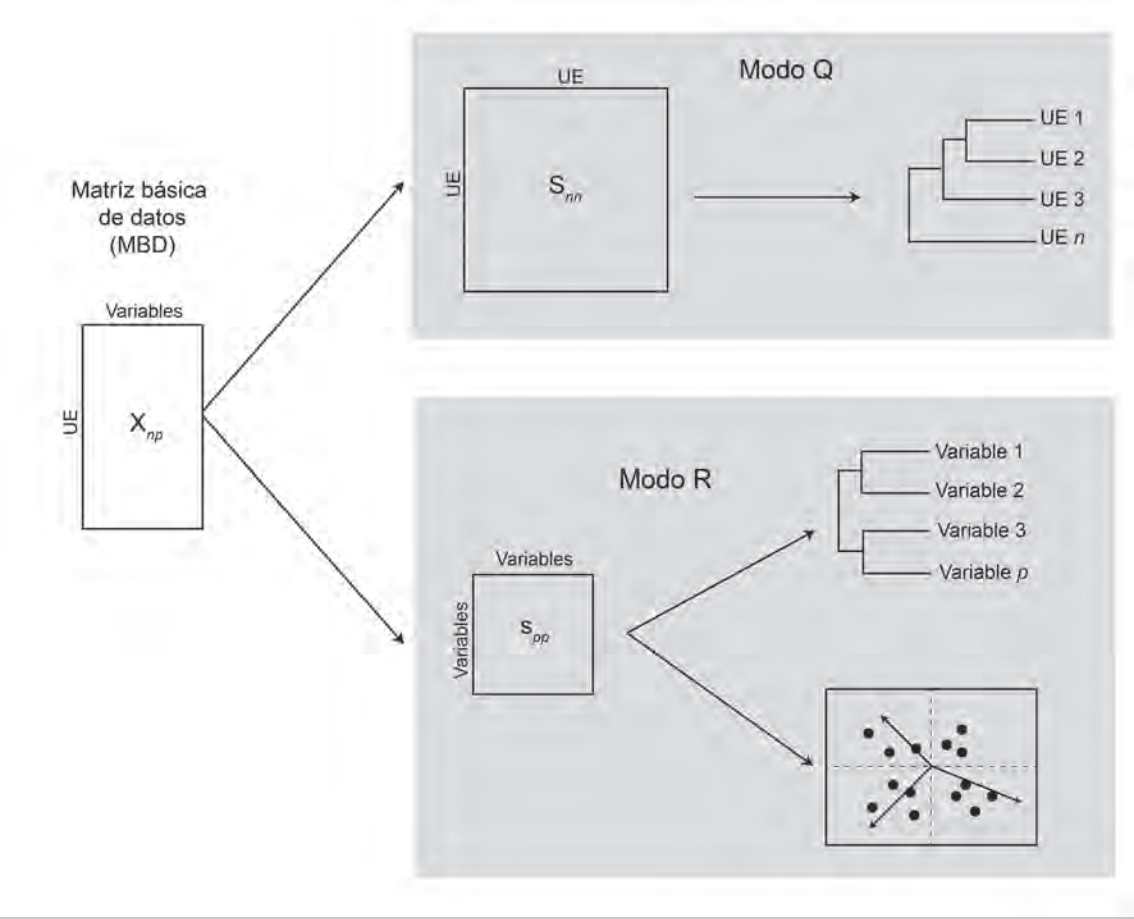
$$S = \sqrt{S^2}$$

$$S = \sqrt{2,24} = 1,50$$

Box 2.2. Modos Q y R

Una MBD puede ser estudiada desde dos puntos de vista (Cattell 1952, Legendre y Legendre 1998): el de asociación de las UE (técnica o modo Q) o el de asociación de las variables (técnica o modo R, no confundir con el software R). Sin embargo, a veces no resulta obvio si un análisis fue realizado mediante el modo Q o R. En la práctica no suele hacerse esta distinción y no es necesaria para realizar el análisis, pero se describen aquí con fines conceptuales.

A modo de ejemplo, el resultado del análisis de agrupamientos (dendrograma) suele construirse mediante el modo Q (ver un ejemplo de dendrograma mediante el modo R en el Cap. 5). En las técnicas de ordenación, si bien suelen graficarse las UE en el espacio, el análisis suele realizarse mediante el modo R. Este modo puede generar ideas e hipótesis acerca del origen y patrones de agrupamiento de las variables. En el estudio de caracteres, por ejemplo, permite determinar posibles complejos adaptativos a partir de los grupos formados (Pigliucci 2003, Pigliucci y Preston 2004, Ordano *et al*. 2008, Singh *et al*. 2012), denominados módulos o pléyades de correlación (*sensu* Terentjev 1931). Otro ejemplo de aplicación del modo R sería una matriz de localidades × especies (variables), donde se reunirían conjuntos de especies que definen un grupo de localidades mediante su co-ocurrencia (Kreft y Jetz 2010).



CAPÍTULO 3
INTRODUCCIÓN AL LENGUAJE R

R (R Core Team 2018) es un lenguaje de programación libre y gratuito, que se ha convertido en una herramienta fundamental y ampliamente utilizada como software estadístico en prácticamente todas las disciplinas del ámbito científico moderno. Esto se debe a que presenta numerosas ventajas: actualmente es el software que ofrece más funciones estadísticas y aplicaciones para la creación de gráficos de alta calidad, posee una amplia comunidad de usuarios, y constituye un lenguaje universal para el intercambio de ideas y líneas de código (Ihaka y Gentleman 1996, Salas 2008).

R fue presentado oficialmente en 1997 y se rige por la licencia general pública (*General Public License* o GPL) de la fundación de software libre (*Free Software Foundation*, sistema operativo GNU, http://www.gnu.org/). Libre hace referencia a la libertad de los usuarios para ejecutar, copiar, distribuir, estudiar, cambiar y mejorar el software. Es muy similar al programa estadístico S-plus (el cual no es gratuito y es distribuido por Insightful Corporation), ya que la implementación base y semántica de ambos son derivados de un lenguaje estadístico llamado S y de un lenguaje llamado Scheme (Ihaka y Gentleman 1996).

En este capítulo se pretende dar una breve base introductoria sobre el manejo de R, lo que permitirá realizar los análisis presentados a lo largo del libro. Al final de cada capítulo se presentarán rutinas con las cuales el lector podrá ejercitar y aplicar a bases de datos propias. Todos los conjuntos de datos y las rutinas están disponibles en: https://fundacionazara.org.ar/analisis-multivariado-para-datos-biologicos/. El lector que se inicia en el software puede sentirse abrumado en un principio, debido a la cantidad de conceptos nuevos y a la propia lógica del software basada en comandos, a la cual no solemos estar acostumbrados. Por lo tanto, deberá armarse de tiempo y paciencia para ir adquiriendo poco a poco los conocimientos y la práctica que el programa requiere. En este sentido, intentaremos utilizar líneas de código que consideramos sencillas, pero debe tenerse en cuenta que no son las únicas y que un mismo análisis puede realizarse de múltiples formas diferentes. Si el lector desea profundizar en aspectos introductorios, se recomienda la lectura de Paradis (2002), Logan (2011), Crawley (2012) y Wickham y Grolemund (2016).

DESCARGA DEL SOFTWARE

El programa puede descargarse del siguiente enlace: https://www.r-project.org/ y está disponible para los sistemas operativos Windows, Unix y GNU/Linux. Como otros software basados en comandos, R tiene una interfaz poco amigable. Por lo tanto, los análisis serán presentados utilizando la interfaz complementaria RStudio, un entorno más amigable (disponible en https://www.rstudio.com/). RStudio presenta herramientas de visualización que facilitan el manejo de datos, ofrece versiones para cualquier sistema operativo y, al igual que R, es gratuito (RStudio 2012). Es importante tener en cuenta que para utilizar RStudio previamente debe instalarse R. Ambos programas pueden configurarse en español o en inglés.

INTRODUCCIÓN A LA INTERFAZ RSTUDIO

La sesión de RStudio, por defecto, se divide en cuatro ventanas básicas que pueden ser personalizadas por el usuario (Fig. 3.1):

1. Ventana inferior izquierda (*Console*). Corresponde a la consola. En ésta se ejecutan los comandos y se visualizan las salidas (resultados de los análisis y operaciones que se ejecuten).
2. Ventana superior izquierda (rutinas y visualización de bases de datos). En esta ventana puede tenerse registro de las rutinas (*scripts*), que constituyen un conjunto de comandos o líneas de código orientados a un análisis específico (por ejemplo, análisis de componentes principales). Cada línea de código puede ejecutarse en la consola con el ícono *Run* (parte superior derecha). Así, el usuario puede ir construyendo sus propias líneas de código e ir corrigiéndolas si hay errores (que se manifiestan en la consola). Por lo tanto, las funciones pueden ejecutarse escribiendo en la consola o en esta ventana. Si a una línea de texto se le antepone el símbolo #, dicha línea no será leída por el software, lo que generalmente es utilizado para agregar notas personales (por ejemplo, *#Primer análisis de mi tesis*). Las rutinas son guardadas en formato .R y pueden ser cargadas por cualquier usuario en su sesión.
3. Ventana superior derecha (*Workspace* y *History*). La pestaña *Workspace* muestra todos los objetos activos, mientras que la pestaña *History* muestra el historial de los comandos ejecutados en la consola en una determinada sesión.
4. Ventana inferior derecha (*Files*, *Plots*, *Packages* y *Help*). La pestaña *Files* lista los archivos que se encuentran en la PC. La pestaña *Plots* muestra los gráficos. La pestaña *Packages* contiene una lista con todos los paquetes instalados, y es la pestaña a partir de la cual pueden instalarse más paquetes (*Install*). Los paquetes son conjuntos de funciones, datos y códigos compilados, orientados a un análisis en particular. La pestaña *Help* contiene un buscador con las ayudas de los paquetes y funciones disponibles en la sesión (es decir, si se quiere ayuda sobre un paquete que está instalado pero no cargado, no se podrá obtener ayuda).

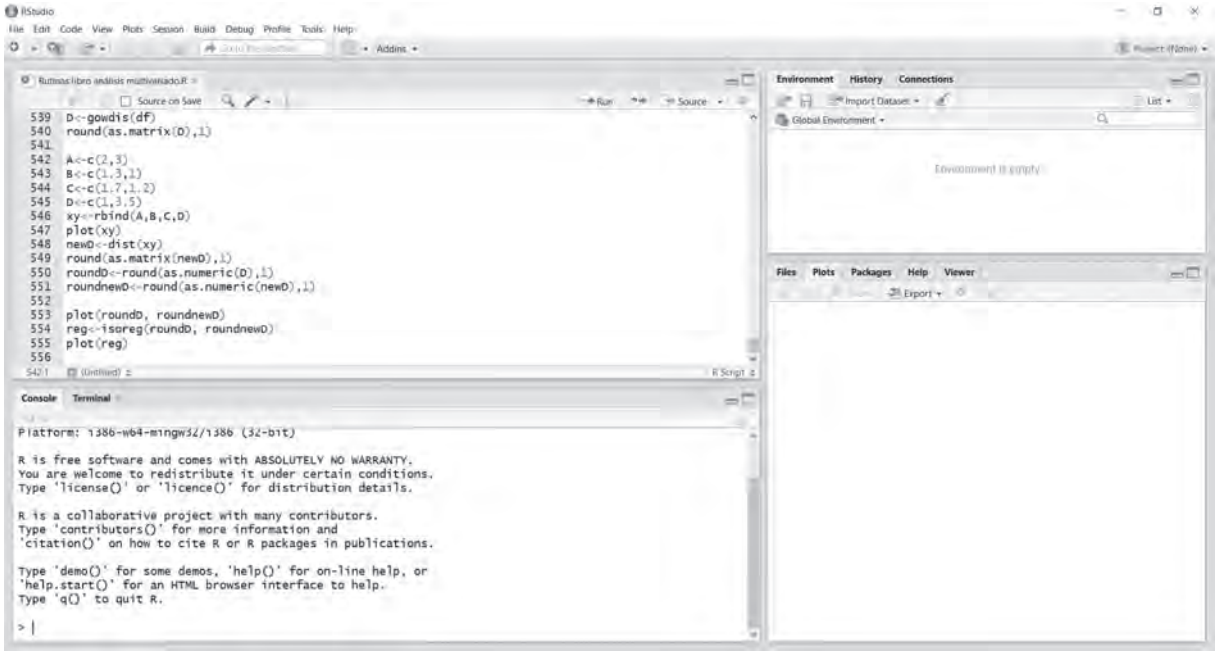


Fig. 3.1. Sesión de RStudio.

OBJETOS: VECTORES Y MARCOS DE DATOS

R es un lenguaje orientado a objetos. Esto significa que cada comando crea un objeto que debe ser nombrado para que permanezca en la memoria, y su nombre debe ser ejecutado para que aparezca en la

consola. El comando más básico es el comando "asignar" o "flecha" `<-`, y es el que permite la creación de nuevos objetos. El nombre del objeto se encuentra a la izquierda del comando "asignar". El objeto puede llevar cualquier nombre (sin embargo, no pueden ser sólo números, comenzar con números, ni contener guión del medio). Debe tenerse en cuenta que R distingue mayúsculas de minúsculas. Además, si se utiliza un mismo nombre para designar un nuevo objeto, el primero será reemplazado por el segundo. Si no se designa un nombre, el objeto no se guardará. Una vez escrito un comando, el mismo puede ejecutarse mediante la tecla *Enter* si se escribió en la consola, o haciendo click en *Run* si se escribió en la ventana de rutinas.

```
> 5 + 5
[1] 10
> a <- 10 # Resultado de 5 + 5
> a
> [1] 10
```

Los caracteres categóricos deben ingresarse entre comillas.

```
> obj <- "A"
> obj
[1] "A"
```

Es importante resaltar que el símbolo > no es parte del comando, es producido por R para indicar que el software está listo para que se ingrese un comando. El número entre corchetes representa un identificador de los valores dentro del objeto, en este caso indica que hay un único elemento cuyo primer valor es `"A"`.

El tipo de objeto más básico es el vector. Un vector es una secuencia de elementos o componentes de un mismo tipo. En este libro usaremos tres tipos básicos: vectores enteros (secuencia de números enteros), vectores numéricos (secuencia de números reales) y vectores de caracteres (en el marco del lenguaje informático, carácter se refiere a una letra o a un signo de puntuación, no confundir carácter con variable).

El separador decimal es el punto, y la coma se utiliza para separar elementos. La función básica es `c` ("concatenar"). Toda función aplicada a un objeto se identifica por que encierra entre paréntesis a dicho objeto. Cabe mencionar que R no distingue espacios, pero los incluiremos a lo largo del libro para una mejor visualización del código. A continuación, se crearán dos objetos (`x` e `y`), uno numérico y otro de caracteres.

```
> x <- c(1.2, 3.5, 7, 10, 250)
> y <- c("Argentina", "Brasil", "Chile")
```

A veces suele suceder que la línea de código está incompleta, por ejemplo cuando nos olvidamos de agregar un paréntesis (aunque en RStudio se agregan por defecto los paréntesis de apertura y cierre). En este caso, al presionar la tecla *Enter* aparecerá un signo + en la línea siguiente, lo que significa que el comando está incompleto. En este punto hay dos opciones: se puede agregar lo que falta y teclear nuevamente *Enter*, o sólo presionar la tecla *Esc*. Esta última opción deshace el comando anterior y permite que volvamos a comenzar desde cero.

Las teclas arriba y abajo (flechas) permiten ver los comandos ejecutados previa y posteriormente, lo cual es útil para ejecutar un comando similar o, simplemente, corregirlo si nos equivocamos, sin la necesidad de volver a escribirlo. La función más importante que vamos a utilizar en este libro es `data.frame()`, que define un marco de datos y se corresponde con la matriz básica de datos (MBD). Cada columna se identifica con un nombre (que corresponde a la variable) y distintas columnas pueden almacenar distintos tipos de datos.

Para el manejo de datos previo a su ingreso en RStudio, puede utilizarse cualquier software para planillas de datos. Sin embargo, hay que tener en cuenta algunas consideraciones: los nombres de las columnas no deben contener espacios (en tal caso utilizar guiones bajos o puntos como separadores) ni

pueden comenzar con números; los datos faltantes no deben dejarse en blanco, sino señalarse con NA (*not available*: dato no disponible). Una vez lista la planilla se recomienda guardarla en formato .txt (texto delimitado por tabulaciones) o .csv (texto delimitado por comas).

En RStudio es fácil importar un marco de datos, y se realiza como en la mayoría de los programas estadísticos. Para esto se debe cliquear *Import Dataset* en la ventana superior derecha y luego en *From Text (Base)...*. Aquí se busca el archivo y se revisa si la MBD es correcta. En este punto puede cambiarse el nombre de la MBD en el campo *Name*. Finalmente, se da clic a *Import*. Como ejemplo, importaremos la MBD de especies del género *Bulnesia* y 43 caracteres morfológicos (Bulnesia.txt). Si las columnas tienen nombres, debe tildarse *Yes* en el campo *Heading.*

Varias funciones nos permiten extraer información de una MBD, como por ejemplo el número de filas –función `nrow()`–, el número de columnas –función `ncol()`– o una descripción general de la MBD –función `str()`–.

```
> nrow(Bulnesia)
[1] 8
> ncol(Bulnesia)
[1] 44
> str(Bulnesia)
'data.frame':       8 obs. of 44 variables:
 $ species: Factor w/ 8 levels "B_arborea","B_bonariensis",..: 1 3 4 2 6 5 8 7
 $ C1 : int 2 2 0 0 1 0 0 2
 $ C2 : num 34.9 35.7 23.5 20.4 40.4 18.8 10.4 21.6
 $ C3 : num 2.1 1.6 2.6 1.3 2 1.3 1.85 1.4
 $ C4 : num 84.9 96.6 13.5 25.9 12.5 28.4 19.7 21.4
 $ C5 : num 56.6 70.8 8.9 17.8 11.3 25 11.6 27.1
 $ C6 : num 7.7 9 1.8 3.4 3.1 5.8 2.7 5.1
 $ C7 : int 13 7 8 14 5 4 10 2
 $ C8 : int 2 2 2 1 1 1 0 2
 $ C9 : int 0 0 1 0 1 1 0 2
 $ C10 : int 2 2 0 2 2 2 2 1
 $ C11 : num 29.6 39.7 5.2 8.9 6.6 13.6 5.7 16.8
 $ C12 : num 8.6 16.4 2.4 2 2.6 7.8 1.9 12
 $ C13 : int 6 6 NA 1 2 3 1 5
 $ C14 : int 0 1 2 2 1 2 2 2
 $ C15 : int 1 1 1 1 2 1 1 1
 $ C16 : int 2 2 1 1 1 1 1 1
 $ C17 : num 17.2 18.1 9 11.5 10 13.2 10.4 3.9
 $ C18 : num 7.1 6.4 7.1 6.8 7.4 5.3 5.9 2.9
 $ C19 : num 3.4 5.8 4.2 3.9 4.4 3 3.1 2.3
 $ C20 : int 2 2 2 2 2 2 2 1
 $ C21 : num 22.4 24.3 9.8 17.8 7.7 8.5 9.3 11.5
 $ C22 : num 17.3 18.5 5.7 10.2 4.6 2.7 4.3 7
 $ C23 : int 16 12 8 10 7 6 5 6
 $ C24 : int 0 0 0 0 1 2 1 2
 $ C25 : int 1 2 0 1 0 0 0 0
 $ C26 : int 2 2 0 1 0 0 0 0
 $ C27 : int 0 0 2 0 1 0 0 0
 $ C28 : int 1 1 1 1 1 1 2 1
 $ C29 : int 2 2 2 1 1 0 0 0
 $ C30 : num 10.2 9.9 7 10.6 7.1 6.2 7.4 4.1
 $ C31 : num 1.5 1.4 2 1.6 2.2 1.6 1.7 1.1
```

```
 $ C32 : num 5.2 53 4.4 4.4 3.1 3.9 4.4 2.9
 $ C33 : int 2 2 1 0 1 1 2 2
 $ C34 : int 2 2 1 2 2 2 2 0
 $ C35 : int 1 2 NA 2 1 1 1 0
 $ C36 : int 2 2 7 1 8 4 4 2
 $ C37 : int 1 1 1 1 1 2 2 1
 $ C38 : num 45.7 56.2 13.4 35.9 22.5 16.3 11.8 51.8
 $ C39 : num 40.8 51.8 12.3 32.6 18.8 13.3 12.9 47.7
 $ C40 : int 2 2 2 2 1 1 1 2
 $ C41 : num 7.7 5.3 0.6 4.8 0.8 0.7 0.4 5.2
 $ C42 : int 1 1 2 1 2 2 2 1
 $ C43 : num 13.4 12.1 2.7 10.8 11 4.9 5.3 13.5
```

En este ejemplo, la función `str()` da como resultado la clase de objeto (`data.frame`), el número de filas (`obs`) y el número de columnas (`variables`). El signo `$` describe cada variable en la MBD. Por ejemplo, la variable `species` describe un factor (variable categórica) con ocho estados (cada una de las especies). En este caso se incluyó el nombre de las UE como primera columna para utilizarlas como etiquetas. De esta forma el programa trata a esta columna como una variable, aunque no lo sea para el análisis. El resto de las variables (C1 a C43) representan variables numéricas o enteras, para las cuales se muestran los primeros valores.

Si se desea visualizar una variable en particular de la MBD o realizar operaciones sobre alguna de ellas, una opción es escribir primero el nombre de la MBD, seguida del signo `$`, seguida del nombre de la variable.

```
> Bulnesia$C4
[1] 84.9 96.6 13.5 25.9 12.5 28.4 19.7 21.4
```

En el caso de obtener o visualizar una variable a partir de un marco de datos, puede utilizarse la notación matricial, es decir, números para identificar las filas y las columnas (variables en este caso). Esto permite extraer no sólo variables, sino también filas o incluso datos específicos (en la jerga denominado indexación o subseteo). Para extraer datos específicos del marco de datos debemos abrir corchetes luego del nombre de la MBD, y escribir un número de filas y columnas separados por una coma. Si queremos por ejemplo extraer el segundo dato (especie *B. carrapo*) de la variable C4:

```
> Bulnesia[2, 4]
[1] 1.6
```

En caso de querer visualizar la variable C4 completa:

```
> Bulnesia[, 4]
[1] 84.9 96.6 13.5 25.9 12.5 28.4 19.7 21.4
```

Observe que entre el primer corchete y la coma no se puso ningún valor. Esto significa que queremos obtener todas las filas, de lo contrario habría que escribir todos los valores de las filas. Si se quiere un grupo de columnas, puede escribirse un vector de números o los nombres de las variables.

```
> Bulnesia[, c(1, 2, 3)]
            species C1   C2
1         B_arborea  2 34.9
2         B_carrapo  2 35.7
3       B_chilensis  0 23.5
4     B_bonariensis  0 20.4
```

```
5          B_retama  1 40.4
6         B_foliosa  0 18.8
7 B_schickendantzii  0 10.4
8      B_sarmientoi  2 21.6

> Bulnesia[,c("species", "C1", "C2")]
            species C1   C2
1         B_arborea  2 34.9
2         B_carrapo  2 35.7
3       B_chilensis  0 23.5
4     B_bonariensis  0 20.4
5          B_retama  1 40.4
6         B_foliosa  0 18.8
7 B_schickendantzii  0 10.4
8      B_sarmientoi  2 21.6
```

También puede interesarnos extraer aquellos valores de una variable con una cierta condición, por ejemplo, mayores a un cierto valor. Esto se realiza mediante los símbolos lógicos > (mayor que), < (menor que), >= (mayor o igual que), <= (menor o igual que), != (distinto de) ó == (igual que). Por ejemplo, si queremos obtener aquellos valores de C1 mayores a 1, debemos indexar el vector de interés (`Bulnesia$C1` en este caso) abriendo corchetes, y luego escribir la condición `Bulnesia$C1 > 1`:

```
> Bulnesia$C1[Bulnesia$C1 > 1]
[1] 2 2 2
```

Cuando indexamos un vector no hay valores de filas y columnas, porque por definición un vector representa una única columna de un marco de datos. Dicho de otra forma, para identificar un elemento de un vector se requiere sólo un número (la posición). Por ejemplo, el dato en la posición 8 de la variable C1 da como resultado 2.

```
> Bulnesia$C1[8]
[1] 2
```

Ahora que sabemos cómo indexar un marco de datos, podemos aplicar cualquier función a una variable, a una fila o a una columna, como por ejemplo, la media –función `mean()`– o el desvío estándar –función `sd()`–.

```
> mean(Bulnesia$C4)
[1] 37.8625
> sd(Bulnesia[, 4])
[1] 0.4652188
```

En este caso, se calcularon la media y el desvío estándar de la misma variable (columna 4) pero utilizando diferente notación.

Actualmente, se está utilizando con mayor frecuencia una estructura de datos muy similar al marco de datos (rectangular, organizada en filas y columnas) denominada *tibble*, perteneciente al tidyverse (Wickam *et al*. 2019). El tidyverse es un conjunto de paquetes que comparten representaciones de datos comunes. Estructuralmente, un *tibble* es también un marco de datos, aunque modifica algunas características antiguas de estos últimos. Por ejemplo, un *tibble* aprovecha mejor el espacio en la pantalla al mostrar información más relevante (solo muestra las primeras 10 filas y aquellas columnas que entran en el ancho de la pantalla). Además, no convierte los caracteres en factores, nunca cambia el nombre de las variables y nunca asigna nombres a las filas. La creación y manipulación de *tibbles* puede hacerse

utilizando el paquete tibble (Müller & Wickham 2019). A lo largo de este libro, trabajaremos con marcos de datos en lugar de *tibbles*, debido a que la mayoría de los usuarios suelen estar más familiarizados con los marcos de datos. Además, los objetos *tibble* son compatibles con estos últimos.

PAQUETES

Como se mencionó anteriormente, los paquetes son conjuntos de funciones, datos y códigos compilados, orientados a un análisis o grupos de análisis en particular. Por ejemplo, el paquete ape (Paradis y Schliep 2018) contiene un conjunto de funciones para realizar distintos análisis filogenéticos. El directorio donde se guarda un paquete se denomina librería. R trae incorporado un conjunto de paquetes estándar que permite realizar operaciones básicas y diversos análisis estadísticos. Una gran cantidad de paquetes (> 10.000) está disponible para su descarga e instalación de acuerdo al tipo de análisis que nos interese realizar. Para instalar un paquete se debe ingresar a la pestaña *Packages* (ventana inferior derecha) y cliquear en *Install*. Aparecerá una ventana con un recuadro donde dice *Packages*, y en el cual se deberá buscar el nombre del paquete. Una vez encontrado se debe cliquear *Install*.

Una vez instalado un paquete, éste debe cargarse –funciones `library()` o `require()`– cada vez que se inicia sesión para poder utilizarlo. Por ejemplo, si queremos cargar el paquete ape:

```
> library(ape)
```

A lo largo del libro, se irán cargando los paquetes que sean necesarios (recuerde que el usuario deberá instalarlos previamente).

DATOS FALTANTES

Cuando un dato no está disponible en la MBD, R utiliza el símbolo NA (*not available*). En muchos casos al aplicar una función el resultado es NA, esto ocurre cuando existe al menos un valor no disponible. Por lo tanto, hay que indicarle al programa que no considere los NA, lo cual varía dependiendo de la función. Por ejemplo, para excluir los datos faltantes en el cálculo de la media y la correlación, deben especificarse los argumentos `na.rm` (eliminar datos faltantes) y `use`, respectivamente:

```
> mean(Bulnesia$C35)
[1] NA
> mean(Bulnesia$C35, na.rm = TRUE)
[1] 1.142857
> cor(Bulnesia$C35, Bulnesia$C36)
[1] NA
> cor(Bulnesia$C35, Bulnesia$C36, use = "complete.obs")
[1] -0.2338821
```

Los datos de tipo lógico sólo tienen dos valores posibles: verdadero (`TRUE`) y falso (`FALSE`), y representan si una condición se cumple (verdadero) o no (falso). En el caso de la media se indica que se deben eliminar los datos faltantes (`na.rm = TRUE`). Otra alternativa para excluir valores no disponibles es eliminar previamente todas las filas con datos faltantes usando la función `na.omit()`. Si aplicamos esta función a la base de datos de *Bulnesia*, se eliminará la fila 3, ya que contiene datos no disponibles. Sin embargo, esto trae aparejado la eliminación de datos que sí pueden ser importantes para el análisis.

AYUDAS

Las ayudas y explicaciones de una función suelen ser poco amigables en R, y suele ser más útil realizar búsquedas en foros especializados y páginas con tutoriales. A la ayuda se puede acceder des-

de la pestaña *Help* (ventana inferior derecha) o ejecutando en la consola el signo de pregunta, seguido del nombre de la función. A modo de ejemplo buscaremos ayuda sobre la función `cor()`.

```
> ?cor
```

Para cada función aparecerá una descripción (*Description*), su uso (*Usage*), una serie de argumentos (*Arguments*), el resultado que arroja (*Value*), referencias (*References*) y ejemplos (*Examples*). Los argumentos constituyen una parte esencial de toda función y, básicamente, representan opciones que el usuario puede especificar. A modo de ejemplo, la función `cor()` calcula el coeficiente de correlación para un par de variables. Si no especificamos ningún argumento, la función utiliza los valores por defecto (para saber cuáles son debemos revisar la ayuda). En este caso, vamos a calcular el coeficiente de correlación entre las variables C1 (argumento `x`) y C2 (argumento `y`) de la MBD `Bulnesia`.

```
> cor(x = Bulnesia$C1, y = Bulnesia$C2)
[1] 0.631622
```

Los argumentos `x` e `y` no pueden estar ausentes ya que no se podría calcular correlación alguna. El nombre de los argumentos puede obviarse si conocemos el orden (es decir que si no los escribimos R asume que la primera variable es `x` y la segunda `y`), aunque suele ser más útil escribirlos, ya que difícilmente recordemos el orden de todos los argumentos de una función.

```
> cor(Bulnesia$C1, Bulnesia$C2)
[1] 0.631622
```

Examinando la ayuda vemos que se encuentra el argumento `method`, el cual permite calcular tres tipos de correlación (`"pearson"`, `"kendall"` y `"spearman"`). Note que el método por defecto es `"pearson"`. Si se desea calcular el coeficiente de correlación de Spearman se debe especificar en la función.

```
> cor(x = Bulnesia$C1, y = Bulnesia$C2, method = "spearman")
[1] 0.6390097
```

ERRORES Y ADVERTENCIAS

Al iniciarse en el uso del programa R es importante no desalentarse, ya que es fácil cometer errores en la ejecución de comandos. En estos casos R arroja un mensaje de error cuando no puede ejecutar la función (y por lo tanto no hay resultado) o cuando queremos llamar un objeto o función que no existe. Por ejemplo, si queremos llamar a la base de datos `Bulnesia` pero la escribimos en minúscula:

```
> bulnesia
Error: object 'bulnesia' not found
```

Algunos de los errores más comunes son:

1. `object 'XXX' not found`. Literalmente el objeto no existe (esto se puede verificar en la ventana superior derecha) y suele ocurrir cuando se intenta llamar de forma incorrecta el nombre del objeto.
2. `could not find function "XXX"`. Es posible que el paquete correspondiente no haya sido cargado o se haya escrito incorrectamente el nombre de la función.
3. `there is no package called 'XXX'`. El nombre del paquete está escrito de forma incorrecta o no ha sido instalado.
4. `unexpected 'X' in`… Falta algún símbolo, como una coma o un paréntesis.

5. `object cannot be coerced to type 'double'.` El objeto no es de la clase numérica. Esto puede verificarse mediante la función `class()`. Es posible que esté intentando aplicar una función a un objeto de tipo carácter o factor.

A diferencia de los errores, los mensajes de advertencia (*warnings*) dan una cierta información al usuario, pero sin detener la ejecución de la función. Dicho de otra forma, la función arroja un resultado, pero puede haber un problema con la información que se ingresó en la función. Por ejemplo, la función `cor()` arroja una advertencia si uno de los vectores tiene un desvío estándar igual a 0, pero devuelve como resultado NA.

```
> x <- c(1, 1, 1, 1, 1)
> y <- c(1, 2, 2.1, 1.9, 3)
> cor(x, y)
[1] NA
Warning message:
In cor(x, y) : the standard deviation is zero
```

EXPORTACIÓN DE ARCHIVOS

Los datos generados en R pueden guardarse en un archivo de varias formas. En general se recomienda exportar cada archivo en formato de texto (como .txt o .csv), ya que así son manejados por todos los programas que trabajan con hojas de cálculo. La función con la cual se pueden exportar más tipos de archivos es `write.table()`, que permite guardar datos mediante la especificación del símbolo decimal a utilizar (argumento `dec`) y el símbolo que se emplea para separar los valores (argumento `sep`), entre otros.

Supongamos que queremos exportar la matriz de correlación entre las variables 2, 3 y 4 de la MBD de Bulnesia, a la que denominamos `cor.bulnesia`.

```
> cor.bulnesia <- cor(Bulnesia[, 2:4])
> cor.bulnesia
           C1        C2         C3
C1 1.00000000 0.6316220 0.03873177
C2 0.63162197 1.0000000 0.42868695
C3 0.03873177 0.4286869 1.00000000
```

Para exportar un archivo también es necesario especificar su ubicación o ruta, explícita o implícitamente, y darle un nombre junto con su extensión (.txt, .csv). Si no se especifica ninguna ubicación el archivo se guardará en el directorio de trabajo actual (*working directory*), al cual se puede acceder mediante la función `getwd()` y modificarse con la función `setwd()` (en este caso, constatar que existe la ubicación especificada). También es posible subir archivos a la nube, como Google Drive y Dropbox.

```
> getwd()
[1] "C:/"
> setwd("C:/R datos")
> getwd()
[1] "C:/R datos"
```

De este modo podemos exportar el objeto generado de dos formas, especificando o no la ruta completa. En primer lugar exportaremos la matriz como formato .csv.

```
> write.table(x = cor.bulnesia, file = "Matriz de correlación Bulnesia.csv",
+             sep = ",", dec = ".")
```

```
> write.table(x = cor.bulnesia, file = "C:/Matriz de correlación
+             Bulnesia.csv", sep = ",", dec = ".")
```

Con el formato .csv se debe tener en cuenta la siguiente consideración. En el lenguaje anglosajón los campos se separan por comas y el símbolo decimal es el punto. En el lenguaje latino los campos están separados por punto y coma y el símbolo decimal es la coma. Por lo tanto, el separador y el símbolo decimal que utilicemos dependerá del idioma en que tengamos configurado R y nuestro sistema operativo. Otras funciones específicas para formato .csv son `write.csv()` y `write.csv2()`.

También podemos exportar el archivo como formato .txt, en el cual los campos están separados por tabulaciones.

```
> write.table(x = cor.bulnesia, file = "C:/R datos/ Matriz de correlación
+             Bulnesia.txt", sep = "\t", dec = ".")
```

Existen otros múltiples formatos que pueden utilizarse para exportar archivos (archivos excel, spss, sas, stata), pero no serán tratados en este libro.

CERRANDO SESIÓN

Una vez que terminamos de utilizar RStudio tenemos la opción de guardar el espacio de trabajo actual, de forma que todos los objetos creados estén disponibles en la próxima sesión. Sin embargo no es recomendable, ya que en poco tiempo puede consumir una gran cantidad de espacio en el disco. Es preferible guardar las rutinas y ejecutarlas cada vez que las necesitemos. De esta forma ocuparemos mucho menos espacio. Para guardar una rutina simplemente se exportan, con la pestaña de la rutina activa, como *File > Save o Save as…*. Este archivo es un archivo formato R.

Finalmente, hay que tener en cuenta que ciertos tipos de análisis pueden arrojar resultados diferentes cada vez que se ejecutan (por ejemplo: escalado multidimensional no métrico, pruebas basadas en permutaciones, *bootstrapping*), por lo que en estos casos es conveniente copiar y pegar los resultados en un archivo aparte.

CAPÍTULO 4
ESTIMACIÓN DEL PARECIDO ENTRE UNIDADES DE ESTUDIO: SIMILITUD

¿Puede expresarse en forma cuantitativa el parecido entre dos unidades de estudio? El parecido o similitud es cuantificable aplicando coeficientes de similitud. Con el uso de estos coeficientes en operaciones matemáticas, pueden calcularse las similitudes respecto a cada par posible de unidades de estudio (UE) de una matriz básica de datos (MBD).

A lo largo de la historia del análisis multivariado se han formulado numerosos coeficientes de similitud (Orlóci 1975, Hubálek 1982, Shi 1993, Legendre y Legendre 1998, Choi *et al*. 2010, Todeschini *et al*. 2012, Schmera y Podani 2018). En este capítulo nos referiremos únicamente a aquellos coeficientes de similitud más utilizados y seguiremos la clasificación de Sneath y Sokal (1973), dividiendo a los mismos en tres grandes grupos: coeficientes de distancia, de asociación y de correlación.

COEFICIENTES DE DISTANCIA

Una representación geométrica de las medidas de distancia podría ser la siguiente: dadas las UEA, B, C, D, y tres variables que las describen X_1, X_2, X_3, se puede construir un espacio tridimensional mediante tres ejes de coordenadas que representan cada variable (Fig. 4.1). Cada UE puede ubicarse en el espacio formado por esos tres ejes, de acuerdo con el valor de cada una de las variables.

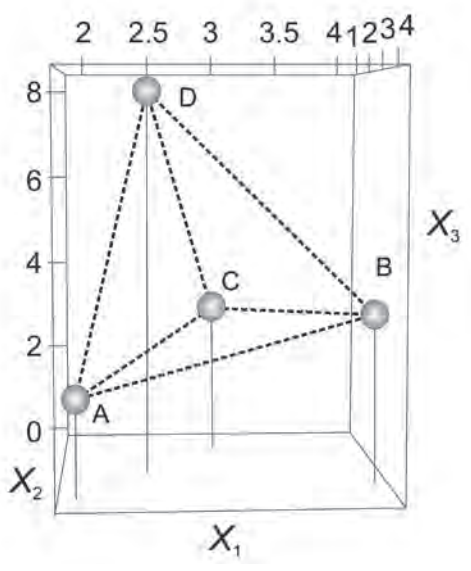


Fig. 4.1. Espacio multivariado de tres dimensiones en el cual se ubican cuatro UE (A a D). Las líneas continuas muestran la posición de las UE en el plano formado por las variables X_1 y X_2, mientras que las líneas discontinuas representan la distancia (euclideana) entre las UE.

Estas medidas se pueden aplicar tanto a datos cuantitativos (continuos y discretos) como cualitativos (presencia-ausencia). A continuación, se presentarán distintos coeficientes junto con el cálculo realizado sobre la base de la siguiente matriz de sitios × abundancia de especies, MBD1 (Tabla 4.1). Esta MBD puede expresarse en términos matemáticos como se muestra en la Tabla 4.2. Si bien esta matriz es la más simple (ya que incluye sólo dos UE), el análisis se extiende a matrices de cualquier dimensión, dado que el cálculo siempre es de a pares de UE.

Tabla 4.1. MBD1 de dos sitios (A, B) × tres especies (sp1 a sp3).

Sitio	sp1	sp2	sp3
A	12	1	0
B	14	5	11

Tabla 4.2. Nomenclatura utilizada para representar la MBD1.

UE	$j = 1$	$j = 2$	$j = 3$
i = A	X_{A1}	X_{A2}	X_{A3}
i = B	X_{B1}	X_{B2}	X_{B3}

Distancia euclideana

La distancia euclideana entre dos UE es una posible cuantificación de sus diferencias, que surge de la fórmula de Pitágoras (Gower 1982). Por ejemplo, si queremos calcular la distancia euclideana en tres dimensiones entre las UEA y B:

$$D = \sqrt{\sum \left(X_{Aj} - X_{Bj} \right)^2}$$

$$D = \sqrt{\left(12 - 14\right)^2 + \left(1 - 5\right)^2 + \left(0 - 11\right)^2}$$

$$D = 11{,}87$$

La distancia euclideana (Fig. 4.1) ha sido utilizada como medida de distancia taxonómica por Sokal (1961) para comparar individuos, especies o cualquier otro taxón sobre la base de caracteres biológicos. Con más de tres variables estaríamos ante un espacio *n*-dimensional. Pero no por esta razón dejaría de tener validez el cálculo de las distancias en ese espacio, ya que a pesar de que es imposible representar gráficamente más de tres ejes, la geometría del espacio tridimensional es aplicable a espacios euclideanos de más de tres dimensiones.

Si bien la distancia euclideana puede aplicarse a matrices de presencia-ausencia, no es recomendable ya que presenta el problema de los "dobles ceros" (Legendre y Legendre 1998, Gagné y Proulx 2009). En la MBD de la Tabla 4.3 la distancia entre A y B es igual a dos, con una especie en común (especie 3), mientras que la distancia entre C y D también es igual a dos, a pesar de no tener especies en común. Esto se debe a su desarrollo matemático, ya que la diferencia entre especies compartidas (1 – 1) es igual a la diferencia entre ausencias compartidas (0 – 0) para los sitios en cuestión, como se muestra en el siguiente cálculo:

$$D_{AB} = \sqrt{(1-0)^2 + (0-1)^2 + (1-1)^2 + (1-0)^2 + (0-1)^2}$$
$$D_{AB} = \sqrt{4} = 2$$

$$D_{CD} = \sqrt{(1-0)^2 + (0-1)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2}$$
$$D_{CD} = \sqrt{4} = 2$$

Tabla 4.3 MBD de cuatro sitios (A a D) × cinco especies (sp1 a sp5).

Sitio	sp1	sp2	sp3	sp4	sp5
A	1	0	1	1	0
B	0	1	1	0	1
C	1	0	0	1	0
D	0	1	0	0	1

Distancia taxonómica

Debido a que el número de variables *p* influye en la estimación de la distancia, Sokal (1961) propuso la distancia taxonómica, que representa el promedio de la distancia euclideana.

$$DT = \frac{D}{\sqrt{p}}$$
$$DT = \frac{11{,}87}{\sqrt{3}}$$
$$DT = 6{,}85$$

Distancia de Manhattan

También conocida como distancia L_1 o del taxista. Esta última denominación refiere al hecho de que la distancia entre dos unidades para dos variables es la distancia en la abscisa (eje *X*) más la distancia en la ordenada (eje *Y*), al igual que la distancia que recorre un taxi entre las cuadras de una ciudad (Fig. 4.2):

$$M = \sum \left| X_{Aj} - X_{Bj} \right|$$
$$M = |12-14| + |1-5| + |0-11|$$
$$M = 17$$

Las barras verticales representan el valor absoluto. Esta distancia funciona mejor que la distancia euclideana para matrices con muchas variables (Aggarwal *et al*. 2001).

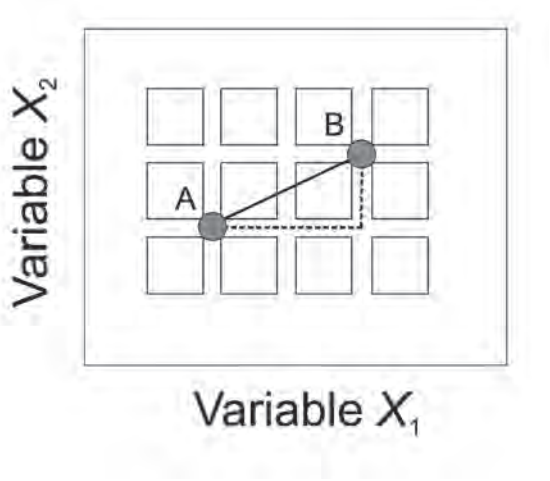


Fig. 4.2. Distancia de Manhattan (línea punteada) entre dos UE (círculos grises). La línea continua representa la distancia euclideana.

Diferencia de carácter promedio (*mean character difference*)

La diferencia de carácter promedio fue descrita por el antropólogo Czekanowski (1909) y propuesta como medida taxonómica por Cain y Harrison (1958). Corresponde a la distancia de Manhattan dividida por el número de variables *p*, lo que permite que no aumente la distancia con el número de variables utilizadas.

$$MCD = \frac{M}{p}$$

$$MCD = \frac{17}{3} = 5{,}67$$

Distancia de Canberra

Lance y Williams (1967) propusieron la distancia de Canberra como una modificación de la distancia de Manhattan:

$$C = \sum \frac{\left|X_{Aj} - X_{Bj}\right|}{X_{Aj} + X_{Bj}}$$

$$C = \frac{\left|12-14\right|}{12+14} + \frac{\left|1-5\right|}{1+5} + \frac{\left|0-11\right|}{0+11}$$

$$C = 1{,}74$$

Para el cálculo de esta distancia deben excluirse los dobles ceros. Para el caso de una matriz de sitios × abundancia de especies, una diferencia dada entre especies abundantes contribuye menos a la distancia que la misma diferencia entre especies raras.

Distancia de Cao

El coeficiente de Cao fue propuesto por Cao *et al.* (1997) como medida de distancia con sesgo mínimo, para sitios o comunidades con baja similitud e intensidades de muestreo variables. No tiene límite superior y puede variar entre sitios con especies no compartidas. Tampoco está definido para cero por utilizar logaritmos, de modo que este se reemplazan por el valor arbitrario de 0,1 (Cao *et al.* 1997):

$$CYd = \frac{1}{p}\sum\left[\frac{\left(X_{Aj}+X_{Bj}\right)\log\left(\frac{X_{Aj}+X_{Bj}}{2}\right)-X_{Aj}\log\left(X_{Bj}\right)-X_{Bj}\log\left(X_{Aj}\right)}{X_{Aj}+X_{Bj}}\right]$$

Donde *p* es el número de especies presente en ambas muestras y log es el logaritmo natural. En nuestro ejemplo de la MBD1 el cálculo sería el siguiente:

$$CYd = \frac{1}{3}\left(\begin{array}{l}\frac{(12+14)\log\left(\frac{12+14}{2}\right)-12\log 14-14\log 12}{12+14}+\\ \frac{(1+5)\log\left(\frac{1+5}{2}\right)-1\log 5-5\log 1}{1+5}+\\ \frac{(0{,}1+11)\log\left(\frac{0{,}1+11}{2}\right)-0{,}1\log 11-11\log 0{,}1}{0{,}1+11}\end{array}\right)$$

$$CYd = \frac{1}{3}\left(\frac{66{,}69-31{,}67-34{,}79}{26}+\frac{6{,}59-1{,}61-0{,}00}{6}+\frac{19{,}02-0{,}24+25{,}33}{11{,}1}\right)$$

$$CYd = \frac{1}{3}\left(0{,}01+0{,}83+3{,}97\right)$$

$$CYd = 1{,}60$$

Distancia chi-cuadrado

Pearson (1900) introdujo la distancia chi-cuadrado en la cual se utilizan datos discretos relativos (divididos por la suma total de cada UE). Representa una distancia euclideana ponderada, donde los pesos w_j están dados por la inversa de la suma de cada columna relativa al total de observaciones (ver Tabla 4.5). Esta medida es la base del análisis de correspondencias (Cap. 6), y al utilizar frecuencias relativas tiene la ventaja de ser independiente del tamaño muestral. Es apropiada para datos de frecuencias, donde todas las variables (especies en este caso) se encuentran expresadas en las mismas unidades. Para calcular este coeficiente las frecuencias absolutas se dividen por el total de cada fila (Tabla 4.4), dando como resultado una nueva matriz con abundancias relativas (Tabla 4.5), donde cada valor es ahora p_{Aj} y p_{Bj}. Los pesos de la MBD1 son 1/0,60, 1/0,14 y 1/0,26 (Tabla 4.5). Este cálculo permite que las especies con mayor abundancia total no influyan tanto en la distancia.

Tabla 4.4. MBD1 de dos sitios (A, B) × tres especies (sp1 a sp3), junto con la suma de las abundancias totales por sitio y especie.

Sitio	sp1	sp2	sp3	Suma
A	12	1	0	13
B	14	5	11	30
Suma	26	6	11	43

Tabla 4.5. MBD1 en la cual la abundancia de cada especie se divide por la abundancia total del sitio (abundancia relativa p_{Aj} y p_{Bj}).

Sitio	*p*1	*p*2	*p*3	Suma
A	p_{A1} = 12/13 = 0,92	p_{A2} = 1/13 = 0,08	p_{A3} = 0/13 = 0,00	1
B	p_{B1} = 14/30 = 0,47	p_{B2} = 5/30 = 0,17	p_{B3} = 11/30 = 0,36	1
Suma	26/43 = 0,60	6/43 = 0,14	11/43 = 0,26	1

$$\chi^2 = \sqrt{\sum w_j \left(p_{Aj} - p_{Bj} \right)^2}$$

$$\chi^2 = \sqrt{\frac{1}{0,60}(0,92-0,47)^2 + \frac{1}{0,14}(0,08-0,17)^2 + \frac{1}{0,26}(0,00-0,36)^2}$$

$$\chi^2 = \sqrt{\frac{0,202}{0,60} + \frac{0,008}{0,14} + \frac{0,130}{0,26}}$$

$$\chi^2 = 0,96$$

La justificación del uso de frecuencias relativas en lugar de absolutas, se hace evidente en el caso extremo donde las abundancias de un sitio son exactamente el doble de las abundancias de otro sitio. Para cualquier coeficiente la distancia será proporcional a esta diferencia, mientras que para la distancia chi-cuadrado la distancia será igual a 0, ya que $p_{Aj} = p_{Bj}$.

A diferencia del resto de los coeficientes, el valor máximo de distancia para chi-cuadrado depende de cada matriz, ya que se calcula como:

$$\sqrt{N \times \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}$$

Donde N es igual al número total de observaciones, n_A es el número de observaciones de la UEA y n_B es el número de observaciones de la UEB.

Distancia de Mahalanobis

Introducida por Mahalanobis (1936), es una distancia euclideana ponderada que tiene en cuenta la covarianza entre variables, está íntimamente relacionada con la correlación y es independiente de sus escalas.

$$MHL = \sqrt{\sum w_{jk} \left(X_{Aj} - X_{Bj} \right) \left(X_{Ak} - X_{Bk} \right)}$$

La covarianza es una medida de relación entre dos variables que puede ser negativa, positiva o nula, y se calcula como:

$$S_{jk} = \frac{\sum \left(X_{ij} - \bar{X}_j \right) \left(X_{ik} - \bar{X}_k \right)}{N-1}$$

En la fórmula aparecen los subíndices j y k para representar dos variables cualquiera, ya que para calcular la covarianza se necesitan dos variables. N es el número de UE de la MBD y $\bar{X}$ es la media de cada variable. Cuando $j = k$ (misma variable) se obtiene la varianza (Box 2.1).

$$S_j^2 = \frac{\sum \left(X_{ij} - \bar{X}_j \right)^2}{N-1}$$

Como ejemplo se muestra la covarianza de las variables 1 y 2 (S_{12}), y la varianza para la variable 1 (S_1^2) de la MBD de la Figura 4.3A.

$$S_{12} = \frac{1}{4}\left[\begin{array}{l}(5,5-3,8)(8,1-6,9)+(2,1-3,8)(5,5-6,9)+(3-3,8)(4,2-6,9)+\\(4,2-3,8)(8,3-6,9)+(4,1-3,8)(8,5-6,9)\end{array}\right]$$

$$S_{12} = 1,91$$

$$S_1^2 = \frac{(5,5-3,8)^2+(2,1-3,8)^2+(3-3,8)^2+(4,2-3,8)^2+(4,1-3,8)^2}{4}$$

$$S_1^2 = 1,67$$

Una vez calculadas las varianzas y covarianzas, se construye una matriz de varianza-covarianza (Fig. 4.3B-C). Esta es una matriz de variables × variables, en la cual las celdas de la diagonal principal contienen las varianzas, y el resto de las celdas contienen las covarianzas entre cada par de variables. Por lo tanto, es una matriz cuadrada (igual número de filas que de columnas) y simétrica (se pueden intercambiar filas por columnas sin alterar la matriz). Luego se calcula una matriz inversa mediante software (Fig. 4.3D) cuyos valores en la fórmula corresponden a los pesos (w_{jk}). Los términos multiplicados por 2, corresponden a las covarianzas que aparecen dos veces en la matriz de varianza-covarianza (al ser simétricas, por ejemplo, $S_{12} = S_{21}$).

$$MHL = \sqrt{\begin{array}{l}1,48(5,5-2,1)^2+1,09(8,1-5,5)^2+0,01(200-225)^2+2(-0,51)(5,5-2,1)(8,1-5,5)+\\2(0,03)(5,5-2,1)(200-225)+2(0,08)(8,1-5,5)(200-225)\end{array}}$$

$$MHL = 2,82$$

Las variables muy asociadas entre sí repiten información, por lo que al ser tratadas de manera independiente sobreestimarían los valores de similitud entre las UE. De esta forma, las variables con alta asociación contribuyen menos a la similitud que aquellas poco asociadas, ya que no son completamente independientes.

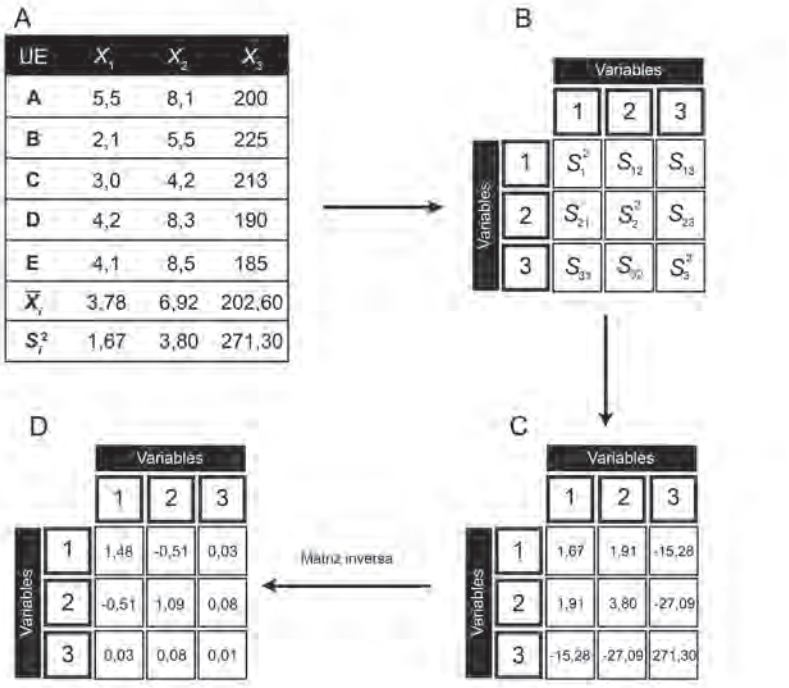


Fig. 4.3. Distancia de Mahalanobis. (A) MBD; (B) modelo de matriz de varianza-covarianza; (C) matriz de varianza-covarianza; (D) matriz inversa de la matriz de varianza-covarianza.

Distancias genéticas

Un caso particular de distancias son las distancias genéticas o evolutivas, que corresponden a la distancia entre pares de bases de secuencias de ADN o ARN. Estas secuencias están constituidas por bases nitrogenadas que se dividen en dos grupos: purinas (A: adenina, G: guanina) y pirimidinas (C: citosina, T: timina, U: uracilo). Las distancias genéticas se construyen a partir de una MBD de taxones × sitios de bases nitrogenadas, y sólo se utilizan en el análisis filogenético.

La diferencia con el resto de los coeficientes de similitud descritos en este capítulo, es que estas distancias suponen un modelo *a priori* de sustitución nucleotídica, que brinda una descripción estadística del fenómeno estudiado (mutaciones). Por lo tanto, cada coeficiente de distancia genética (existen al menos 20) asume una serie de supuestos para realizar el análisis. Este tema será abordado en un contexto filogenético en el Capítulo 8.

COEFICIENTES DE ASOCIACIÓN

Se han propuesto innumerables coeficientes de asociación (Legendre y Legendre 1998). En general, los coeficientes de asociación varían entre 0 y 1, donde 0 representa la mínima similitud y 1 la máxima. En el caso más simple, la similitud entre dos UE está basada en datos de presencia-ausencia. Éstos pueden describir la presencia-ausencia de determinadas condiciones ambientales, especies o caracteres. Las observaciones pueden resumirse en una tabla de frecuencias de 2×2, en la cual pueden darse cuatro combinaciones posibles (Fig. 4.4A):

1. Presencia de un taxón, una variable o una condición ambiental en ambas UE, codificada como (1, 1) = *a*.
2. Presencia de un taxón, una variable o una condición ambiental en la primera UE y ausencia en la segunda, codificada como (1, 0) = *b*.
3. Ausencia de un taxón, una variable o una condición ambiental en la primera UE y presencia en la segunda, codificada como (0, 1) = *c*.
4. Ausencia de un taxón, una variable o una condición ambiental en ambas UE, codificada como (0, 0) = *d*.

Dicho de otra forma, *a* y *d* representan las concordancias, mientras que *b* y *c*, las discordancias. En términos de teoría de conjuntos, *a* representa la intersección entre ambas UE, *b* y *c* representan las diferencias de cada UE, y *d* representa el complemento de la unión de A y B (Fig. 4.4B).

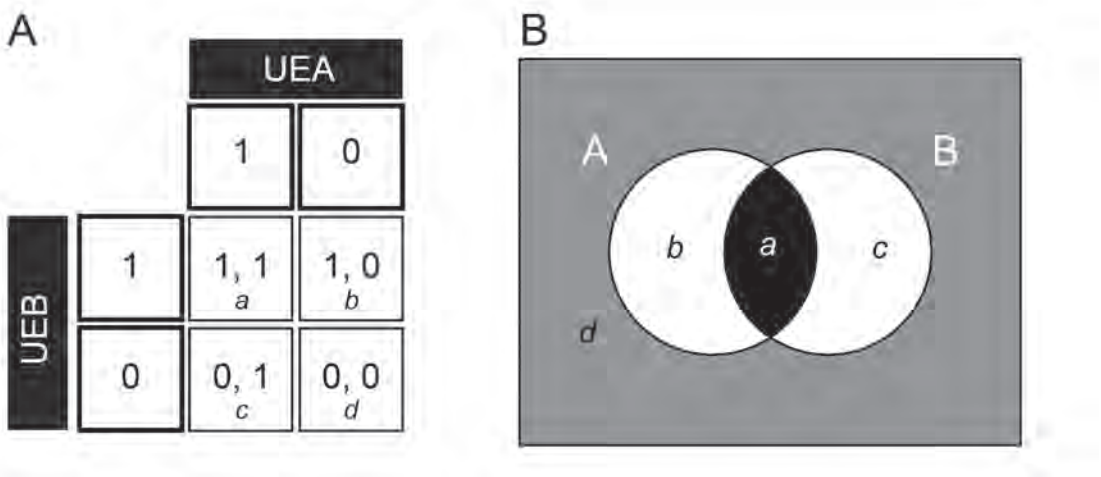


Fig. 4.4. Las cuatro combinaciones posibles de presencia-ausencia entre dos UE (*a* = presencias compartidas, *b* y *c* = presencia en una sola de las UE, y *d* = ausencias compartidas). (A) Combinaciones en forma de MBD; (B) combinaciones en términos de teoría de conjuntos.

Presentaremos a continuación distintos coeficientes de asociación junto con el cálculo realizado sobre la siguiente MBD2 hipotética, constituida por las UEA y B, y 10 variables presencia-ausencia (Tabla 4.6):

Tabla 4.6. MBD2 de dos sitios (A, B) × 10 especies (sp1 a sp10).

Sitio	sp1	sp2	sp3	sp4	sp5	sp6	sp7	sp8	sp9	sp10
A	0	1	1	0	1	0	1	0	1	0
B	1	1	0	0	1	1	0	0	1	1

En este caso, a = 3 (sp2, sp5 y sp9), b = 2 (sp3 y sp7), c = 3 (sp1, sp6, y sp10) y d = 2 (sp4 y sp8). El resultado de la suma de a, b, c y d es el número total de variables utilizadas (p = 10).

COEFICIENTES DE ASOCIACIÓN QUE UTILIZAN DATOS BINARIOS

Simple matching

Este coeficiente propuesto por Sokal y Michener (1958) es el más simple, y asume que no hay diferencias entre presencias y ausencias compartidas (a y d):

$$SM = \frac{a+d}{a+b+c+d}$$
$$SM = \frac{3+2}{3+2+3+2}$$
$$SM = 0,50$$

Rogers y Tanimoto

Una variante del *simple matching* es el coeficiente de Rogers y Tanimoto (1960), el cual da más peso a las diferencias (b y c):

$$RT = \frac{a+d}{a+2b+2c+d}$$

Al aplicarse este coeficiente a las UEA y B de la MBD2 resulta:

$$RT = \frac{3+2}{3+4+6+2}$$
$$RT = 0,33$$

Russell y Rao

Russell y Rao (1940) propusieron una medida que considera las dobles presencias (a) sobre el total de las variables utilizadas:

$$RR = \frac{a}{a+b+c+d}$$
$$RR = \frac{3}{3+2+3+2}$$
$$RR = 0,30$$

Por ejemplo, para una MBD de dos UE con a = 3 y d = 3, el coeficiente da un valor de RR = 3/6, mientras que si a = 3 y d = 10 el coeficiente da un valor de RR = 3/13. Observe que las dobles ausencias (d), en este caso, son un elemento en contra de la similitud.

Kulczynski

Kulczynski (1928) propuso un coeficiente que se calcula como el cociente entre las dobles presencias (*a*) y las discordancias (*b* y *c*):

$$K = \frac{a}{b+c}$$
$$K = \frac{3}{2+3}$$
$$K = 0{,}60$$

Este coeficiente tiene un límite inferior de 0 (no hay similitud) y no tiene límite superior. Arbitrariamente se asigna un valor máximo de 9999,999. Esto se debe a que si dos UE son completamente iguales, entonces $b = 0$ y $c = 0$, dando como resultado un cociente indeterminado.

En caso de matrices de especies × sitios, este coeficiente es apropiado cuando existe mucha diferencia entre las riquezas de especies entre dos áreas (Moline y Linder 2006).

Sokal y Sneath

El coeficiente de Sokal y Sneath (1963) da más peso a las concordancias que a las discordancias (*a* y *d*):

$$SS = \frac{2a+2d}{2a+b+c+2d}$$
$$SS = \frac{6+4}{6+2+3+4}$$
$$SS = 0{,}66$$

Hamann

El coeficiente de Hamann (1961) penaliza las diferencias:

$$H = \frac{(a+d)-(b+c)}{a+b+c+d}$$
$$H = \frac{(3+2)-(2+3)}{3+2+3+2}$$
$$H = 0{,}00$$

Este coeficiente, a diferencia de la mayoría, varía entre –1 y +1, y equivale a la mínima y máxima similitud, respectivamente.

Jaccard

Algunos coeficientes no consideran a las ausencias compartidas (*d*) como elemento en favor de la similitud, ya que podría haber tantas ausencias conjuntas como se quisiera. Jaccard (1900) propuso el siguiente coeficiente:

$$J = \frac{a}{a+b+c}$$
$$J = \frac{3}{3+2+3}$$
$$J = 0{,}37$$

Dice-Sørensen

Una variante del coeficiente de Jaccard es el coeficiente de Sørensen (1948), basado en Dice (1945). Este coeficiente confiere mayor peso a las dobles presencias (*a*) y excluye las dobles ausencias (*d*), ya que es posible considerar que las dobles presencias son más informativas que las discordancias (*b* y *c*).

$$DS = \frac{2a}{2a+b+c}$$
$$DS = \frac{6}{6+2+3}$$
$$DS = 0,54$$

Simpson

Para el coeficiente de Simpson (1943) dos UE pueden tener máxima similitud, aunque no posean exactamente las mismas presencias.

$$S = \frac{a}{a+\min(b,c)}$$

Por ejemplo, si un sitio está anidado dentro de otro, ambos tendrán similitud igual a 1. A modo de ejemplo considere la siguiente MBD3 (Tabla 4.7):

Tabla 4.7. MBD3 de tres sitios (A a C) × seis especies (sp1 a sp6)

Sitio	sp1	sp2	sp3	sp4	sp5	sp6
A	0	1	0	0	0	0
B	1	1	1	1	1	0
C	0	1	1	1	0	1

Para la similitud entre A y B: $a = 1$, $b = 4$ y $c = 0$. Como el mínimo de b y c es 0, entonces:

$$S = \frac{1}{1+0} = 1$$

La similitud es máxima a pesar de que los sitios no son exactamente iguales. Para la similitud entre B y C: $a = 3$, $b = 2$ y $c = 1$. Por lo tanto:

$$S = \frac{3}{3+1} = 0,75$$

COEFICIENTES DE ASOCIACIÓN QUE PUEDEN UTILIZAR DATOS BINARIOS Y CUANTITATIVOS

Bray-Curtis

Este coeficiente fue descrito por Odum (1950) y luego por Bray y Curtis (1957).

$$BC = \frac{\sum \left| X_{Aj} - X_{Bj} \right|}{\sum \left(X_{Aj} + X_{Bj} \right)}$$

Como primer ejemplo, utilizaremos la matriz con datos de abundancia MBD1 (Tabla 4.1):

$$BC = \frac{|12-14| + |1-5| + |0+11|}{(12+14) + (1+5) + (0+11)}$$

$$BC = 0,39$$

Como segundo ejemplo, utilizaremos la matriz de presencia-ausencia utilizada para el coeficiente de Simpson (Tabla 4.7) para las UEA y B:

$$BC_{AB} = \frac{|0-1| + |1-1| + |0-1| + |0-1| + |0-0|}{(0+1) + (1+1) + (0+1) + (0+1) + (0+1) + (0+0)}$$

$$BC_{AB} = \frac{1+1+1}{1+2+1+1+1}$$

$$BC_{AB} = \frac{3}{6} = 0,5$$

Observe que para datos binarios:

$$BC = \frac{2a}{2a+b+c}$$

En este caso el coeficiente de Bray-Curtis se convierte en el coeficiente de Dice-Sørensen.

Morisita

Se lo utiliza sólo para datos cuantitativos, es casi independiente del tamaño muestral, excepto para muestras muy pequeñas (Krebs 1999) y se basa en valores relativos. Para una mejor interpretación describiremos el coeficiente aplicado a una MBD de sitios × especies. Tiene en cuenta los siguientes parámetros:

1. El número de individuos de la especie *j* presentes en el sitio A, n_{Aj}.
2. El número de individuos de la especie *j* presentes en el sitio B, n_{Bj}.
3. El número total de individuos en el sitio A, $N_A = \Sigma n_{Aj}$.
4. El número total de individuos en el sitio B, $N_B = \Sigma n_{Bj}$.

A los fines prácticos volvemos a utilizar la MBD1 (Tabla 4.8), cuyos valores se relativizan al total de cada sitio (Tabla 4.9).

Tabla 4.8. MBD1 de dos sitios (A, B) × tres especies (sp1 a sp3), junto con las sumas totales por sitio.

Sitio	sp1	sp2	sp3	Suma
A	12	1	0	13
B	14	5	11	30
Suma	26	6	11	43

Tabla 4.9. MBD1 en la cual la abundancia de cada especie se divide por la abundancia total del sitio (abundancia relativa).

Sitio	*p*1	*p*2	*p*3	Suma
A	12/13 = 0,92	1/13 = 0,08	0/13 = 0,00	1
B	14/30 = 0,47	5/30 = 0,17	11/30 = 0,36	1

$$M = \frac{2\sum p_{Aj}\, p_{Bj}}{\sum p_{Aj}\left(\frac{n_{Aj}-1}{N_A-1}\right)+\sum p_{Bj}\left(\frac{n_{Bj}-1}{N_B-1}\right)}$$

$$M = \frac{2\left(0{,}92\times 0{,}47+0{,}08\times 0{,}17+0{,}00\times 0{,}36\right)}{0{,}92\left(\frac{12-1}{13-1}\right)+0{,}08\left(\frac{1-1}{13-1}\right)+0{,}00\left(\frac{0-1}{13-1}\right)+0{,}47\left(\frac{14-1}{30-1}\right)+0{,}17\left(\frac{5-1}{30-1}\right)+0{,}36\left(\frac{11-1}{30-1}\right)}$$

$$M = \frac{0{,}89}{0{,}84+0{,}21+0{,}02+0{,}12}$$

$$M = 0{,}74$$

Morisita-Horn o simplificado de Morisita

Horn (1966) propuso una modificación del coeficiente de Morisita, el cual puede utilizarse también para datos binarios. Wolda (1981) comparó varios coeficientes de similitud cuantitativos y encontró que este índice no estaba fuertemente influenciado por la riqueza de especies y el tamaño muestral. Una desventaja de este coeficiente es que distorsiona el resultado cuando existe una especie muy abundante.

Sólo utiliza las proporciones al cuadrado en el denominador, a diferencia de Morisita que utiliza las proporciones y el número de individuos totales:

$$MH = \frac{2\sum p_{Aj}\, p_{Bj}}{\sum p_{Aj}^2+\sum p_{Bj}^2}$$

$$MH = \frac{2\left(0{,}92\times 0{,}47+0{,}08\times 0{,}17+0{,}00\times 0{,}36\right)}{0{,}92^2+0{,}08^2+0{,}00^2+0{,}47^2+0{,}17^2+0{,}36^2}$$

$$MH = \frac{0{,}89}{0{,}85+0{,}01+0{,}22+0{,}03+0{,}13}$$

$$MH = 0{,}72$$

Gower

Gower (1971) propuso una medida de similitud que puede combinar diferentes tipos de variables (cuantitativas y/o cualitativas) de acuerdo con sus propiedades matemáticas. La similitud entre dos UE para *p* variables es el promedio de las similitudes calculadas para todas las UE:

$$G = \frac{1}{p}\sum S_{ABj}$$

Para cada variable *j* la "similitud parcial" S_{AB} entre dos UE se calcula de diferentes maneras según el tipo de dato:

1. *Datos binarios (presencia-ausencia).* Similitud igual a 1 (concordancias, *a* y *d*) ó 0 (discordancias, *b* y *c*). Gower (1971) propuso dos variantes para considerar la similitud de las dobles ausencias (*d*): *S* = 1 (forma simétrica) y *S* = 0 (forma asimétrica).

2. *Datos cuantitativos (continuos y discretos).* Para cada variable, primero se calcula la diferencia absoluta entre ambas UE, $|X_{Aj} - X_{Bj}|$. Luego este valor se divide por el rango de la variable analizada. Dado que este valor representa una distancia, debe restársela a 1 para transformarse en similitud:

$$S_{ABj} = 1 - \frac{\left|X_{Aj} - X_{Bj}\right|}{\max\left(X_j\right) - \min\left(X_j\right)}$$

3. *Datos cualitativos nominales.* Se convierten a variables ficticias (*dummies*), es decir que cada categoría de la variable se convierte en una nueva variable codificada con ceros y unos (ver en el Cap. 2 *Codificación de datos cuantitativos*). Se tratan de igual forma que las variables binarias, *S* = 1 (concordancias) y *S* = 0 (discordancias), y los dobles ceros se pueden tratar de forma simétrica o asimétrica.

4. *Datos cualitativos ordinales.* Cuando se utilizan datos ordinales como si fueran cuantitativos, se podrían utilizar las diferencias entre valores de las categorías. En estos casos, es importante estar seguro de que las distancias entre categorías consecutivas sean comparables en magnitud. Por ejemplo, con variables codificadas como 1, 2 y 3, la distancia puede ser utilizada sólo si la diferencia entre 1 y 2 es equivalente a la distancia entre 2 y 3. Si dicha diferencia es sustancial, estas distancias no son comparables y deben transformarse. Podani (1999) propuso un método para cuantificar la similitud en datos ordinales, donde se transforman los valores a rangos *r*:

$$S_{ABj} = 1 - \frac{\left|r_{Aj} - r_{Bj}\right| - \frac{T_{Aj} - 1}{2} - \frac{T_{Bj} - 1}{2}}{\max\left(r_j\right) - \min\left(r_j\right) - \frac{\max\left(T_j\right) - 1}{2} - \frac{\min\left(T_j\right) - 1}{2}}$$

T_{Aj} y T_{Bj} son el número total de UE que tienen el mismo rango de A y B, respectivamente, para la variable. Los valores max(T_j) y min(T_j) son los números totales de UE que tienen el rango máximo y mínimo, respectivamente.

El coeficiente de Gower no realiza ningún cálculo para las variables con datos faltantes. La presencia de un dato faltante se describe mediante el valor w_j que representa la presencia o ausencia de información: $w_j = 0$, cuando hay un dato faltante para al menos una UE o una ausencia compartida; $w_j = 1$, cuando no hay datos faltantes. La forma final del coeficiente de Gower es:

$$G = \frac{\sum w_j S_{ABj}}{\sum w_j}$$

Tomamos como ejemplo una nueva MBD4 hipotética (Tabla 4.10) con una variable continua (longitud de la hoja), una variable categórica (color de la flor) con dos estados (roja y blanca), una variable binaria (presencia-ausencia de raíz secundaria) con un dato faltante (NA) y una variable ordinal (densidad de glándulas). La densidad de glándulas corresponde a una variable cuantitativa (glándulas/cm^2), transformada a una escala ordinal (1 = 0–10 glándulas/cm^2, 2 = 11–100 glándulas/cm^2).

Tabla 4.10. MBD4 de tres especies de plantas (sp1 a sp3) × cinco caracteres morfológicos.

Especie	Longitud de la hoja (cm)	Flor roja	Flor blanca	Raíz secundaria	Densidad de glándulas (cm^2)
sp1	10,0	0	0	NA	2
sp2	9,7	0	1	0	1
sp3	5,0	1	1	1	2

Como se dijo anteriormente, el cálculo se realiza de a una variable a la vez dependiendo del tipo de dato.

Similitud parcial - longitud de la hoja (variable 1). Como hay información para ambas UE, $w_1 = 1$, y es una variable continua:

$$S_{AB1} = 1 - \frac{|10 - 9{,}7|}{10{,}0 - 5{,}0}$$

$$S_{AB1} = 0{,}94$$

Similitud parcial - flor roja (variable 2). En este caso consideraremos las dobles ausencias como asimétricas, por lo que la similitud $S_{AB2} = 0$. Si bien hay información para ambas UE, representa una ausencia compartida, por lo que $w_2 = 0$.

Similitud parcial - flor blanca (variable 3). Como hay información para ambas UE, $w_3 = 1$. Como es una discordancia, $S_{AB3} = 0$.

Similitud parcial - raíz secundaria (variable 4). Como hay un dato faltante para la especie A, $w_4 = 0$.

Similitud parcial - densidad de glándulas (variable 5). Como hay información para ambas UE, $w_5 = 1$, y al ser una variable ordinal, se transforman los valores a rangos (Tabla 4.11).

Tabla 4.11. Transformación de la variable "densidad de glándulas" a rangos.

Especie	Densidad de glándulas	Rango
A	2	2,5
B	1	1
C	2	2,5

En teoría debería haber tres rangos (1, 2 y 3). Como las especies A y C tienen el mismo valor (empate), no se puede establecer cuál tiene el rango 2 y cuál el rango 3, por lo que se les asigna el valor promedio (2,5). Si hubiera otra especie con mayor valor que el resto, debería tomar el rango 4 (porque 1, 2 y 3 ya se utilizaron).

En este caso, $T_{Aj} = 2$ (número de veces que aparece el rango de la especie A), $T_{Bj} = 1$ (número de veces que aparece el rango de la especie B), $\max(T_j) = 2$ (número de veces que aparece el rango máximo) y $\min(T_j) = 1$ (número de veces que aparece el rango mínimo).

$$S_{AB5} = 1 - \frac{|2{,}5 - 1| - \frac{2-1}{2} - \frac{1-1}{2}}{2{,}5 - 1 - \frac{2-1}{2} - \frac{1-1}{2}}$$

$$S_{AB5} = 1 - \frac{1{,}5 - 0{,}5}{1{,}5 - 0{,}5}$$

$$S_{AB5} = 0$$

Finalmente, realizamos el cálculo de la similitud de la siguiente forma:

$$G = \frac{w_1 S_{AB1} + w_2 S_{AB2} + w_3 S_{AB3} + w_4 S_{AB4} + w_5 S_{AB5}}{w_1 + w_2 + w_3 + w_4 + w_5}$$

$$G = \frac{1 \times 0,94 + 0 \times 0 + 1 \times 0 + 0 \times 0 + 1 \times 0}{1 + 0 + 1 + 0 + 1}$$

$$G = \frac{0,94}{3}$$

$$G = 0,31$$

COEFICIENTES DE CORRELACIÓN

Es posible cuantificar la similitud mediante el ángulo entre dos vectores que parten del origen de las coordenadas, y pasan por las UEA y B en el espacio euclideano. El coeficiente de correlación es función de ese ángulo (Fig. 4.5). Existen numerosos coeficientes de correlación, por ejemplo, para variables categóricas, ordinales o combinaciones de variables continuas y categóricas (ρ de Spearman, τ de Kendall, coeficiente *V* de Cramér, correlación policórica, correlación biserial puntual) aunque el más empleado es el de correlación producto-momento de Pearson (Galton 1877, Pearson 1895, Zar 1999), introducido en la taxonomía numérica por Michener y Sokal (1957) y utilizado para datos cuantitativos. Sin embargo, en la práctica el coeficiente de correlación suele utilizarse para analizar la similitud entre variables (modo R, ver Box 2.2). La correlación de Pearson se calcula como el cociente entre la covarianza y el producto del desvío estándar de cada variable (raíz cuadrada de la varianza):

$$r = \frac{S_{jk}}{S_j S_k}$$

$$r = \frac{\sum \left(X_{ij} - \bar{X}_j \right) \left(X_{ik} - \bar{X}_k \right)}{\sqrt{\sum \left(X_{ij} - \bar{X}_j \right)^2} \sqrt{\left(X_{ik} - \bar{X}_k \right)^2}}$$

Donde $\bar{X}_j$ y $\bar{X}_k$ son las medias de las variables *j* y *k*, respectivamente. En la fórmula final no aparecen los tamaños muestrales, dado que están multiplicando y dividiendo. El coeficiente de correlación de Pearson varía entre –1 (correlación negativa perfecta) y +1 (correlación positiva perfecta). Un valor igual a 0 indica que no hay correlación (variables independientes). Para datos centrados (se le resta la media a cada variable de forma que $\bar{X}_j$ y $\bar{X}_k$ sean iguales a 0), el coeficiente de correlación corresponde al coseno del ángulo θ entre los dos vectores, que se conoce como similitud del coseno (Fig. 4.5).

$$r = \cos\theta = \frac{\sum X_{ij} X_{ik}}{\sqrt{\sum X_{ij}^2} \sqrt{\sum X_{ik}^2}}$$

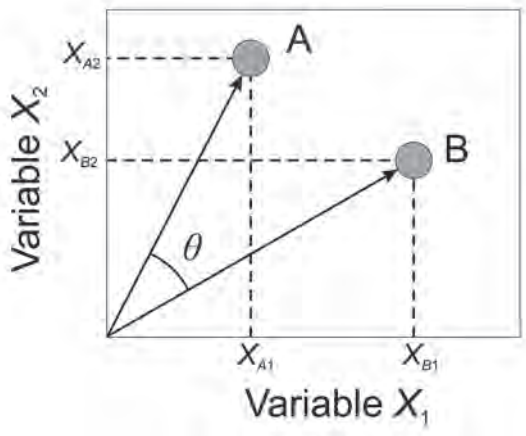


Fig. 4.5. Coeficiente de correlación de Pearson entre dos UE (círculos grises). Corresponde al coseno del ángulo θ entre los dos vectores correspondiente a cada UE.

Si $\theta = 0°$, $r = 1$ (vectores en la misma dirección y sentido, correlación positiva); si $\theta = 90°$, $r = 0$ (vectores perpendiculares, correlación nula); si $\theta = 180°$, $r = –1$ (vectores en la misma dirección pero en sentidos opuestos, correlación negativa). Si aplicamos el coeficiente de correlación de Pearson entre las UEA y B de la Tabla 4.5, primero debemos obtener los promedios para cada UE:

$$\overline{X}_A = \frac{12+1+0}{3} = 4,3$$

$$\overline{X}_B = \frac{14+5+11}{3} = \frac{30}{3} = 10$$

Ahora, podemos utilizar la fórmula de la correlación:

$$r = \frac{(12-4,3)(14-10)+(1-4,3)(5-10)+(0-4,3)(11-10)}{\sqrt{(12-4,3)^2+(1-4,3)^2+(0-4,3)^2}\sqrt{(14-10)^2+(5-10)^2+(11-10)^2}}$$

$$r = \frac{7,7\times 4,0+(-3,3)(-5,0)+(-4,3)1,0}{\sqrt{7,7^2+(-3,3)^2+(-4,3)^2}\sqrt{4,0^2+(-5,0)^2+1,0^2}}$$

$$r = 0,70$$

ELECCIÓN DEL COEFICIENTE DE SIMILITUD: Y AHORA ¿QUÉ HAGO CON MIS DATOS?

La elección del coeficiente de similitud que se va a utilizar está supeditada al tipo de datos que contiene la MBD y al objetivo del estudio. Por ejemplo, los coeficientes de asociación en general sólo se aplican a datos binarios, mientras que el coeficiente de correlación de Pearson no da buenos resultados con este tipo de datos. En la Tabla 4.12 se resumen algunas características de estos coeficientes, incluyendo propiedades, tipo de dato al que se aplica, valor mínimo y máximo de similitud.

Se ha demostrado empíricamente (Sneath y Sokal 1973) que en aquellos estudios en los cuales la MBD presenta datos binarios y nominales multiestado, y predominan las variables con datos binarios, es conveniente transformar los datos nominales multiestado en datos binarios y utilizar coeficientes de asociación. En aquellos estudios en los cuales predominan las variables con datos nominales multiestado, es aconsejable la estandarización (Box 4.1) y la utilización de coeficientes de distancia y correlación.

Por último, es importante resaltar que dos valores de similitud entre dos UE que surjan de dos MBD diferentes no son comparables. Sólo son comparables los valores de similitud obtenidos en un mismo estudio (que surjan de la misma MBD). Por ejemplo, el valor de un coeficiente de Jaccard de 0,52 entre dos especies de plantas no es comparable con un valor de 0,80 entre dos especies de mamíferos y, por lo tanto, no puede afirmarse que esas dos especies de plantas son menos similares entre sí que las dos especies de mamíferos.

Tabla 4.12. Resumen de los coeficientes de similitud tratados en este capítulo.

Tipo de coeficiente	Coeficiente	Propiedades	Tipos de datos	Mínima similitud	Máxima similitud
Distancia	Euclideana	Refleja distancias absolutas	Cuantitativos continuos y discretos	$+\infty$	0
	Manhattan	Refleja distancias absolutas. Funciona mejor que la distancia euclideana para matrices con muchas variables			
	Taxonómica	Penalizan por el número de variables utilizadas			
	Diferencia de carácter promedio				
	Canberra	Las diferencias entre las especies más abundantes contribuyen más a la similitud que las diferencias entre las especies raras			
	Mahalanobis	Penaliza por el desvío estándar de las variables utilizadas			
	Cao	Útil para sitios con poco sesgo, baja similitud y distintas intensidades de muestreo	Cuantitativos discretos		
	Chi-cuadrado	Tiene en cuenta las abundancias relativas. Las diferencias entre las especies más abundantes contribuyen más a la similitud que las diferencias entre las especies raras		0	$\sqrt{N \times \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$

Tipo de coeficiente	Coeficiente	Propiedades	Tipos de datos	Mínima similitud	Máxima similitud
Asociación	*Simple Matching*	Las concordancias y discordancias contribuyen de igual manera a la similitud	Binarios	0	1
	Rogers y Tanimoto	Las discordancias contribuyen en mayor medida a la similitud que las concordancias			
	Sokal y Sneath	Las concordancias contribuyen en mayor medida a la similitud que las discordancias			
	Russell y Rao	Las ausencias compartidas son un elemento en contra de la similitud			
	Jaccard	Las ausencias compartidas no son tenidas en cuenta para el cálculo de la similitud			
	Dice-Sørensen	Da mayor peso a las presencias compartidas que Jaccard			
	Simpson	Dos UE anidadas o iguales tienen similitud máxima			
	Kulczynski	Penalizan las discordancias		0	$+\infty$
	Hamann			–1	1
	Bray-Curtis	Las ausencias compartidas no son tenidas en cuenta para el cálculo de la similitud. Tiene en cuenta los valores absolutos	Cuantitativos y binarios	0	1
	Morisita-Horn	Tiene en cuenta los valores relativos			
	Morisita	Tiene en cuenta los valores relativos. Es casi independiente del tamaño muestral, excepto para muestras muy pequeñas. Es uno de los más recomendados	Cuantitativos		
	Gower	Único que admite datos nominales multiestado y ordinales	Cuantitativos y cualitativos		
Correlación	Pearson	Analiza la relación lineal (positiva, negativa o nula) entre las UE	Cuantitativos	–1	1

MATRIZ DE SIMILITUD

Los resultados obtenidos de la aplicación de cualquiera de los coeficientes de similitud para cada par posible de UE de una MBD, constituyen la matriz de similitud (MS) (Fig. 4.6). De acuerdo al coeficiente utilizado se denominan matriz de distancia, de asociación o de correlación.

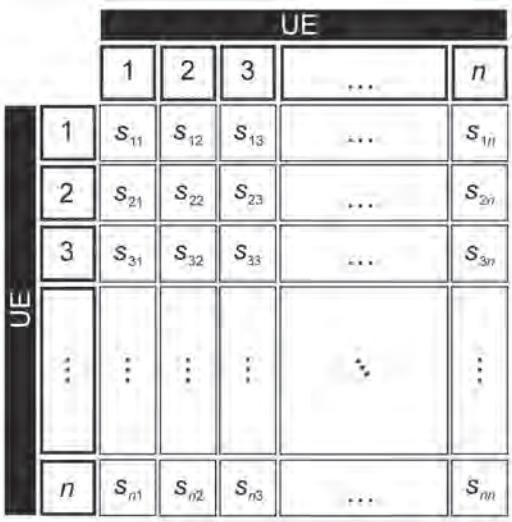


Fig. 4.6. Matriz de similitud de $n \times n$ UE. Cada celda corresponde a un valor de similitud S_{ij}.

La MS es una matriz de $n \times n$ UE; cada celda de la matriz S_{ij} representa el valor de similitud entre la UE *i* y la UE *j*. Las UE ocupan tanto las filas como las columnas, siguiendo el mismo orden en ambas; de esta manera se logra comparar cada UE consigo misma y con las restantes. Así estructurada la matriz, cada valor de la diagonal principal (S_{11}, S_{22}, S_{33}, ..., S_{nn}) representa a cada UE comparada consigo misma; este valor es trivial y corresponde al de la máxima similitud; 1 en el caso de la mayoría de los coeficientes de asociación y correlación, y 0 en los coeficientes de distancia (Tabla 4.12). La similitud entre la UE1 y la UE2 (S_{12}) es la misma que entre la UE2 y la UE1 (S_{21}), por lo que la parte superior derecha de la matriz es la imagen especular de la parte inferior izquierda.

La MS tiene, por lo tanto, dos características principales: es cuadrada (posee la misma cantidad de filas que de columnas) y es simétrica (se pueden intercambiar filas por columnas sin alterar la matriz ya que $S_{ij} = S_{ji}$). Es por este motivo que en la práctica se suele mostrar sólo el triángulo inferior izquierdo o superior derecho de la matriz.

Las Tablas 4.13 y 4.14 representan la MS obtenida de la aplicación de la distancia taxonómica y del coeficiente de correlación, respectivamente, sobre la MBD del ejemplo en el género *Bulnesia*.

Tabla 4.13. Matriz de similitud basada en la distancia taxonómica sobre la MBD de especies de *Bulnesia*, previamente estandarizada. BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*.

Especie	BAR	BCA	BCH	BBO	BRE	BFO	BSC	BSA
BAR	0,00	0,69	1,65	1,20	1,58	1,59	1,70	1,67
BCA	0,69	0,00	1,80	1,31	1,73	1,66	1,84	1,74
BCH	1,65	1,80	0,00	1,22	1,01	1,18	1,23	1,57
BBO	1,20	1,31	1,22	0,00	1,22	1,12	1,13	1,53
BRE	1,58	1,73	1,01	1,22	0,00	1,00	1,13	1,53
BFO	1,59	1,66	1,18	1,12	1,00	0,00	0,74	1,28
BSC	1,70	1,84	1,23	1,13	1,13	0,74	0,00	1,56
BSA	1,67	1,74	1,57	1,53	1,53	1,28	1,56	0,00

Tabla 4.14. Matriz de similitud basada en el coeficiente de correlación de Pearson sobre la MBD de especies de *Bulnesia*. BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*.

Especie	BAR	BCA	BCH	BBO	BRE	BFO	BSC	BSA
BAR	1,00	0,73	–0,40	0,25	–0,48	–0,61	–0,49	–0,11
BCA	0,73	1,00	–0,50	0,17	–0,56	–0,46	–0,54	0,02
BCH	–0,40	–0,50	1,00	–0,16	0,35	–0,08	0,03	–0,25
BBO	0,25	0,17	–0,16	1,00	–0,28	–0,33	–0,07	–0,37
BRE	–0,48	–0,56	0,35	–0,28	1,00	0,17	0,11	–0,24
BFO	–0,61	–0,46	–0,08	–0,33	0,17	1,00	0,54	–0,04
BSC	–0,49	–0,54	0,03	–0,07	0,11	0,54	1,00	–0,33
BSA	–0,11	0,02	–0,25	–0,37	–0,24	–0,04	–0,33	1,00

Box 4.1. Escalas y transformaciones

Es frecuente que en una MBD coexistan variables en diferentes escalas y/o con distintas unidades de medida. Por ejemplo, la variable "diámetro del grano de polen" puede estar expresada en micrones, mientras que la variable "altura de la planta" puede estar expresada en centímetros. Si igualamos las escalas y unidades expresando la altura de la planta en micrones, habría que ubicar en la MBD una cifra enorme que distorsionaría los cálculos. También pueden presentarse variables en distintas unidades de medida (por ejemplo, longitud y peso). Podría darse incluso que una MBD contenga variables con igual escala e igual unidad de medida, pero que por una gran variación en una o más variables sea necesario la transformación de la matriz para hacer comparables los datos. En estos casos, es necesario transformar la MBD para poder expresar todos los valores en una misma escala y poder compararlos (Legendre y Legendre 1998).

La transformación más simple es el centrado, en el cual se resta la media (Box 2.1) de una variable a cada UE (Quinn y Keough 2002):

$$x_{ij} = X_{ij} - \overline{X}_j$$

Esta transformación expresa los valores en términos de distancia con respecto a la media. Como resultado, la media de esta nueva variable es 0, eliminando los efectos de magnitud debido a la posición de la media en las diferentes variables de la MBD (Legendre y Legendre 1998). Sin embargo, las UE siguen siendo expresadas en las mismas unidades que en la MBD. Por lo tanto, el centrado es aconsejable en aquellas MBD con variables medidas en las mismas unidades (ver Cap. 6).

La transformación más utilizada es la denominada estandarización o normalización, que consiste en expresar los valores de la MBD en unidades de desvío estándar (Legendre y Legendre 1998, Quinn y Keough 2002):

$$z_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j}$$

El numerador corresponde al centrado, mientras que el denominador corresponde al desvío estándar (Box 2.1). De aquí se observa que la diferencia entre el centrado y la estandarización es que el primero mantiene las unidades originales de la variable, mientras que la segunda modifica las unidades a desvío estándar. La estandarización permite que variables con diferentes escalas y unidades de medida no afecten el análisis. Esta nueva variable, cuyos valores son denominados *z-scores*, tiene

tres propiedades (Legendre y Legendre 1998): (1) su media es 0, (2) su desvío estándar es 1, y (3) no tiene dimensiones debido a que las unidades del numerador y denominador se simplifican.

La estandarización suele utilizarse en la mayoría de las técnicas multivariadas, incluyendo los análisis de agrupamientos (Cap. 5) y el análisis de componentes principales (ver Cap. 6). Algunos coeficientes de similitud (Cap. 4) y técnicas de ordenación, como el análisis de correspondencias, transforman las variables a cantidades relativas, lo que permite la comparación entre diferentes variables (o alternativamente UE) sin necesidad de transformar la MBD, pues la transformación la genera la misma técnica al momento de realizar los cálculos (Cap. 6).

Otra transformación posible consiste en convertir los datos a rangos. Este tipo de transformación es la que genera mayor pérdida de información, porque se asume que la distancia entre un valor y el siguiente de una variable es el mismo (ver en este capítulo *Datos ordinales*). Sin embargo, reduce la influencia de valores extremos en la MBD y es utilizada por técnicas de ordenación no métricas (Cap. 6).

En la siguiente tabla se muestra un ejemplo de diferentes tipos de transformaciones de una variable con tres valores: 2, 3 y 7. La media es 4 y el desvío estándar 2,64.

Valor observado	Valor centrado	Valor estandarizado
2	2 – 4 = –2	–2/2,64 = –0,76
3	3 – 4 = –1	–1/2,64 = –0,38
7	7 – 4 = 3	3/2,64 = 1,14

El tipo de transformación aplicado a la MBD dependerá de las escalas y unidades de las variables y del tipo de análisis que se quiera realizar. Existen muchas otras transformaciones que pueden ser aplicadas a una MBD (para más detalles ver Legendre y Legendre 1998).

COEFICIENTES DE SIMILITUD EN R

En R hay diversos paquetes que permiten calcular numerosos coeficientes de similitud. En particular se presentarán dos paquetes que consideran un gran número de coeficientes, los paquetes vegan (Oksanen *et al.* 2018) y proxy (Meyer y Buchta 2019). Otra función que cuenta con varios coeficientes de similitud es la función `dist.binary()` del paquete ade4 (Chessel *et al.* 2004). Se debe tener en cuenta que, en este último caso, los coeficientes de asociación se calculan como la raíz cuadrada de 1 menos la distancia, `s <- sqrt(1 - d)`, donde `d` es una medida de distancia. Por lo tanto, si se prefiere obtener el resultado en la escala original, debe elevarse al cuadrado y restársele a 1, `d = 1 - s^2`. Para calcular distancias entre secuencias de ADN se puede utilizar la función `dist.dna()` del paquete ape (Paradis y Schliep 2018) o la función `dist.genpop()` del paquete adegenet (Jombart 2008).

A continuación vamos a reconstruir las MBD utilizadas anteriormente, uniendo dos sitios como filas –función `rbind()`– y luego convirtiendo el resultado en un marco de datos –función `data.frame()`–. Primero necesitamos construir la MBD1 (Tabla 4.1).

```
> sitioA <- c(12, 1, 0)
> sitioB <- c(14, 5, 11)
> MBD1 <- data.frame(rbind(sitioA, sitioB))
> MBD1
       X1 X2 X3
sitioA 12  1  0
sitioB 14  5 11
```

R tiene incorporados varios coeficientes de distancia, por lo que no es necesario descargar ningún paquete. Entre estos coeficientes se encuentran la distancia euclideana, la distancia de Manhattan y la distancia de Canberra.

```
> D <- dist(MBD1, method = "euclidean")
> D
         sitioA
sitioB 11.87434
> M <- dist(MBD1, method = "manhattan")
> M
       sitioA
sitioB     17
> C <- dist(MBD1, method = "canberra")
> C
        sitioA
sitioB 1.74359
```

La distancia taxonómica corresponde al cociente entre la distancia euclideana y la raíz cuadrada del número de variables, que representa el número de columnas de la MBD.

```
> p <- ncol(MBD1)
> DT <- D/sqrt(p)
> DT
         sitioA
sitioB 6.855655
```

De forma similar, la diferencia de carácter promedio puede calcularse como el cociente entre la distancia de Manhattan y el número de variables.

```
> MCD <- M/p
> MCD
         sitioA
sitioB 5.666667
```

Para las distancias de Cao y chi-cuadrado utilizaremos el paquete vegan (Oksanen *et al.* 2018) y las funciones `vegdist()` y `decostand()`, respectivamente. La función `vegdist()` presenta un gran número de coeficientes de distancia (que pueden indagarse con el comando `?vegdist`), y algunas de las cuales serán presentadas como coeficientes de asociación. Se debe tener en cuenta que en algunos casos (ver abajo) algunos coeficientes de asociación pueden convertirse a coeficientes de distancia, en cuyo caso deben calcularse como `a = 1 - d`, donde `d` es un coeficiente de distancia y `a` un coeficiente de asociación.

```
> library(vegan)
> CYd <- vegdist(MBD1, method = "cao")
> CYd
         sitioA
sitioB 1.604435

> chi <- dist(decostand(MBD1, method = "chi.square"), method = "euclidean")
> chi
          sitioA
sitioB 0.9632178
```

Para calcular la distancia de Mahalanobis usaremos la MBD de la Figura 4.3 y la función `mahalanobis.dist()` del paquete StatMatch (D'Orazio 2019)

```
> df <- data.frame(x1 = c(5.5, 2.1, 3.0, 4.2, 4.1),
```

```
+                  x2 = c(8.1, 5.5, 4.2, 8.3, 8.5),
+                  x3 = c(200, 225, 213, 190, 185))
> library(StatMatch)
> MHL <- mahalanobis.dist(data.x = df)
> MHL
         1        2        3         4         5
1 0.000000 2.825219 2.812660 2.1642847 2.6619660
2 2.825219 0.000000 2.823694 2.2707872 2.6096782
3 2.812660 2.823694 0.000000 2.3835520 2.5351330
4 2.164285 2.270787 2.383552 0.0000000 0.5371456
5 2.661966 2.609678 2.535133 0.5371456 0.0000000
```

En este caso, el resultado es la MD entre cada par de UE. Para el cálculo de los coeficientes de asociación construiremos la matriz de datos binarios MBD2 (Tabla 4.6).

```
> sitioA <- c(0, 1, 1, 0, 1, 0, 1, 0, 1, 0)
> sitioB <- c(1, 1, 0, 0, 1, 1, 0, 0, 1, 1)
> MBD2 <- data.frame(rbind(sitioA, sitioB))
> MBD2
       X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
sitioA  0  1  1  0  1  0  1  0  1   0
sitioB  1  1  0  0  1  1  0  0  1   1
```

Para calcular los coeficientes de *simple matching*, Rogers y Tanimoto, Russell y Rao, Kulczynski, Hamann, Jaccard, Dice-Sørensen y Simpson utilizaremos el paquete proxy (Meyer y Buchta 2019).

```
> library(proxy)
> SM <- simil(MBD2, method = "simple matching")
> SM
       sitioA
sitioB    0.5

> RT <- simil(MBD2, method = "Tanimoto")
> RT
          sitioA
sitioB  0.3333333

> RR <- simil(MBD2, method = "Russel")
       sitioA
sitioB    0.3

> K <- simil(MBD2, method = "Kulczynski1")
> K
       sitioA
sitioB    0.6

> H <- simil(MBD2, method = "Hamman")
> H
       sitioA
sitioB      0
```

```
> J <- simil(MBD2, method = "Jaccard")
> J
       sitioA
sitioB  0.375

> DS <- simil(MBD2, method = "Sorensen")
> DS
          sitioA
sitioB 0.5454545

> S <- simil(MBD2, method = "Simpson")
> S
     sitioA
sitioB  0.6
```

Para calcular el coeficiente de Sokal y Sneath no hay un paquete disponible. Por lo tanto, lo calcularemos manualmente obteniendo primero las sumas de las columnas –función `colSums()`–.

```
> sum.col <- colSums(MBD2)
> sum.col
 X1  X2  X3  X4  X5  X6  X7  X8  X9 X10
 1    2   1   0   2   1   1   0   2   1
```

A partir de aquí, es fácil calcular las presencias compartidas *a* como aquellas sumas iguales a 2 (ya que hay dos unos por cada columna), las ausencias compartidas *d* como aquellas sumas iguales a 0, y las presencias sólo en uno de los sitios (*b* + *c*) como aquellas sumas iguales a 1. Para esto, indexamos el vector `sum.col` y le indicamos que seleccione aquellos valores con la condición igual a 2, igual a 0 e igual a 1. Luego, contamos el número de elementos de dicho vector con la función `length()`. En este caso no es necesario distinguir entre *b* y *c*, lo que facilita el cálculo.

```
> a <- length(sum.col[sum.col == 2])
> d <- length(sum.col[sum.col == 0])
> bc <- length(sum.col[sum.col == 1])
> SS <- (2*a + 2*d)/(2*a + bc + 2*d)
> SS
[1] 0.6666667
```

A continuación, vamos a calcular los coeficientes de Bray-Curtis, Morisita y Morisita-Horn sobre la MBD1 mediante la función `vegdist()` del paquete vegan (Oksanen *et al.* 2018). En el caso de los coeficientes de Morisita y Morisita-Horn, la asociación debe calcularse como 1 – distancia. En el caso del coeficiente de Bray-Curtis la formulación ya está expresada como asociación, a pesar de que la función `vegdist()` hace referencia a distancia.

```
> BC <- vegdist(MBD1, method = "bray")
> BC
          sitioA
sitioB 0.3953488

> M <- 1 - vegdist(MBD1, method = "morisita")
> M
          sitioA
sitioB 0.7363863
```

```
> MH <- 1 - vegdist(MBD1, method = "horn")
> MH
         sitioA
sitioB 0.71663
```

El coeficiente de Gower se puede calcular utilizando el paquete FD (Laliberté *et al.* 2014) sobre la MBD4 (Tabla 4.10), que presenta una variable continua (X1), tres variables binarias (X2 a X4) y una variable ordinal (X5).

```
> X1 <- c(10, 9.7, 5.0)
> X2 <- c(0, 0, 1)
> X3 <- c(0, 1, 1)
> X4 <- c(NA, 0, 1)
> X5 <- c(2, 1, 2)
> MBD4 <- data.frame(X1, X2, X3, X4, X5)
> MBD4
    X1 X2 X3 X4 X5
1 10.0  0  0 NA  2
2  9.7  0  1  0  1
3  5.0  1  1  1  2
```

Si la MBD contiene variables ordinales, debemos primero convertirlas a factor –función `factor()`–, con los estados ordenados (argumento `ordered = TRUE`).

```
> MBD4$X5 <- factor(X5, ordered = TRUE)
```

Por otro lado, en el caso de los datos binarios podemos especificar si deseamos que las dobles ausencias se traten como 1 (simétricas) ó 0 (asimétricas), utilizando el argumento `asym.bin`. En éste último debe indicarse el número de columnas correspondientes a las variables binarias de interés (columna 2 en nuestro caso). En el caso de las variables ordinales, podemos especificar el método utilizado para calcular su similitud parcial (argumento `ord`). En nuestro ejemplo vamos a utilizar el método de Podani (1999). Finalmente, al calcularse como un coeficiente de distancia debe restársele a de 1.

```
> library(FD)
> G <- 1 - gowdis(MBD4, asym.bin = 2, ord = "podani")
> G
          1         2
2 0.3133333
3 0.2500000 0.2120000
```

Por último, vamos a calcular la correlación momento-producto de Pearson sobre la MBD1, que está disponible por defecto en R.

```
> sitioA <- c(12, 1, 0)
> sitioB <- c(14, 5, 11)
> r <- cor(sitioA, sitioB, method = "pearson")
> r
[1] 0.704634
```

Todas las funciones vistas anteriormente son aplicables a una MBD de cualquier tamaño. En los casos en que haya más de dos UE, el resultado siempre será una MS.

CAPÍTULO 5
VISUALIZANDO SIMILITUDES ENTRE UNIDADES DE ESTUDIO: ANÁLISIS DE AGRUPAMIENTOS

La matriz de similitud (MS) es insuficiente para expresar relaciones entre la totalidad de las unidades de estudio (UE), pues sólo expone similitudes entre pares de dichas unidades. Con el fin de reconocer las relaciones entre todas las UE analizadas, disponemos de una gran variedad de técnicas de análisis de la MS (Sneath y Sokal 1973, Legendre y Legendre 1998, Quinn y Keough 2002), que tienen como objetivo sintetizar la información que ésta contiene. En este capítulo y en el siguiente presentaremos dos de las técnicas más utilizadas: el análisis de agrupamientos (*cluster analysis*; Cap.5) y los métodos de ordenación (*ordination*; Cap. 6).

El análisis de agrupamientos comprende técnicas que, siguiendo reglas arbitrarias, forman grupos de UE que se asocian por su grado de similitud. Esta definición es poco precisa y ello se debe a dos factores: primero, al escaso acuerdo entre los investigadores acerca de cómo reconocer los límites entre grupos y segundo, a la enorme variedad de técnicas propuestas. Las numerosas técnicas de análisis de agrupamientos han sido estudiadas por Ball (1965), Williams y Dale (1965), Wishart (1969), Spence y Taylor (1970), Cormack (1971), Hartigan (1975), Romesburg (1984), Jain *et al.* (1999), y Kaufman y Rousseeuw (2009).

Box 5.1. Clasificación de las técnicas de análisis de agrupamientos

a) Técnicas exclusivas *vs.* técnicas no exclusivas.
Técnicas exclusivas. Originan grupos donde las UE son exclusivas del grupo del cual forman parte y no pueden pertenecer a otro grupo que se halle en un mismo nivel o rango.
Técnicas no exclusivas. Originan grupos donde las UE pueden pertenecer a más de un grupo en un mismo nivel o rango.

b) Técnicas jerárquicas *vs.* técnicas no jerárquicas.
Técnicas jerárquicas. Originan grupos que presentan rangos, en los cuales las UE o grupos de UE subordinados forman parte de un grupo mayor o inclusivo.
Técnicas no jerárquicas. Originan grupos que no exhiben rangos.

c) Técnicas aglomerativas *vs.* técnicas divisivas. Algunos tipos de análisis (por ejemplo *K*-medias) caen por fuera de esta dicotomía.

Técnicas aglomerativas. Son aquellas en las que partiendo de *n* UE separadas, se las reune en sucesivos grupos (siempre en número menor que *n* para llegar finalmente a un solo grupo que contiene a las *n* UE).
Técnicas divisivas. Son aquellas en las que partiendo de un grupo que contiene a las *n* UE se las divide en subgrupos.

d) Técnicas secuenciales *vs.* técnicas simultáneas.
Técnicas secuenciales. Cada grupo es formado de a uno por vez hasta que se agota el conjunto total de UE.
Técnicas simultáneas. Los grupos son formados simultáneamente en una sola operación.

e) Técnicas directas *vs.* técnicas iterativas.
Técnicas directas. Construyen una clasificación de forma directa y la solución a la cual se llega se acepta como óptima. Una vez que una UE es asignada a un grupo en un cierto nivel, este agrupamiento no se modifica en los pasos posteriores (y es considerado localmente óptimo).
Técnicas iterativas. Buscan soluciones óptimas globales del análisis, son sujetas a autocorrección mediante un proceso iterativo que va mejorando la medida de optimización utilizada. La pertenencia de las UE a un determinado grupo puede ir cambiando a medida que el análisis transcurre.

f) Técnicas no supervisadas *vs.* técnicas supervisadas.
Técnicas no supervisadas. El número de grupos que se quiere establecer es definido *a posteriori* por el investigador.
Técnicas supervisadas. El número de grupos que se quiere establecer es definido *a priori* por el investigador.

De las técnicas presentadas en el Box 5.1 las más utilizadas son las exclusivas, jerárquicas, aglomerativas, secuenciales, directas y no supervisadas, las cuales se combinan caracterizando a la mayoría de las técnicas de agrupamientos que describiremos en este capítulo. Dentro de las mismas hemos elegido, por ser las más sencillas, las del llamado "grupo par" (*pair group*), en las cuales solamente puede ser admitida una UE o un grupo de UE por nivel. Esto significa que los grupos formados en cualquier etapa de los agrupamientos contienen sólo dos miembros.

A continuación describiremos el método general para las técnicas del grupo par, con sus variantes:

1. Se examina la MS para identificar el mayor valor de similitud presente, descartando la diagonal principal. Se identifican así las dos UE que formarán el primer grupo. En algunos casos puede haber más de un valor máximo de similitud, es decir, otro par o pares de UE presentan igual valor que el anterior. En este caso se construyen dos o más grupos por separado.
2. Se construye una nueva MS derivada de la MS original, que al menos puede obtenerse por tres métodos diferentes: ligamiento simple (*simple linkage*), completo (*complete linkage*) y promedio (*average linkage*).

3. Se busca el próximo valor de mayor similitud en la MS. En las primeras etapas del proceso de agrupamiento, el hallazgo de este nuevo valor puede llevar a:
   a. La formación de nuevos grupos;
   b. la incorporación de una UE a un grupo ya existente para formar un nuevo grupo, o
   c. la fusión de los grupos preexistentes.
4. Se repite la tercera etapa del proceso hasta que todos los grupos estén unidos y en ellos se incluya la totalidad de las UE.

Otro método del grupo par es el método de Ward, que opera de modo diferente ya que no se calcula a partir de una MS y será explicado más adelante en este capítulo.

TÉCNICAS EXCLUSIVAS, JERÁRQUICAS, AGLOMERATIVAS, SECUENCIALES, DIRECTAS Y NO SUPERVISADAS

Todas estas técnicas dan como resultado un dendrograma (Fig. 5.1 a 5.9), una representación gráfica en forma de árbol que muestra las relaciones jerárquicas entre las UE según sus valores de similitud. Sin embargo, estos valores no se corresponden a los de la MBD original, sino que se encuentran distorsionados por la imposibilidad de representar una matriz de *p* variables en dos dimensiones (ver en este capítulo *Medida de la distorsión*). Los valores de similitud se expresan como una escala que acompaña al dendrograma.

Ligamiento simple

Los pasos para el ligamiento aplicados a la MBD de *Bulnesia* (Tabla 2.11) se encuentran resumidos en la Figura 5.1. El valor de mayor similitud encontrado en la MS (en este caso, el menor valor de distancia) es el que poseen *B. arborea* y *B. carrapo* que forman un grupo, unido por un valor de similitud de 0,68 (Fig. 5.1A).

Así se genera una primera matriz derivada que considera a *B. arborea* y *B. carrapo* como una nueva UE (Fig. 5.1B). Como la técnica utilizada es la del ligamiento simple, los valores que se extraen de la MS derivan de la elección del valor de similitud numéricamente menor (el más parecido) entre el par *B. arborea-B. carrapo* y las restantes UE. Por ejemplo, la similitud entre la nueva UE (*B. arborea-B. carrapo*) y *B. schickendantzii* puede ser 1,65 (*B. arborea-B. schickendantzii*) o 1,80 (*B. carrapo-B. schickendantzii*) por lo tanto se toma como valor de similitud de la nueva UE y *B. schickendantzii* como 1,65.

En la primera matriz derivada se observa que el próximo valor de mayor similitud es el que poseen *B. foliosa* y *B. schickendantzii*, que se unen formando un nuevo grupo con un valor de similitud de 0,73 (Fig. 5.1B).

La segunda matriz derivada considera no sólo a *B. arborea* y *B. carrapo* como una UE, sino también a *B. foliosa* y *B. schickendantzii* (Fig. 5.1C) y se extraen los menores valores de similitud existentes entre el nuevo grupo y el anterior, así como entre el grupo anterior y las restantes UE. En esta matriz se observa que el valor de mayor similitud es el que poseen *B. retama* y el grupo *B. foliosa-B. schickendantzii*. Esto significa que *B. retama* se une a dicho grupo originando un nuevo grupo con un valor de similitud de 0,99 (Fig. 5.1C).

La tercera matriz derivada contiene los valores de similitud del grupo *B. arborea-B. carrapo* y del grupo *B. retama-B. foliosa-B. schickendantzii* entre sí y de ambos con las restantes UE. Como se observa, el valor de mayor similitud de esta matriz es de 1,01 y une a *B. chilensis* al grupo *B. retama-B. foliosa-B. schickendantzii* (Fig. 5.1D).

La cuarta matriz derivada contiene los valores de similitud del grupo *B. arborea-B. carrapo* y del grupo *B. chilensis-B. retama-B. foliosa-B. schickendantzii* entre sí y de ambos con las restantes UE. Se observa que el nuevo valor de mayor similitud es de 1,12 y es a este nivel que se une *B. bonariensis* al grupo *B. chilensis-B. retama-B. foliosa-B. schickendantzii* (Fig. 5.1E).

La quinta matriz derivada considera el valor de similitud entre el grupo *B. arborea-B. carrapo* con respecto al grupo *B. chilensis-B. retama-B. foliosa-B. schickendantzii-B. bonariensis* y el de ambos con respecto a *B. sarmientoi*. El resultado de este paso es la unión del grupo *B. arborea-B. carrapo* al grupo *B. chilensis-B. retama-B. foliosa-B. schickendantzii-B. bonariensis* con una similitud de 1,19 (Fig. 5.1F). La sexta y última matriz derivada nos da el valor (1,27) al que se une *B. sarmientoi* a las restantes UE (Fig. 5.1G).

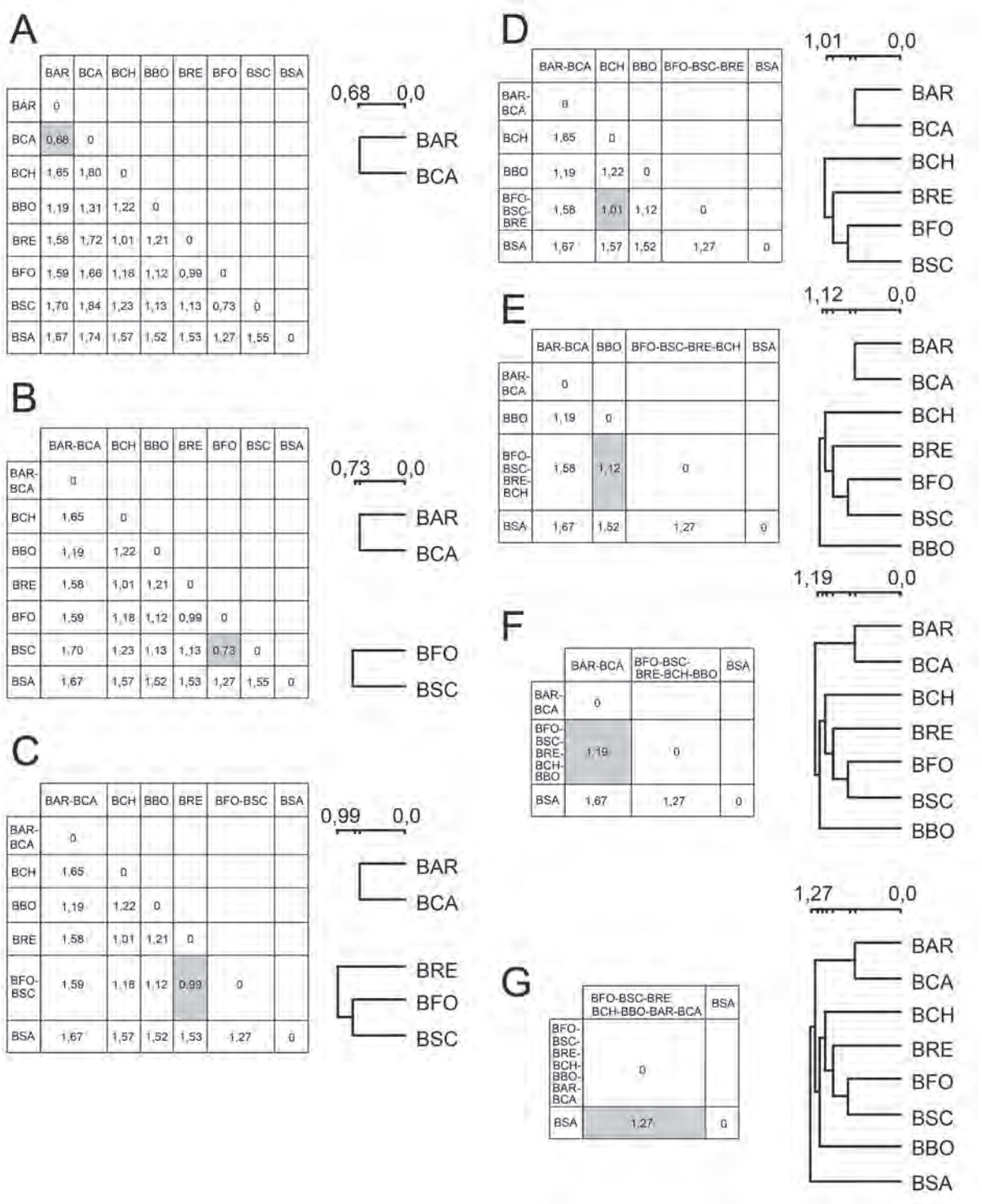


Fig. 5.1. Análisis de agrupamientos (coeficiente de distancia taxonómica, ligamiento simple) sobre la MBD estandarizada de especies de *Bulnesia*. Modificada de Crisci y López Armengol (1983). BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*. Cada letra representa un paso (incorporación de una UE) en la construcción del dendrograma.

Ligamiento completo

Para la construcción de la primera MS derivada en la que se utiliza el primer grupo formado como una nueva UE, el valor de la similitud entre la nueva UE y cada una de las otras UE analizadas corresponde al valor de menor similitud (a diferencia del ligamiento simple donde se considera el valor de mayor similitud).

El valor de mayor similitud hallado en la MS original es el que poseen *B. arborea* y *B. carrapo*, que como ya vimos forman un grupo con un valor de similitud de 0,68 (Fig. 5.2A). La primera matriz derivada considera a *B. arborea* y *B. carrapo* como un grupo y una nueva UE. Los valores obtenidos en la MS provienen de la elección del valor de similitud menor (numéricamente mayor) entre el par *B. arborea-B. carrapo* y las restantes UE. Por ejemplo, la similitud entre la nueva UE (*B. arborea-B. carrapo*) y *B. schickendantzii* puede ser 1,65 (*B. arborea- B. schickendantzii*) o 1,80 (*B. carrapo- B. schickendantzii*), por lo tanto se toma como valor de similitud de la nueva UE y *B. schickendantzii* 1,80. En esta primera matriz derivada se observa que el próximo valor de mayor similitud es el que poseen *B. foliosa* y *B. schickendantzii,* que se unen formando un nuevo grupo a un valor de similitud de 0,73 (Fig. 5.2B).

La segunda matriz derivada considera no sólo a *B. arborea* y *B. carrapo* un grupo, sino también a *B. foliosa* y *B. schickendantzii*, y se extraen los menores valores de similitud existentes entre el nuevo grupo y el anterior, y el nuevo grupo y las restantes UE. En esta matriz se observa que el valor de mayor similitud es el que poseen *B. retama* y *B. chilensis* dando origen a un tercer grupo, con un valor de similitud de 1,01 (Fig. 5.2C).

La tercera matriz derivada considera el mayor valor de similitud existente entre cada grupo comparado con los restantes grupos, y cada grupo comparado con las restantes UE. El valor de mayor similitud es el que poseen el par *B. foliosa-B. schickendantzii* y *B. bonariensis*. Esto significa que *B. bonariensis* se une a éstos originando un nuevo grupo con un valor de similitud de 1,13 (Fig. 5.2D).

La cuarta matriz derivada contiene los valores de similitud de los grupos *B. arborea-B. carrapo, B. chilensis-B. retama* y del grupo *B. bonariensis-B. foliosa-B. schickendantzii* entre sí y con las restantes UE. El valor de mayor similitud es de 1,23 entre el grupo *B. chilensis-B. retama* y el grupo *B. bonariensis-B. foliosa-B. schickendantzii*; ambos se fusionan en un nuevo grupo (Fig. 5.2E).

La quinta matriz derivada considera el valor de similitud entre el grupo *B. arborea-B. carrapo* con respecto al grupo *B. chilensis-B. retama-B. bonariensis-B. foliosa-B. schickendantzii*, y de ambos con *B. sarmientoi*. El valor de mayor similitud es de 1,57 y relaciona a *B. sarmientoi* con *B. chilensis-B. retama-B. bonariensis-B. foliosa-B. schickendantzii* (Fig. 5.2F). La sexta y última matriz derivada une el par *B. arborea-B. carrapo* con el grupo restante con un valor de similitud de 1,84 (Fig. 5.2G).

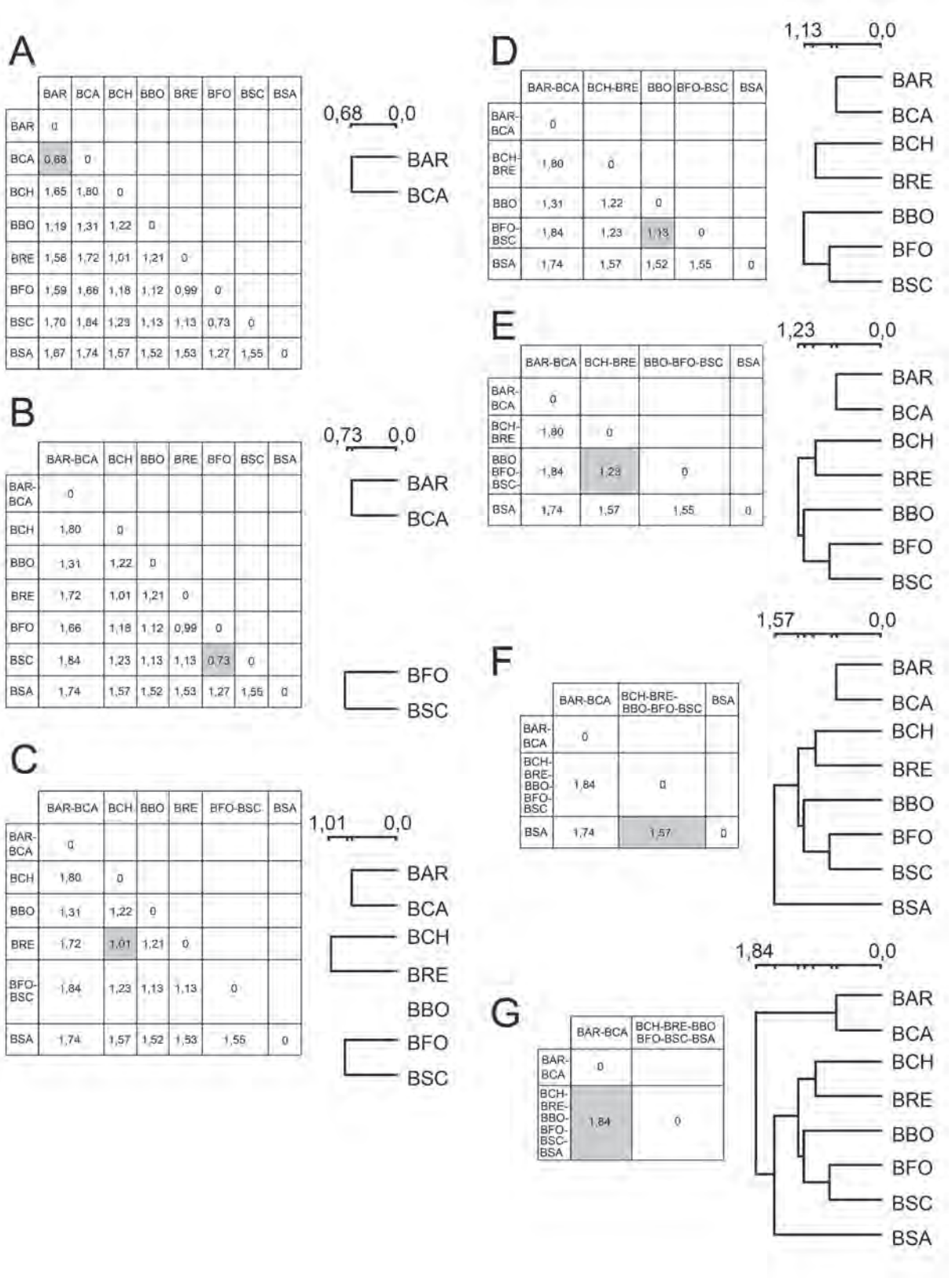


Fig. 5.2. Análisis de agrupamientos (coeficiente de distancia taxonómica, ligamiento completo) sobre la MBD estandarizada de especies de *Bulnesia*. Modificada de Crisci y López Armengol (1983). BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*. Cada letra representa un paso (incorporación de una UE) en la construcción del dendrograma.

Ligamiento promedio

En este caso, para la construcción de la MS derivada en la que se utiliza el primer grupo formado como una nueva UE, el valor de la similitud entre la nueva UE y cada una de las otras UE analizadas es el valor promedio de similitud. El valor de mayor similitud hallado en la MS original es el que poseen *B. arborea* y *B. carrapo*, que como ya vimos forman un grupo con un valor de similitud de 0,68 (Fig. 5.3A).

Como existen varios tipos de medias, es posible contar con más de una técnica de ligamiento promedio. La más utilizada es la media aritmética no ponderada (UPGMA, *unweighted pair-group method with arithmetic averages*). Si el candidato a incorporarse es un grupo en sí mismo, el valor de similitud será un promedio de los valores de similitud entre los pares posibles de UE provenientes uno de cada grupo.

La primera matriz derivada considera a *B. arborea* y *B. carrapo* como un grupo con respecto a las restantes UE. La técnica utilizada es la del ligamiento promedio no ponderado (UPGMA). Los valores que se vuelcan a la matriz derivada provienen de la media aritmética extraída de los valores de similitud del grupo *B. arborea-B. carrapo* con respecto a las demás UE. En esta primera matriz derivada se observa que el próximo valor de mayor similitud es el que poseen *B. foliosa* y *B. schickendantzii* que se unen formando un nuevo grupo, con un valor de similitud de 0,73 (Fig. 5.3B).

La segunda matriz derivada considera no sólo a *B. arborea* y *B. carrapo* como una UE, sino también a *B. foliosa* y *B. schickendantzii*. Se extraen las medias entre el nuevo grupo y el anterior, y el nuevo grupo y las restantes UE. La similitud entre el grupo ya formado *B. arborea-B. carrapo* con respecto al nuevo grupo *B. foliosa-B. schickendantzii* se calcula utilizando el método no ponderado de la siguiente manera:

a) *B. arborea-B. foliosa* = 1,59 (Fig. 5.3A)

b) *B. carrapo-B. foliosa* = 1,66 (Fig. 5.3A)

c) *B. arborea-B. schickendantzii* = 1,70 (Fig. 5.3A)

d) *B. carrapo-B. schickendantzii* = 1,84 (Fig. 5.3A)

e) Suma = 6,79

f) Promedio = 6,79/4 = 1,69

Si recurrimos a un método que atribuya peso al candidato ingresante, utilizaríamos un método ponderado de la siguiente forma:

a) (*B. arborea-B.carrapo*)-*B. foliosa* = 1,62 (Fig. 5.3B)

b) (*B. arborea-B.carrapo*)-*B. schickendantzii* = 1,77 (Fig. 5.3B)

c) Suma = 3,39

d) Promedio ponderado = 3,39/2 = 1,69

Como puede apreciarse, con el método no ponderado se retorna en cada paso a la MS original, en cambio con el método ponderado se utiliza la matriz derivada inmediatamente anterior.

Retornemos a la segunda matriz derivada, donde se observa que el valor de mayor similitud es el que poseen *B. retama* y *B. chilensis*, dando origen a un tercer grupo con un valor de similitud de 1,01 (Fig. 5.3C).

La tercera matriz derivada considera los promedios de similitud existentes entre cada grupo comparado con los restantes grupos, y cada grupo comparado con las restantes UE. El valor de mayor similitud es el que poseen el par *B. foliosa*-B. *schickendantzii* y *B. bonariensis*. Esto significa que *B. bonariensis* se une a dicho grupo originando un nuevo grupo con un valor de similitud de 1,12 (Fig. 5.3D).

La cuarta matriz derivada contiene los valores de similitud de los grupos *B. arborea-B. carrapo*, *B. chilensis-B. retama* y del grupo *B. bonariensis-B. foliosa-B. schickendantzii* entre sí y las restantes UE. Como se observa, el valor de mayor similitud es de 1,16 entre el grupo *B. chilensis-B. retama* y el grupo *B. bonariensis-B. foliosa-B. schickendantzii* (Fig. 5.3E), ambos se fusionan en un gran grupo.

La quinta matriz derivada considera el valor de similitud entre el grupo *B. arborea-B. carrapo* con respecto al grupo *B. chilensis-B. retama-B. bonariensis-B. foliosa-B. schickendantzii* y de ellos con *B.*

sarmientoi, siendo 1,48 el valor de mayor similitud (Fig. 5.3F). *B. sarmientoi* pasa a formar parte del grupo constituido por *B. chilensis-B. retama-B. bonariensis-B. foliosa-B. schickendantzii*. La sexta y última matriz derivada da un valor de 1,62 (Fig. 5.3G), al que se une el grupo *B. arborea-B. carrapo* y el grupo constituido en el paso anterior.

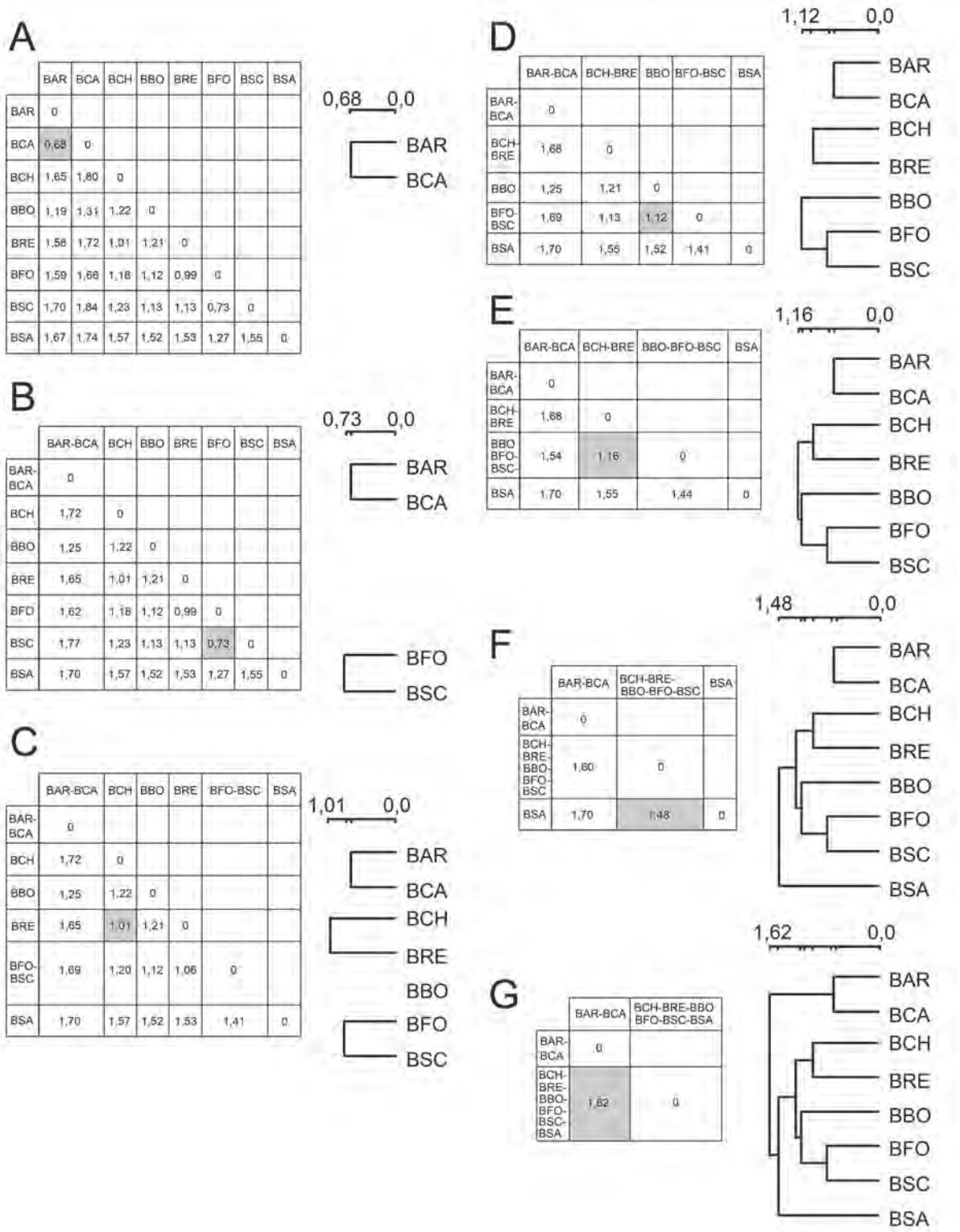


Fig. 5.3. Análisis de agrupamientos (coeficiente de distancia taxonómica, ligamiento promedio no ponderado) sobre la MBD estandarizada de especies de *Bulnesia*. Modificada de Crisci y López Armengol (1983). BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*. Cada letra representa un paso (incorporación de una UE) en la construcción del dendrograma.

Método de Ward

En el análisis multivariado existe otro método de análisis de agrupamientos, el cual puede ser aplicado directamente a la MBD sin pasar por una MS. El método de Ward (1963) se basa en el método de las medias o centroides. Para la formación de grupos este método minimiza la suma del cuadrado del error *E*, que se utiliza en el análisis de la varianza (Legendre y Legendre 1998). *E* funciona como un coeficiente de similitud y es una parte inalterable del método, ya que no se puede elegir otro coeficiente de similitud (Romesburg 1984). Kuiper y Fisher (1975) mostraron que este método recuperaba mejor los agrupamientos generados por simulaciones que los tres tipos de ligamientos.

Básicamente, el método busca minimizar la variación dentro de cada grupo con respecto a los nuevos grupos que se van formando (Romesburg 1984, Murtagh y Legendre 2014). Como otros métodos de agrupamientos aglomerativos, éste sigue una serie de pasos que comienza con tantos grupos como UE se estén analizando. Por lo tanto, la distancia de una UE al centroide (promedio) del grupo con respecto a cada variable original es 0. Así, la suma de todas las distancias también es 0. En cada paso el método de Ward encuentra el par de UE o grupos de UE que minimizan la distancia euclideana entre las UE y los centroides de su grupo correspondiente.

Para la siguiente MBD (Tabla 5.1) de cinco UE × dos variables, en primer lugar se seleccionan dos UE para conformar el primer grupo (por ejemplo, si tomamos como primer grupo las UE1 y 2), mientras que las restantes UE (3, 4 y 5) forman grupos individuales. Para cada variable se calcula la suma de las distancias al cuadrado entre la UE y la media de su respectivo grupo (Fig. 5.4, Tabla 5.2), lo que corresponde a la suma del cuadrado del error. Se repite el procedimiento para cada combinación posible y se selecciona el grupo con menor valor de *E*. Como ejemplo, calculamos el error para la conformación de grupos 12, 3, 4 y 5 (Fig. 5.4A), cuyas medias son 150 (variable 1) y 125 (variable 2) para el grupo 12, 300 (variable 1) y 100 (variable 2) para el grupo 3, 300 (variable 1) y 150 (variable 2) para el grupo 4, y 50 (variable 1) y 100 (variable 2) para el grupo 5. Este procedimiento se repite para todas las combinaciones de grupos posibles y se selecciona aquel agrupamiento que minimiza el error (Fig. 5.4A).

Tabla 5.1. MBD hipotética de cinco UE × dos variables (modificada de Romesburg 1984).

UE	Variable 1	Variable 2
1	100	50
2	200	200
3	300	100
4	300	150
5	50	100

Tabla 5.2. Ejemplo de la aplicación del método de Ward por pasos. Las celdas en gris muestran los grupos seleccionados en cada paso.

Paso	Combinación	Grupo	Posibles combinaciones de subgrupos			Suma del cuadrado del error *E*
1	1	12	3	4	5	16250
	2	13	2	4	5	21250
	3	14	2	3	5	25000
	4	15	2	3	4	2500
	5	23	2	4	5	10000
	6	24	1	3	5	6250
	7	25	1	3	4	16250
	8	34	1	2	5	1250
	9	35	1	2	4	31250
	10	45	1	2	3	32500
2	1	34	12	5		17500
	2	34	15	2		3750
	3	34	25	1		17500
	4	134	2	5		31667
	5	234	1	5		11667
	6	534	1	2		43333,3
3	1	3415	2			56875
	2	34	152			24583,3
	3	234	15			14167

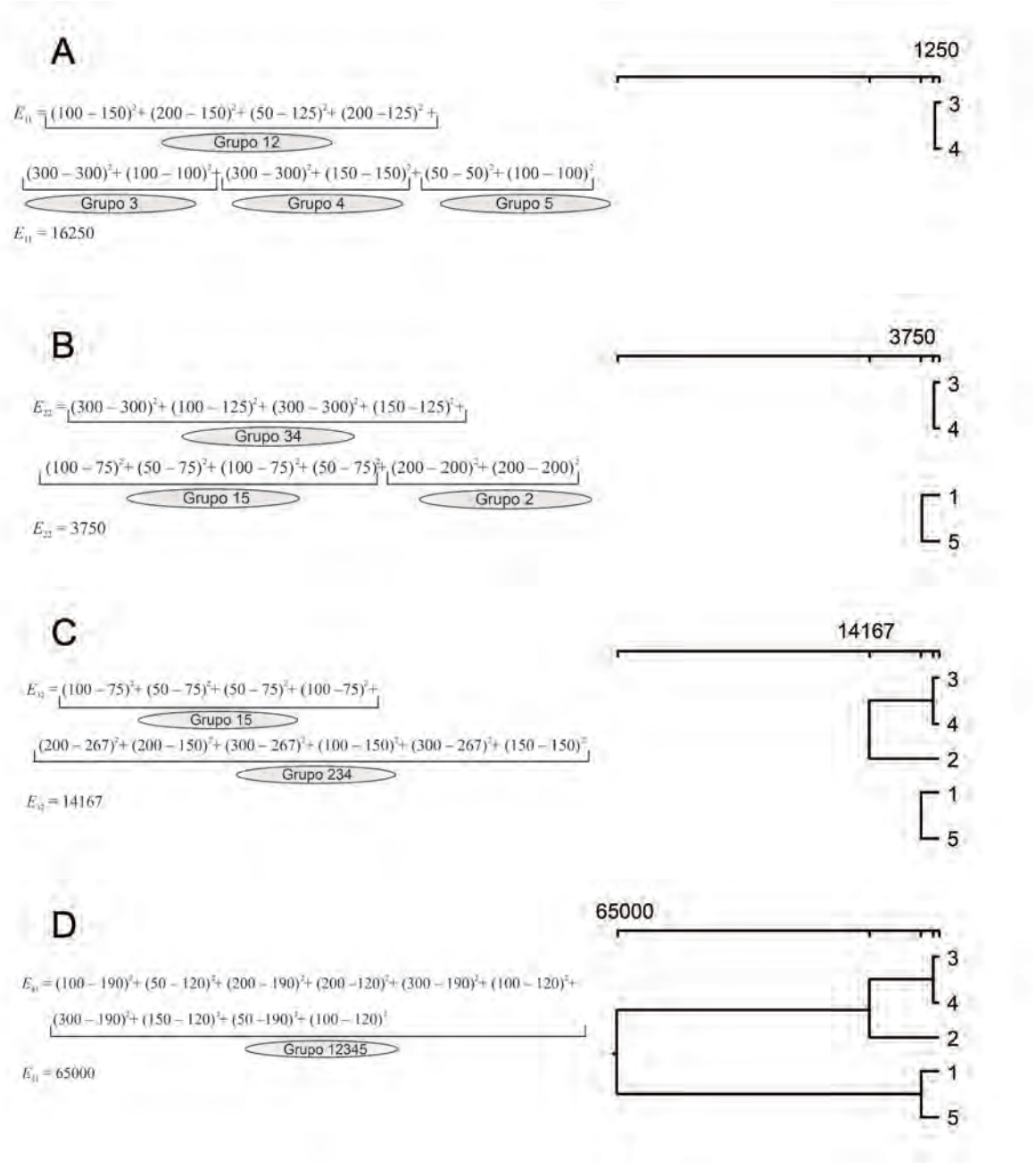


Fig. 5.4. Ejemplos de cálculo del error *E* para cada paso de la Tabla 5.2 y la construcción del dendrograma. (A) Paso 1; (B) paso 2; (C) paso 3; (D) paso 4.

Una vez determinado el primer grupo con menor valor de *E*, se procede a determinar el segundo grupo de la misma forma. En el ejemplo hay dos alternativas (Fig. 5.4B): (1) dejar el grupo 34 como un grupo independiente y generar los grupos 12, 15 ó 25, o (2) incorporar al grupo 34 las UE1, 2 ó 5. Luego de calcular los valores de *E* para cada combinación, se elige formar el grupo 15 debido a que presenta el menor valor de *E* (Fig. 5.4B).

A continuación se sigue el mismo razonamiento que en los pasos anteriores. Se muestra un ejemplo para el paso 3, combinación 3 (Fig. 5.4C). Para finalizar, se forma un grupo que contenga a todas las UE (Fig. 5.4D). En la figura 5.5 se muestra el dendrograma resultante de la aplicación del método de Ward en la MBD de especies de *Bulnesia*.

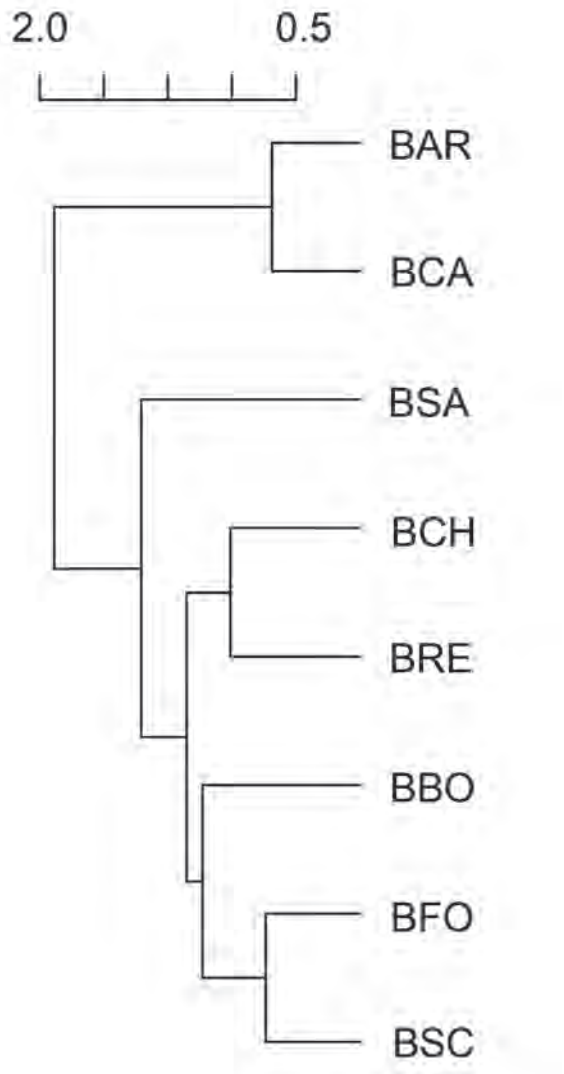


Fig. 5.5. Análisis de agrupamientos basado en el método de Ward sobre la MBD estandarizada de especies de *Bulnesia*. BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*. La escala horizontal representa distancia.

A diferencia del ligamiento simple, los métodos de ligamiento completo, ligamiento promedio y el método de Ward evitan el denominado efecto cadena (Romesburg 1984, du Toit *et al.* 1986). Este efecto se produce en aquellos casos donde hay un incremento progresivo en las distancias entre las UE. En estos casos el ligamiento simple irá incorporando la UE con mayor similitud de a una a la vez, dando como resultado un dendrograma con aspecto de "peine" (Fig. 5.6; du Toit *et al.* 1986, Robidoux y Pritchard 2014).

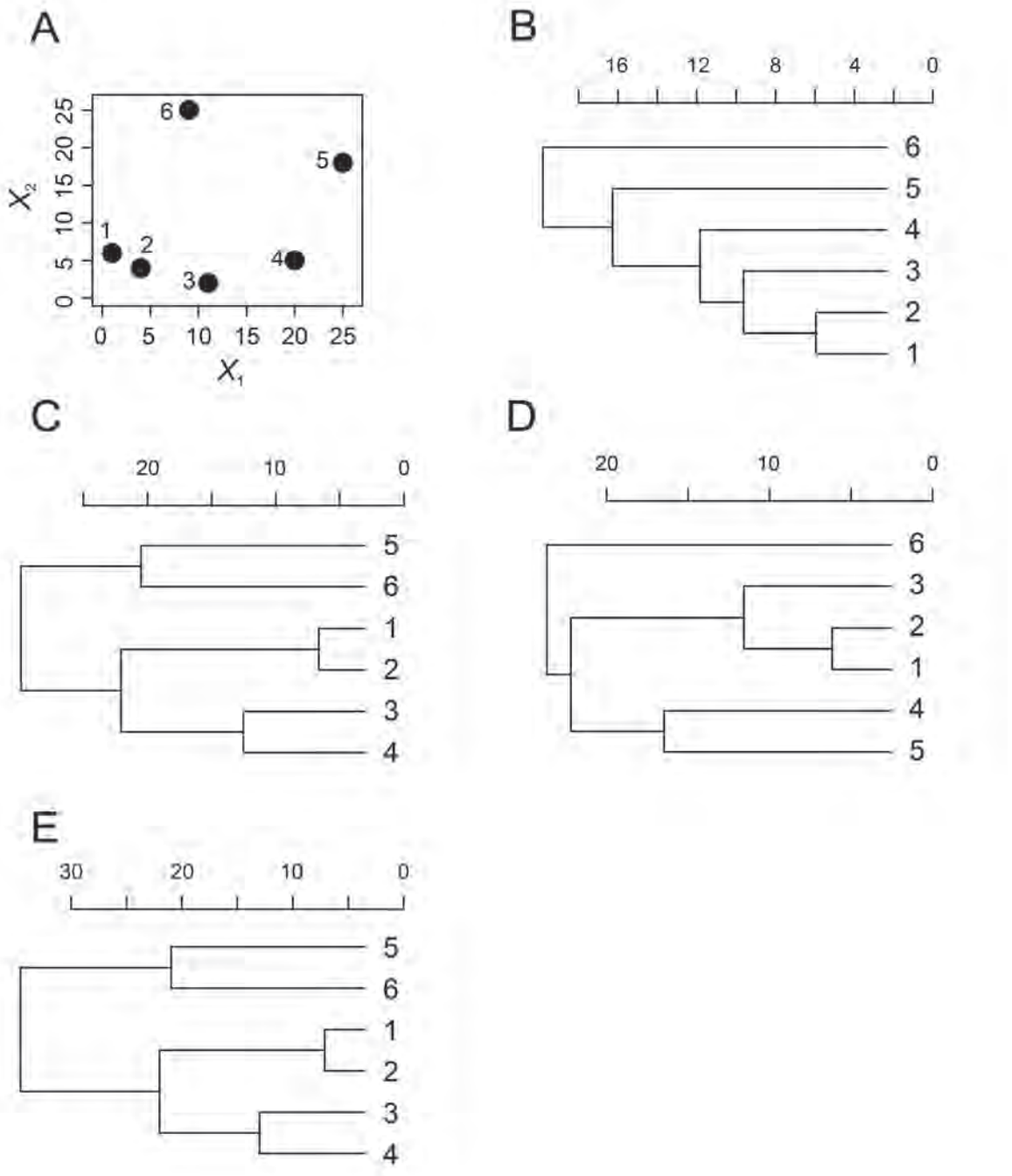


Fig. 5.6. Efecto cadena en la construcción de un dendrograma. Modificada de Romesburg (1984). (A) Se muestran seis UE (1 a 6), dos variables (X_1, X_2) y los dendrogramas generados por (B) ligamiento simple, (C) ligamiento completo, (D) ligamiento promedio y (E) el método de Ward. Las escalas horizontales representan distancias.

Medida de la distorsión

Si se examinan con atención las técnicas de construcción de un dendrograma, se comprenderá que es imposible que el mismo sea un reflejo exacto de la MS. Algunas de las relaciones de similitud serán necesariamente distorsionadas al realizar una representación bidimensional de la matriz. En este sentido, se han propuesto varias técnicas para medir el grado en que el dendrograma representa los valores de la MS. La técnica más conocida es la del coeficiente de correlación cofenética (CCC), que consiste en

construir una nueva MS a partir de los valores del dendrograma, a la que se denomina "matriz cofenética" (Sokal y Rohlf 1962, Farris 1969a, Rohlf 1970). Luego, se calcula el coeficiente de correlación de Pearson entre la MS que dio origen al dendrograma y la matriz cofenética que representa el dendrograma. Una alta correlación entre ellas es señal de escasa distorsión. Generalmente los valores oscilan entre 0,60 y 0,95. Se ha demostrado empíricamente (Sneath y Sokal 1973) que los valores superiores a 0,80 indican una buena representación de la MS por parte del dendrograma y que la técnica del ligamiento promedio es la que origina menor distorsión (Saraçli *et al.* 2013).

Volviendo al ejemplo de *Bulnesia*, tomamos la MS obtenida a partir del coeficiente de distancia y el dendrograma resultante de esa matriz empleando la técnica del ligamiento promedio. El valor de similitud entre *B. arborea* y *B. retama* en la MS es de 1,58, mientras que en el dendrograma esa relación tiene una similitud de 1,62. Por lo tanto, en la matriz cofenética la similitud entre *B. arborea-B. retama* será de 1,62. Si volcamos en la matriz cofenética todos los valores de similitud entre pares de UE tal como están representados en el dendrograma, podremos calcular la distorsión aplicando el CCC entre ambas matrices. La Figura 5.7 muestra las dos matrices y el dendrograma, junto al valor del CCC = 0,90. Los valores del CCC para los restantes dendrogramas se muestran en la Tabla 5.3.

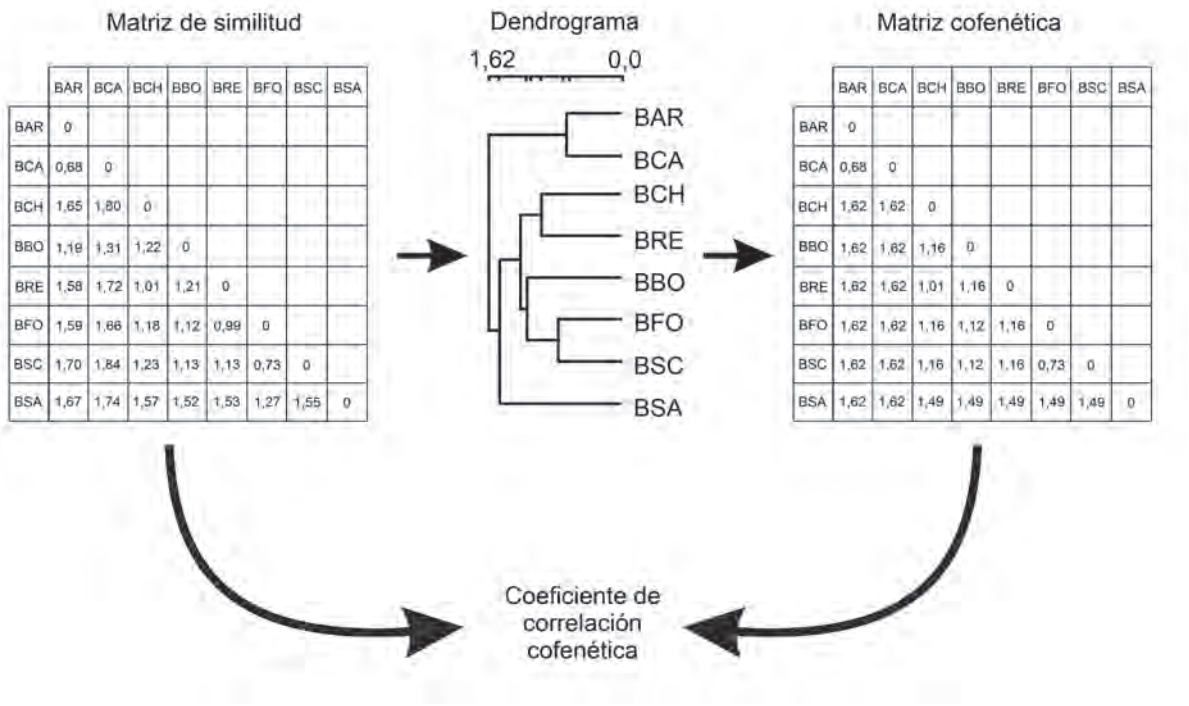


Fig. 5.7. Coeficiente de correlación cofenética. MS de especies de *Bulnesia*, dendrograma y matriz cofenética reconstruida a partir del dendrograma. El coeficiente de correlación cofenética (CCC) corresponde a la correlación de Pearson entre la MS y la matriz cofenética. Modificada de Crisci y López Armengol (1983).

Tabla 5.3. Coeficientes de correlación cofenética (CCC) para los distintos dendrogramas basados en la MBD de especies de *Bulnesia*.

Dendrograma	Técnica	CCC
Fig. 5.1	Ligamiento simple	0,82
Fig. 5.2	Ligamiento completo	0,89
Fig. 5.3	Ligamiento promedio	0,90
Fig. 5.5	Método de Ward	0,87

LAS VARIABLES COMO UNIDADES DE ESTUDIO: MODO R

Como se señaló en el Box 2.2, es posible examinar la MBD asociando las variables en lugar de las UE, considerando a las variables como UE. Esta variante, denominada modo R, se aplica con el propósito de establecer cuáles variables forman grupos y por lo tanto, están altamente correlacionadas, y cuáles variables son relativamente independientes entre sí.

Si tomamos el ejemplo de *Bulnesia* y retornamos a la MBD original (Tabla 2.11), podemos aplicar el coeficiente de correlación de Pearson entre cada par posible de caracteres (columnas en este caso). Obtendremos una MS de 43 variables × 43 variables mediante la aplicación del coeficiente de correlación de Pearson. Sobre la MS resultante aplicamos la técnica del análisis de agrupamientos para construir el dendrograma mediante ligamiento promedio no ponderado (Fig. 5.8).

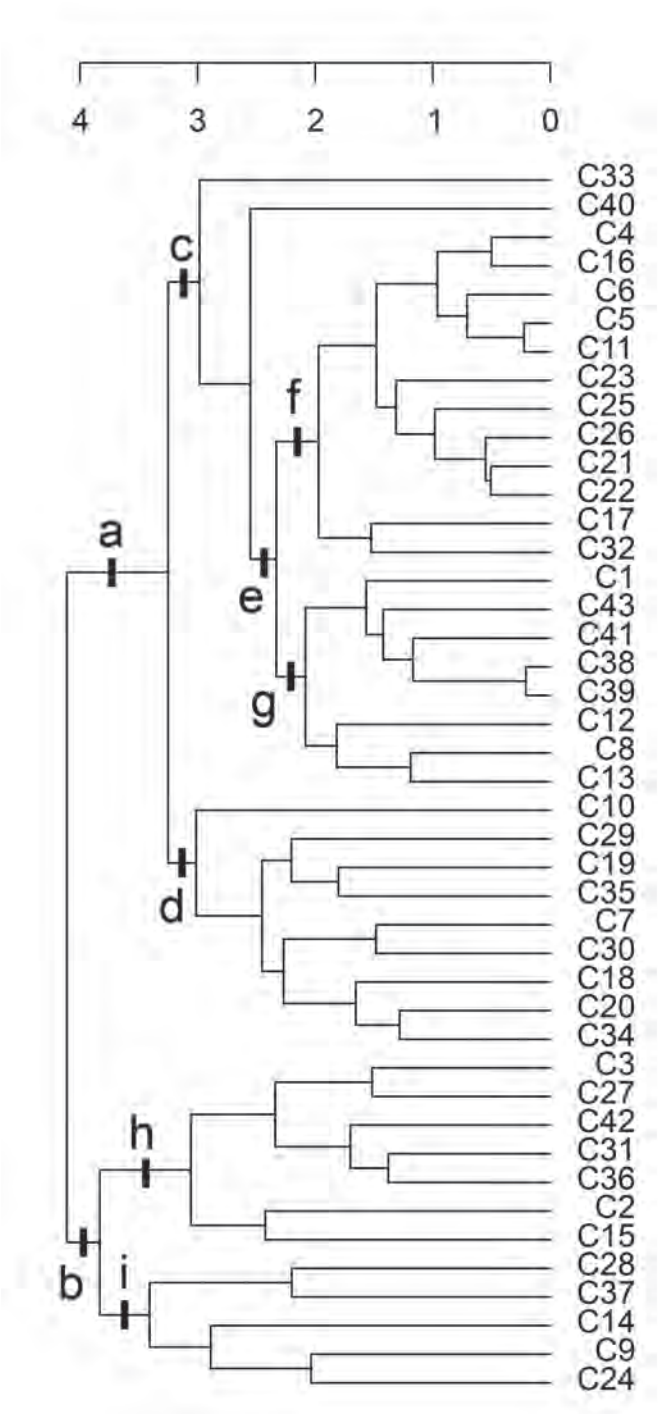


Fig. 5.8. Modo R. Dendrograma sobre la MS (distancia euclideana) utilizando ligamiento promedio aplicado a la MBD estandarizada de especies de *Bulnesia*, donde cada UE representa una variable. Los códigos corresponden a los caracteres de la Tabla 2.8. Las letras representan subgrupos de variables.

INTERPRETACIÓN DEL DENDROGRAMA

La interpretación de un dendrograma es una operación sencilla de realizar. En primer lugar se reconocen visualmente los grandes grupos, es decir los que se han originado a bajos niveles de similitud. Luego, se analizan dichos grupos separándolos en subgrupos hasta llegar a los grupos que presentan la máxima similitud entre las UE. En el dendrograma de la Fig. 5.3 (modo Q, coeficiente de distancia, ligamiento promedio) se reconoce un amplio grupo constituido por *B. retama*, *B. chilensis*, *B. bonariensis*, *B. foliosa*, *B. schickendantzii* y *B. sarmientoi*, y un grupo relativamente aislado del grupo anterior formado por *B. arborea* y *B. carrapo*. Dentro del primero se observa una UE aislada de las demás (*B. sarmientoi*) y dos subgrupos: el primero constituido por *B. chilensis* y *B. retama*, y el segundo por *B. bonariensis*, *B. foliosa* y *B. schickendantzii*. Dentro de este último subconjunto, *B. foliosa* y *B. schickendantzii* se encuentran estrechamente asociados.

Examinemos ahora un dendrograma algo más complicado, como el de la Figura 5.8. Este dendrograma considera a las variables como UE y, a diferencia del dendrograma anterior, no intenta establecer relaciones de similitud entre UE, sino determinar el grado de similitud entre variables. El objetivo es deducir a partir de esa relación, el origen, el valor selectivo y los patrones de variación de los caracteres. A modo de ejemplo se muestra el dendrograma con algunos de los grupos que pueden reconocerse (Fig. 5.8).

A un bajo nivel de similitud se originan dos grandes grupos: *a* y *b*. El grupo *a* se subdivide en *c* y *d*. El grupo *b* se subdivide en *h* e *i*. El grupo *c* contiene dos variables (33 y 40) y al grupo *e* (4, 16, 6, 5, 11, 23, 25, 26, 21, 22, 17, 32, 1, 43, 41, 38, 39, 12, 8, 13). Este último se subdivide en *f* (4, 16, 6, 5, 11, 23, 25, 26, 21, 22, 17, 32) y *g* (1, 43, 41, 38, 39, 12, 8, 13). Se continúa así el análisis de los otros grupos.

Esta interpretación es una descripción del dendrograma y no debe confundirse con el paso final (formulación de las inferencias), en el que entran en juego el juicio del investigador y los conocimientos que éste posee sobre el grupo. En esta etapa final habrá que formular hipótesis sobre las causas biológicas y metodológicas que producen las similitudes encontradas.

¿CÓMO DETERMINAR EL NÚMERO ÓPTIMO DE GRUPOS?

Como se vio anteriormente, un mismo dendrograma puede utilizarse para definir distintos números de grupos y usualmente la decisión depende del investigador. En la práctica se seleccionan aquellos grupos que tienen sentido biológico, por lo que es en gran parte arbitrario y subjetivo. Sin embargo, podemos preguntarnos ¿cuál es el número óptimo de agrupamientos? Este problema no es trivial y ha sido ampliamente discutido en la literatura, de hecho existen más de 30 métodos denominados reglas de detención (*stopping rules*; Sugar y James 2003, Wishart 2005).

Método del codo

Podemos pensar que existe una buena partición cuando los grupos son homogéneos. En términos matemáticos esto significa una baja variación intra-grupo y una alta variación entre grupos. El objetivo es definir grupos de forma tal que el cociente entre la variación intra-grupo (que corresponde a la suma de cuadrados del error) y la variación entre grupos sea mínimo, por lo que la calidad de la partición puede cuantificarse con la variación intra-grupo (Fraley y Raftery 1998). El método del "codo" (*elbow method*) analiza la variación intra-grupo como función de la cantidad de grupos: el número óptimo de grupos es aquel que, al ir subdividiendo los grupos, los subgrupos resultantes no disminuyen de manera significativa la variación intra-grupo.

Al aplicar este método al dendrograma de *Bulnesia* (Fig. 5.3) calculamos primero la variación intra-grupo para todo el dendrograma (primer corte). El segundo corte reconoce dos grupos, en este punto se calcula cuánto disminuye la variación intra-grupo (diferencia entre el primer y segundo corte). El tercer corte reconoce tres grupos, nuevamente se calcula la diferencia en la variación intra-grupo entre este corte y el anterior. Se repite esta serie de pasos hasta que haya tantos grupos como sea posible (número de UE – 1, ya que el último corte es trivial y corresponde a una UE por grupo). Se grafica el número de corte (1, 2, 3, etc.) *vs.*

la disminución en la variación intra-grupo al pasar de un corte al siguiente (Fig. 5.9). El número óptimo de grupos es aquel que genera la mayor disminución de la variación intra-grupo, la que se visualiza como un "codo". Note que la variación intra-grupo siempre disminuye a medida que se generan más grupos. Cuando hay tantos grupos como UE la variación intra-grupo es 0; el objetivo es obtener un número óptimo de grupos con una variación intra-grupo relativamente baja, pero sin el resultado trivial de una UE por agrupamiento.

Hay que tener en cuenta que este "codo" no siempre se logra identificar de forma inequívoca, en particular en aquellos casos donde la variación intra-grupo no disminuye de forma abrupta. Esto es lo que sucede en el caso de *Bulnesia*; quizás tres o cuatro grupos sea el número óptimo de grupos que se pueden identificar (Fig. 5.9).

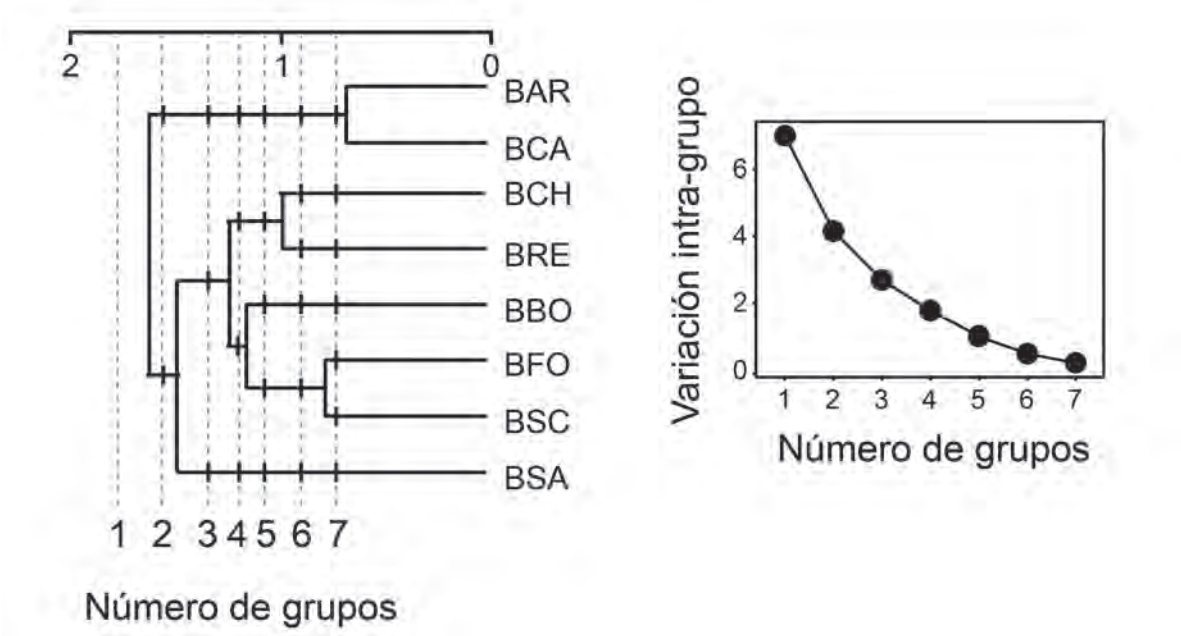


Fig. 5.9. Número óptimo de grupos. Método del "codo" aplicado al dendrograma de especies de *Bulnesia* (coeficiente de distancia taxonómica, ligamiento promedio no ponderado). Las líneas discontinuas muestran el número de grupos que son seleccionados (cortes), y las barras negras indican los grupos que son formados en cada partición. BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*.

TÉCNICAS EXCLUSIVAS, NO JERÁRQUICAS, SIMULTÁNEAS, ITERATIVAS Y SUPERVISADAS

K-medias

En los métodos vistos anteriormente el número *K* de grupos es determinado *a posteriori* por el investigador (no supervisado). En *K*-medias el número de grupos se establece *a priori* del análisis (supervisado). El problema a resolver es el siguiente: dada una MBD, determinar el agrupamiento de las UE en *K* grupos, de forma tal que las UE dentro de cada grupo sean más parecidas entre sí que a las UE de otros grupos (Jain y Dubes 1988, Jain 2010). El método de *K*-medias fue desarrollado independientemente por Steinhaus (1956), Fisher (1958), Ball y Hall (1965), MacQueen (1967) y Lloyd (1982). A diferencia de los métodos anteriores, *K*-medias es un método no jerárquico, no aglomerativo y no divisivo. Este método cumple con dos condiciones simples (Fig. 5.10):

1. El centro de un grupo es la media (centroide) de todas las UE pertenecientes al grupo.
2. Cada UE está más cerca de su propio centroide que de los centroides de otros grupos.

El procedimiento conceptual es el siguiente:

1. Considerar *a priori* una cantidad *K* de grupos.
2. Elegir *K* puntos aleatorios que funcionarán como centroides iniciales de cada grupo.
3. Asignar cada UE a su centroide más cercano. Cada UE pertenece ahora a uno de los *K* grupos.
4. Calcular un nuevo centroide para cada grupo.
5. Repetir los pasos 3 y 4 hasta que la última solución no cambie con respecto a la anterior (convergencia).

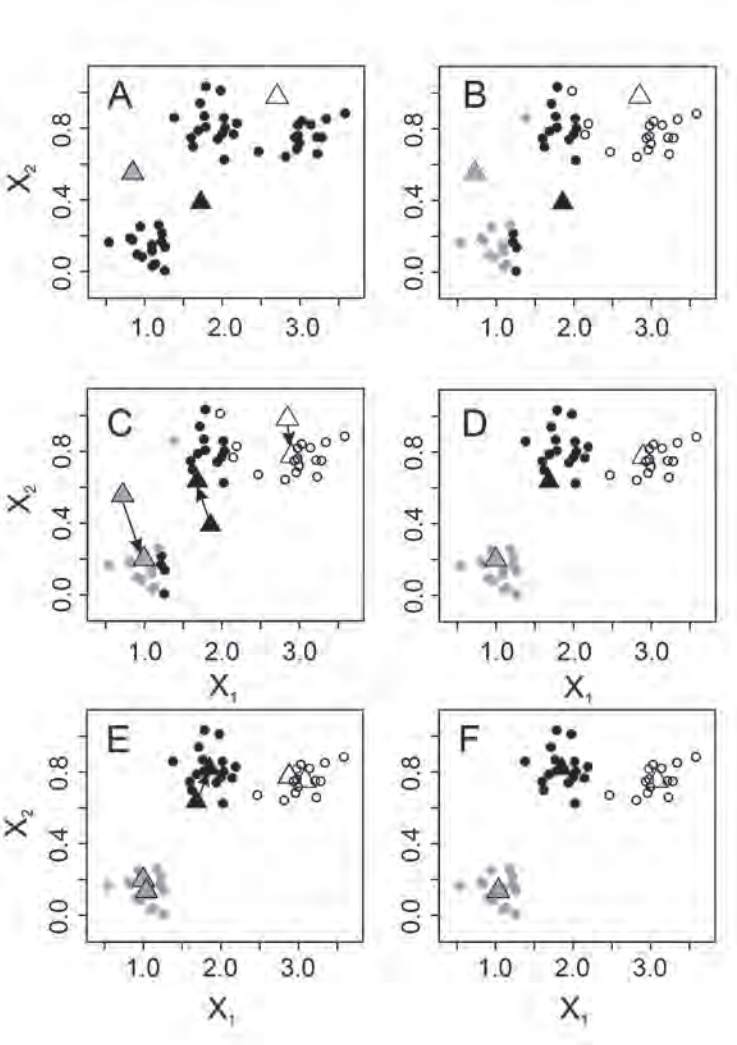


Fig. 5.10. Método de *K*-medias. (A) El primer paso consiste en especificar un número de grupos *K a priori* (en este ejemplo, dos variables X_1, X_2 y $K = 3$) y elegir tantos puntos aleatorios como número de grupos se distingan (representados por triángulos) que funcionarán como centroides iniciales; (B) se calculan las distancias euclideanas entre cada UE y los centroides iniciales, y se asigna cada UE a su centroide más cercano; (C) se calcula un nuevo centroide para cada grupo; (D-F) se vuelve a repetir el procedimiento hasta que el centroide no modifique su posición. Las flechas muestran las trayectorias de los centroides.

Al igual que en el método de Ward, en *K*-medias se busca se busca minimizar la suma de errores al cuadrado. El mayor problema de este método es que la solución depende de la posición inicial de los centroides de cada grupo. Esto no sucede con el método de Ward, que procede iterativamente por aglomeración jerárquica. Por este motivo se recomienda probar con varias configuraciones iniciales. El método de *K*-medias también es sensible a la presencia de UE atípicas (con valores que se alejan mucho del resto de las UE).

K-medias puede aplicarse tanto a la MBD como a la MS. Se recomienda utilizar la MBD cuando la matriz tiene un gran número de UE, dado que la MS se vuelve muy grande. En este caso, se calcula la distancia de cada UE al centroide, en lugar de todas las distancias posibles entre pares de UE. Sin embargo, la desventaja de usar la MBD es que las distancias utilizadas sólo pueden ser euclideanas, lo cual no es apropiado para conteos de especies u otros problemas ecológicos. Cuando la distancia euclideana

no es apropiada, se puede calcular otro coeficiente de similitud y aplicar un análisis de coordenadas principales o un escalado multidimensional no métrico (ver Cap. 6), a través del cual se obtiene una matriz con nuevas coordenadas cartesianas para cada UE. Esta nueva MBD puede utilizarse para construir grupos con *K*-medias.

Si bien el número de grupos es determinado *a priori*, se puede evaluar cuál es el número óptimo de grupos utilizando el método del "codo" visto anteriormente. Como en el caso de los dendrogramas, existe una gran diversidad de métodos para estimar el número óptimo de grupos (Kane 2012, Kodinariya y Makwana 2013). Algunos ejemplos de *K*-medias aplicados a la Biología se pueden ver en Rueda *et al*. (2010), Battey *et al*. (2018) y da Silva *et al*. (2018).

ANÁLISIS DE AGRUPAMIENTOS EN R

El análisis de agrupamientos representa un conjunto de técnicas que se encuentran disponibles con la instalación del programa R (funciones `hclust()` para dendrogramas y `kmeans()` para *K*-medias). Sin embargo, hay algunos paquetes que agregan una gran diversidad de métodos para representar los grupos (Kassambara 2017a), así como también facilidades gráficas, como el paquete cluster (Maechler *et al*. 2018) y FactoMineR (Lê *et al*. 2008), métodos para calcular el número óptimo de grupos, como los paquetes factoextra (Kassambara y Mundt 2017), fpc (Hennig 2018) y NbClust (Charrad *et al*. 2014), y otros que incluso permiten calcular valores de probabilidad para el soporte de los grupos, como el paquete pvclust (Suzuki y Shimodaira 2015). Como ejemplo utilizaremos la MBD de *Bulnesia* (Bulnesia.txt) sobre la cual aplicaremos los cinco métodos de agrupamiento vistos en este capítulo (ligamiento simple, completo, promedio, método de Ward y *K*-medias).

Agrupamiento jerárquico (modo Q)

1. Estandarización de la matriz básica de datos

La función `scale()` permite estandarizar (media 0 y varianza 1) cualquier variable del marco de datos. En nuestro caso debemos recordar excluir la primera columna ya que contiene los nombres de las especies. Se etiquetarán las filas del marco de datos con los nombres de las especies, de lo contrario en el dendrograma las terminales aparecerán referidas con el número de fila (ya que por defecto el programa utiliza números para identificarlas).

```
> Bulnesia <- read.table("C:/R datos/Bulnesia.txt", header = TRUE)
> rownames(Bulnesia) <- Bulnesia$species
> z <- scale(Bulnesia[, -1])
```

2. Cálculo de la matriz de similitud

Utilizaremos la distancia taxonómica, que corresponde al cociente entre la distancia euclideana y la raíz cuadrada del número de variables.

```
> N <- ncol(z)
> S <- dist(z, method = "euclidean")/sqrt(N)
```

3. Selección del método de agrupamiento

Utilizaremos la función `hclust()` y el argumento `method` para indicar qué método aplicar (ligamiento simple, completo, promedio o método de Ward).

```
> clusterS <- hclust(S, method = "single")
> clusterC <- hclust(S, method = "complete")
> clusterA <- hclust(S, method = "average")
> clusterW <- hclust(S, method = "ward.D2")
```

4. Construcción del dendrograma

El gráfico resultante se realiza mediante la función `plot()`. Por ejemplo, si queremos graficar el dendrograma calculado por UPGMA (Fig. 5.11):

```
> plot(clusterA, hang = -1, main = "")
```

El argumento `hang` indica en que posición deben ubicarse las etiquetas en el dendrograma, un valor de -1 alinea las etiquetas entre sí por fuera de la escala del dendrograma. El argumento `main` describe el título del gráfico entre comillas (en este caso ninguno).

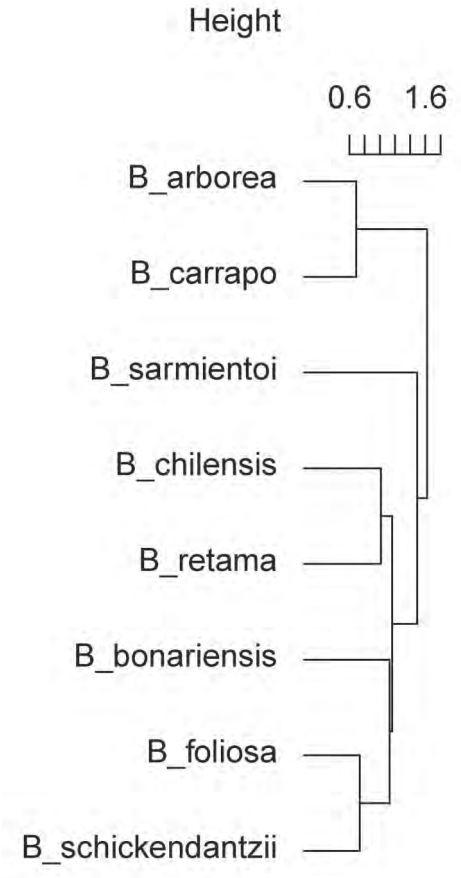


Fig. 5.11. Dendrograma resultante utilizando la MBD de especies de *Bulnesia* (modo Q), función `hclust()` y `method = "average"`.

5. Medida de la distorsión

En primer lugar se calcula la matriz cofenética –función `cophenetic()`–, que resulta de obtener una nueva MS a partir del dendrograma.

```
> mat.cof <- cophenetic(clusterA)
```

Por último, se calcula la correlación de Pearson entre la matriz cofenética y la MS original.

```
> cor(mat.cof, S, method = "pearson")
[1] 0.9043222
```

6. Estimación del número óptimo de grupos

Utilizaremos la función `fviz_nbclust()` del paquete factoextra (Kassambara y Mundt 2017) para realizar un gráfico del número de grupos *vs.* la variación intra-grupo (Fig. 5.12). La función toma como argumentos la MBD (en este caso estandarizada `z`), el método para realizar los agrupamientos (agrupamiento jerárquico `hcut`), la matriz de similitud `S`, el método para calcular el número óptimo de grupos (`wss`, *within sum of squares*) y el número máximo de grupos a considerar (en este caso siete, ya que analizamos ocho UE).

```
> library(factoextra)
> fviz_nbclust(z, FUN = hcut, diss = S, method = "wss", k.max = 7,
+              linecolor = "black") + xlab("Número de grupos") +
+              ylab("Variación intra-grupo"))
```

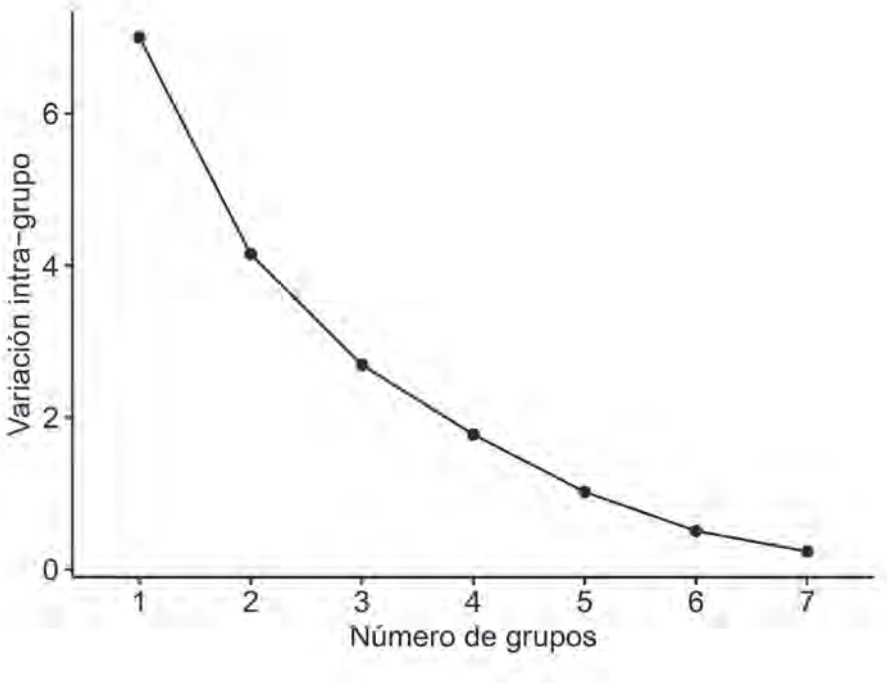


Fig. 5.12. Número de grupos *vs.* variación intra-grupo obtenido con la función `fviz_nbclust()` sobre el agrupamiento jerárquico de la MBD de *Bulnesia*.

En la Figura 5.12 se observa que cuando se pasa de uno a dos grupos y de dos a tres grupos, hay una disminución abrupta en la variación intra-grupo. Sin embargo, cuando se pasa de tres a cuatro grupos esta variación no disminuye demasiado (lo cual es en parte subjetivo). Por lo tanto, podemos decir que el número óptimo de grupos es tres e incluyen a: (1) *B. arborea* y *B. carrapo*, (2) *B. sarmientoi* y (3) *B. chilensis*, *B. retama*, *B. bonariensis*, *B. foliosa* y *B. schickendantzii*. Estos grupos pueden ser visualizados sobre el dendrograma (Fig. 5.13) con la función `fviz_dend()`, indicando el número de grupos (`k = 3`), diferenciando estos últimos con colores (argumento `k_colors`) y añadiéndoles rectángulos (`rect = TRUE y rect_fill = TRUE`).

```
> fviz_dend(clusterA, k = 3, k_colors = c("gray40", "black", "gray70"),
+           rect = TRUE, rect_fill = TRUE, horiz = TRUE,
+           color_labels_by_k = FALSE)
```

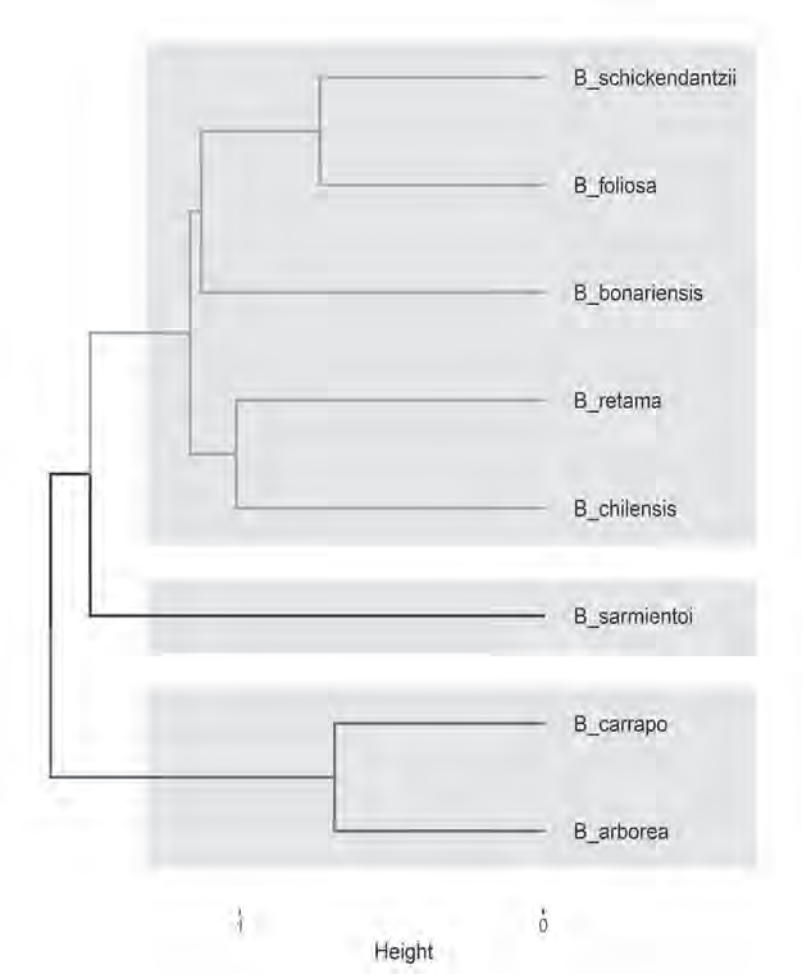


Fig. 5.13. Dendrograma resultante utilizando la MBD de especies de *Bulnesia* (modo Q), función `hclust()` y `method = "average"`, donde se muestra el número óptimo de grupos (tres en este caso).

Agrupamiento jerárquico (modo R)

Si se quiere obtener un dendrograma en el cual se agrupen las variables en lugar de las UE (Fig. 5.14), simplemente se transpone la matriz (se intercambian filas por columnas) con la función `t()` y se realiza el análisis como se describió en el modo Q.

```
> tz <- t(z)
> N <- ncol(tz)
> tS <- dist(tz, method = "euclidean")/sqrt(N)
> clusterA <- hclust(tS, method = "average")
> plot(clusterA, hang = -1, main = "")
```

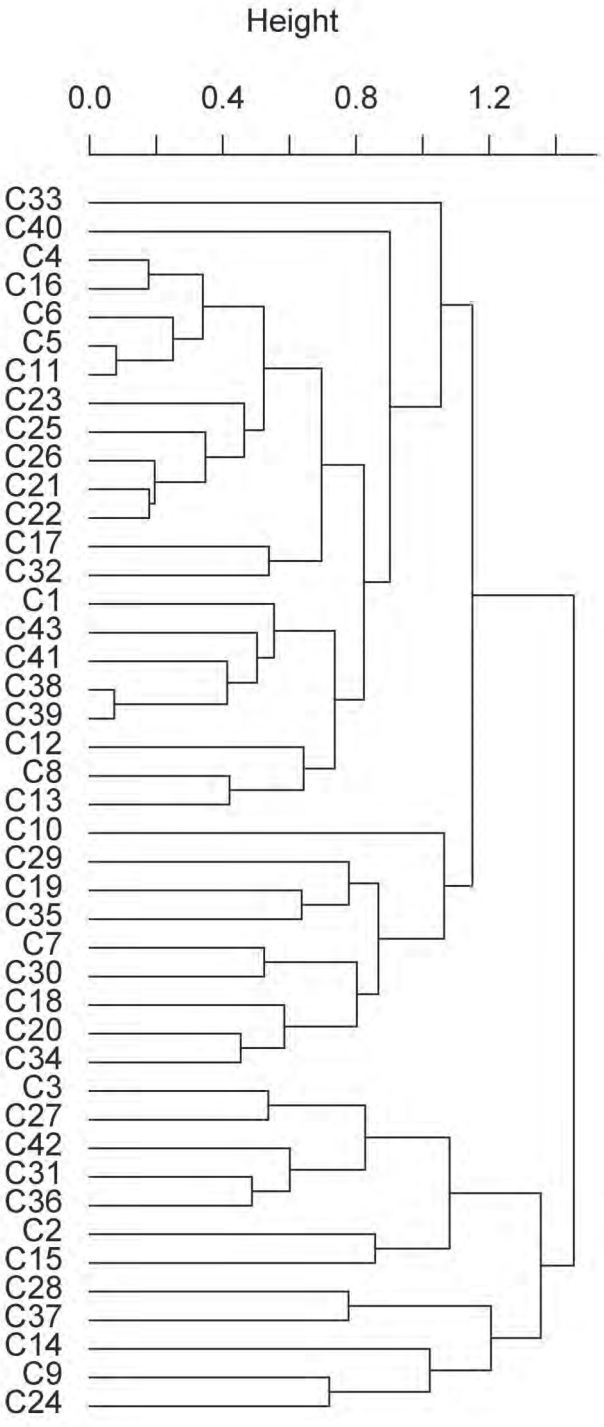


Fig. 5.14. Dendrograma resultante utilizando la MBD de especies de *Bulnesia* (modo R), función `hclust()` y `method = "average"`. Los códigos corresponden a los caracteres de la Tabla 2.8.

Mapa de calor

Los modos Q y R se pueden combinar simultáneamente en lo que se conoce como mapa de calor (*heatmap* o *heat map*). Éste es una representación gráfica de una MBD o de una MS, donde el valor de cada celda se representa con un color. Además, se pueden agregar los dendrogramas de las UE (modo Q) y de las variables (modo R) en sus márgenes (Wilkinson y Friendly 2009). Cada celda tiene un color que representa un valor en la escala de esa variable (Fig. 5.15). Así, el mapa facilita la inspección de la estructura de las UE, las variables y los grupos en conjunto y es muy útil para la visualización de grandes matrices, como por ejemplo las de datos de expresión génica (Weinstein 2008, Schroeder *et al.* 2013,

Gu *et al.* 2016). Representaremos un mapa de calor sobre la base de la MBD de *Bulnesia* con la función `heatmap()`, junto con la leyenda –funciones `rasterImage()` y `text()`–.

```
> heatmap(z, col = rev(heat.colors(256)))
> leyenda <- as.raster(matrix(heat.colors(256), ncol = 1))
> rasterImage(leyenda, xleft = 0.85, xright = 0.9, ybottom = 0.85,
+             ytop = 0.95)
> text(x = 0.92, y = c(0.85, 0.9, 0.95), labels = c(-2, 0, 2))
> text(x = 0.875, y = 1, label = "z", cex = 2)
```

Otras funciones para realizar mapas de calor más estéticos (Kassambara 2017a) son la función `heatmap.2()` del paquete gplots (Warnes *et al.* 2019) y la función `Heatmap()` del paquete ComplexHeatmap (Gu *et al.* 2016).

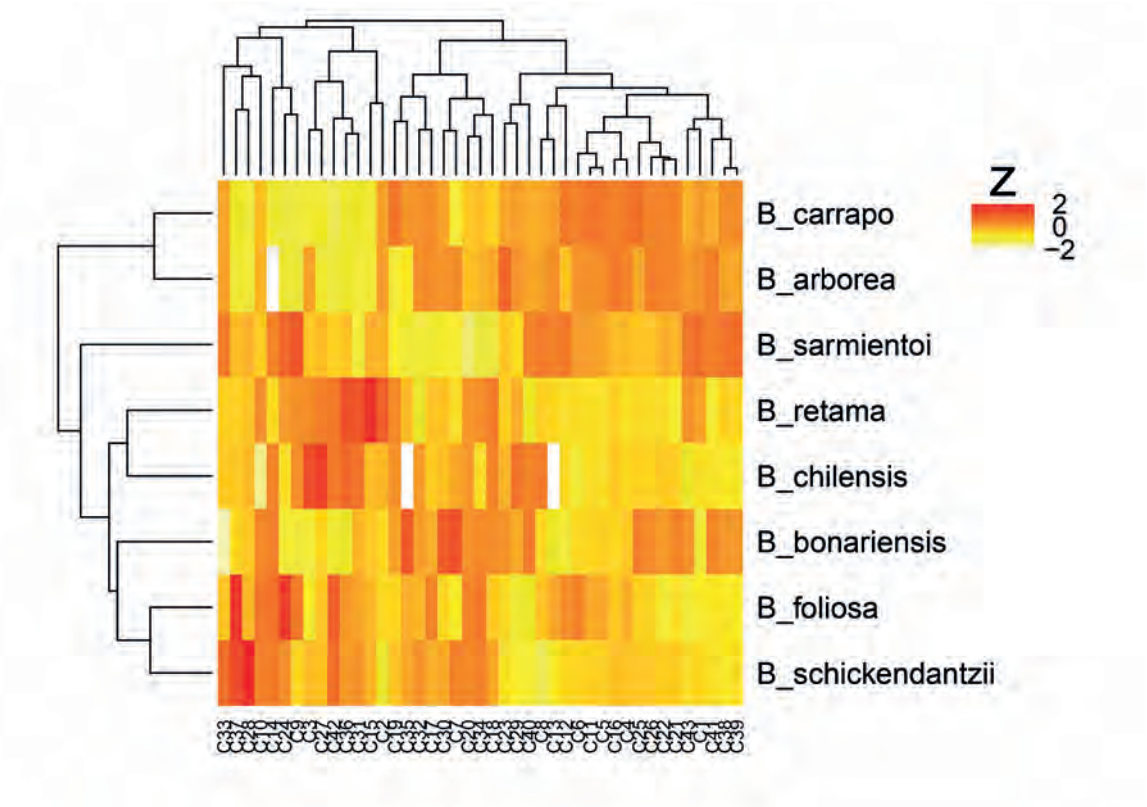


Fig. 5.15. Mapa de calor de la MBD de especies de *Bulnesia*. La leyenda muestra los valores estandarizados de las variables.

A modo de ejemplo, el grupo formado por *B. foliosa* y *B. schickendantzii* tiene valores altos para las variables C33 (presencia de ápice lacinado en la escama estaminal), C37 (pubescencia del fruto) y C28 (presencia de una escama suplementaria junto al estambre) y valores bajos para las variables C41 (longitud del carpóforo), C38 (longitud del fruto) y C39 (ancho del fruto), mientras que el grupo formado por *B. carrapo* y *B. arborea* muestra el patrón opuesto.

K-medias

Utilizando la MBD de *Bulnesia*, vamos a suponer que existen tres grupos de especies (que eventualmente podrían ser tres tribus). El objetivo es asignar cada una de las ocho especies a uno de los tres grupos sobre la base de los caracteres de la matriz.

K-medias se puede realizar con la función `kmeans()`, y para nuestro ejemplo $K = 3$ (argumento `centers`). También consideraremos 1000 configuraciones iniciales de centroides aleatorios dadas por el

argumento `nstart` (recuerde que una sola configuración inicial puede no ser globalmente óptima). El programa luego seleccionará la configuración con la menor variación intra-grupo. Una vez estandarizada la MBD eliminaremos las columnas 13 y 35, dado que tienen algunas celdas con datos faltantes. Otra alternativa es aplicar la función directamente a la MS.

```
> z2 <- z[, -c(13, 35)]
> kmedias <- kmeans(z2, centers = 3, nstart = 1000)
> kmedias
K-means clustering with 3 clusters of sizes 5, 1, 2

Cluster means:
          C1         C2          C3         C4          C5         C6
1 -0.6811087 -0.2945578 0.03224289  -0.5374264 -0.59883897 -0.5752844
2  1.1351812 -0.4021141 -0.69859595 -0.4953048 -0.06711973  0.1079885
3  1.1351812  0.9374515  0.26869075  1.5912183  1.53065728  1.3842168
           C7         C8          C9         C10         C11        C12
1  0.07633863 -0.5040161 -0.03360108 -0.03360108 -0.61530266 -0.6195359
2 -1.37996752  0.8400269  1.84805914 -0.84002688  0.08223852  0.9713258
3  0.49913719  0.8400269 -0.84002688  0.50401613  1.49713738  1.0631769
         C14        C15        C16        C17       C18         C19
1  0.3968627  0.2121320 -0.5400617 -0.1843517  0.263063 -0.03960993
2  0.6614378 -0.3535534 -0.5400617 -1.6985518 -2.180877 -1.36304770
3 -1.3228757 -0.3535534  1.6201852  1.3101551  0.432781  0.78054868
         C20        C21        C22        C23         C24        C25
1  0.3535534 -0.4972123 -0.5433553 -0.4153158  0.05640761 -0.3968627
2 -2.4748737 -0.3643203 -0.2954365 -0.7368505  1.41019019 -0.6614378
3  0.3535534  1.4251909  1.5061065  1.4067147 -0.84611411  1.3228757
         C26        C27        C28        C29         C30        C31
1 -0.4639103  0.3024097  0.2121320 -0.2160247 -0.06767164  0.5336251
2 -0.6822210 -0.5040161 -0.3535534 -1.0801234 -1.64741614 -1.5716355
3  1.5008863 -0.5040161 -0.3535534  1.0801234  0.99288717 -0.5482449
         C32        C33        C34        C36        C37        C38
1 -0.1835326 -0.5040161  0.2352075  0.4118439  0.3240370 -0.6452574
2 -1.4912023  0.8400269 -2.1840699 -0.6864065 -0.5400617  1.1068284
3  1.2044326  0.8400269  0.5040161 -0.6864065 -0.5400617  1.0597293
         C39        C40        C41        C42        C43
1 -0.6535218 -0.4347413 -0.6016092  0.5612486 -0.5350207
2  1.1457064  0.7245688  0.7199875 -0.9354143  1.0094175
3  1.0609514  0.7245688  1.1440292 -0.9354143  0.8328430

Clustering vector:
        B_arborea      B_carrapo       B_chilensis    B_bonariensis
                3              3                 1                1
         B_retama      B_foliosa B_schickendantzii     B_sarmientoi
                1              1                 1                2

Within cluster sum of squares by cluster:
[1] 101.704591  0.000000  9.122475
 (between_SS / total_SS = 61.4 %)
```

```
Available components:

[1] "cluster"      "centers"          "totss"       "withinss"
[5] "tot.withinss" "betweenss"        "size"        "iter"
[9] "ifault"
```

Esta salida da como resultado: (a) una matriz (*Cluster means*) cuyos valores corresponden al promedio de cada variable (estandarizada) por grupo (filas 1 a 3), (b) números que indican a qué grupo pertenece cada especie (*Clustering vector*), el porcentaje de variación explicada por el agrupamiento (*Within cluster sum of squares by cluster*) calculada como variación entre grupos (*between_SS*) sobre variación total (*total_SS*), y (d) los distintos componentes guardados en el objeto `kmedias`. Para identificar puntualmente a qué grupo fue asignada cada especie, escribimos el siguiente comando:

```
> kmedias$cluster
       B_arborea      B_carrapo       B_chilensis      B_bonariensis
               3              3                 1                  1
        B_retama      B_foliosa B_schickendantzii       B_sarmientoi
               1              1                 1                  2
```

Este método da como resultado el mismo agrupamiento que asumimos con el dendrograma mediante el método del "codo" (*B. arborea* y *B. carrapo* forman el grupo 3, *B. sarmientoi* el grupo aislado 2 y las restantes especies forman el grupo 1).

Si bien no es necesario (debido a que por definición el número de grupos se establece *a priori*) algunos usuarios suelen establecer *a posteriori* el número óptimo de grupos (Heikinheimo *et al*. 2007, Walsh *et al*. 2012, Kassambara 2017a). En el ejemplo de *Bulnesia* podemos usar nuevamente la función `fviz_nbclust()`, con la diferencia de aplicar el método *K*-medias, que muestra el número de agrupamientos *vs.* la suma de cuadrados intra-grupo (Fig. 5.16). En este caso no parece tan evidente la presencia de un "codo", quizás cinco o seis grupos sean suficientes.

```
> fviz_nbclust(z2, FUNcluster = kmeans, method = "wss", k.max = 7,
+              linecolor = "black") +  xlab("Número de grupos") +
+              ylab("Variación intra-grupo")
```

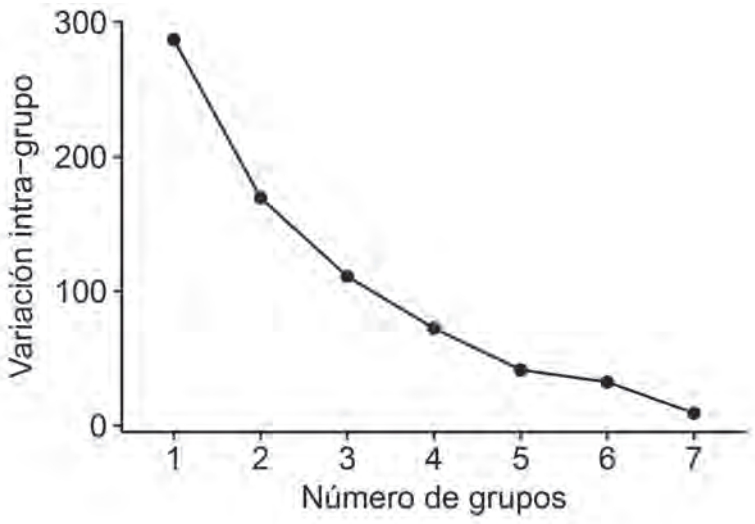


Fig. 5.16. Gráfico del número de grupos *vs.* variación intra-grupo obtenidos con la función `fviz_nbclust()` sobre el método de *K*-medias aplicado a la MBD de *Bulnesia*.

Finalmente ¿cómo podríamos visualizar el resultado de *K*-medias? Una opción sería realizar una figura con las UE y colorearlas de acuerdo al grupo al que pertenecen. El problema es que si la MBD contiene más de tres variables, deberíamos decidir qué variables utilizar. Una solución para reducir el número de dimensiones es utilizar alguna técnica de ordenación, como el análisis de componentes principales (ver Cap. 6), y luego utilizar los primeros componentes para realizar el gráfico (Kassambara 2017a). Para esto, utilizaremos la función `fviz_cluster()` del paquete factoextra (Kassambara y Mundt 2017). Además, agregaremos un gráfico de estrella (Fig. 5.17) que añade segmentos entre las UE y sus respectivos centroides, y une mediante líneas las UE más extremas de cada grupo (argumento `star.plot = TRUE`). Los grupos 1 y 2 contienen dos y una UE, respectivamente, por lo que la estrella sólo se visualiza en el grupo 3. El argumento `repel` se utiliza para evitar el solapamiento entre etiquetas, mientras que el argumento `palette` indica los colores para distinguir los grupos, y el argumento `ggtheme` permite personalizar el fondo del gráfico.

```
> fviz_cluster(kmedias, data = z2, star.plot = TRUE, repel = TRUE,
+              palette = c("black", "black", "black"),
+              ggtheme = theme_minimal())
```

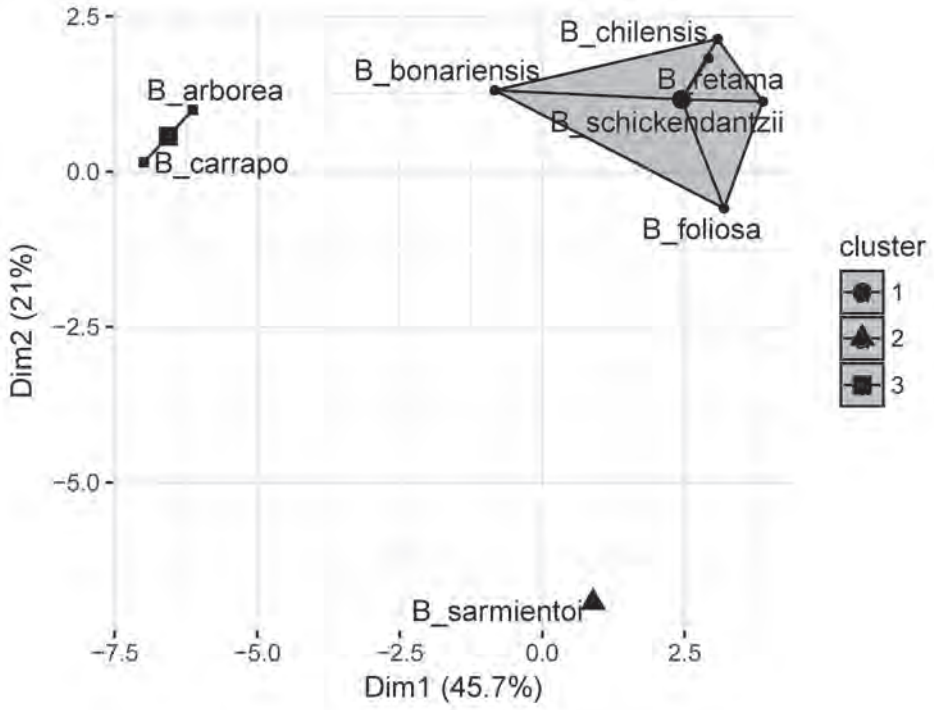


Fig. 5.17. Resultado de *K*-medias mostrado a través de un análisis de componentes principales (ver Cap. 6). Se muestran las especies y los tres grupos junto con sus centroides (círculo, triángulo y cuadrado de mayor tamaño). Los porcentajes de los ejes indican la variación explicada por cada componente.

CAPÍTULO 6
REDUCCIÓN DE DIMENSIONES: MÉTODOS DE ORDENACIÓN

Para encontrar el patrón de relaciones entre la totalidad de las unidades de estudio (UE) se utilizan también los métodos de ordenación. Podemos imaginar un espacio multidimensional donde ubicar a las UE y donde cada dimensión represente una variable, es decir, tantas dimensiones como variables. Los métodos de ordenación reducen, sin gran pérdida de información, el número total de dimensiones (p), y de esa manera facilitan la representación de las UE y sus relaciones en función de las variables empleadas. A diferencia de los análisis de agrupamientos, la mayoría de los métodos de ordenación no trazan límites en el espacio que separen a los grupos, tarea que usualmente corresponde al investigador. Otra diferencia con los análisis de agrupamientos es que estos últimos sólo representan gráficamente a las UE o a las variables, pero no a las dos en simultáneo (excepto en el mapa de calor). En un espacio de ordenación, las relaciones entre las UE están reflejadas en la posición en que se disponen en ese espacio. Si el espacio reducido (menores dimensiones que las del espacio original) refleja en buena medida el espacio de p dimensiones, cuanto más cerca se encuentran entre sí dos UE, más estrechamente relacionadas están.

Se han propuesto numerosas técnicas de ordenación (Gower 1966, Sneath y Sokal 1973, Whittaker 1973, Benzécri 1980, Legendre y Legendre 1998, Quinn y Keough 2002, Greenacre y Primicerio 2014), entre las que mencionaremos: el análisis de componentes principales (PCA), el análisis de correspondencias (CA), el análisis de coordenadas principales (PCoA), análisis discriminante (DA) y el escalado multidimensional no métrico (NMDS). Todos los métodos de ordenación producen una representación de las UE en el espacio euclideano, preservando las distancias originales (cualquiera sea el coeficiente utilizado) de la mejor forma posible.

Las técnicas de ordenación o escalamiento multidimensional pueden clasificarse en escalamiento multidimensional métrico o no métrico (Legendre y Legendre 1998). Los términos escalamiento u ordenación hacen referencia a que las UE pueden ubicarse en un espacio de menores dimensiones que el espacio original (de la matriz), mientras se preservan de la mejor forma posible las relaciones entre las UE. Los términos métrico y no métrico hacen referencia a si la representación final está basada en la medida de distancia original o en el rango de esta medida de distancia, respectivamente. Independientemente del método utilizado, ambos dan siempre como resultado una representación de las relaciones entre las UE en un espacio euclideano. A diferencia de los análisis de agrupamientos vistos en el Capítulo 5 (excepto K-medias) los métodos de ordenación son no jerárquicos.

ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales (PCA) tiene su origen en los trabajos realizados a principios del siglo XX por Pearson (1901), pero fue Hotelling (1933) quien consolidó su uso para representar según un modelo lineal, un conjunto de variables mediante un número reducido de variables hipotéticas, denominadas

componentes principales. Estos componentes no están correlacionados entre sí y por lo tanto, se interpretan independientemente unos de otros, y cada uno de ellos contiene una parte de la variabilidad total la matriz básica de datos (MBD) original. El primer componente (PC1) es el que contiene la mayor variabilidad. De la variabilidad restante, el segundo componente (PC2) es el que incluye más información. El tercer componente (PC3) posee la mayor variabilidad no contenida en los componentes anteriores. Así se continúa hasta que toda la variabilidad ha sido distribuida diferencialmente entre los componentes. Cada componente contiene información de todas las variables pero en diferentes proporciones. Examinaremos en términos generales los principios básicos de esta técnica. Para un examen más detallado y matemáticamente más riguroso, ver Legendre y Legendre (1998), Jolliffe (2002), Abdi y Williams (2010), Jolliffe y Cadima (2016) y Husson *et al.* (2017).

En su concepción más simple, los fundamentos subyacentes del PCA pueden visualizarse con un modelo geométrico constituido por dos variables, es decir, dos dimensiones ($p = 2$). En la Figura 6.1A se han ubicado 10 UE (individuos de una especie de pez) en el espacio bidimensional originado por dos variables (longitud y peso corporal), con una correlación hipotética de 0,70. De la figura se desprende que ambas variables están correlacionadas positivamente, dado que a medida que aumenta la longitud también aumenta el peso. La dispersión de las UE en relación con cualquiera de los dos ejes depende de la escala utilizada. Por lo tanto, es necesario que ambas variables estén expresadas en la misma unidad de medida. Para ello es conveniente estandarizar los caracteres a media 0 y varianza 1 (ésto se logra restando su media y dividiendo por el desvío estandar; ver Box 1.4; Fig. 6.1B). Una consecuencia importante de la estandarización, es que las UE están referidas ahora a un nuevo par de ejes ortogonales (perpendiculares entre sí) que se cortan en un punto que corresponde al promedio de cada variable (Fig. 6.1B). Asimismo, esto permite que el análisis se base sobre la matriz de correlación entre las variables (modo R). Cada UE tendrá un nuevo par de coordenadas definidas en función de unidades de desvío estándar.

El objetivo ahora es hallar aquel eje sobre el cual existe la máxima variación en las UE. Una opción es utilizar la abscisa o la ordenada, pero a simple vista se advierte que ninguno de estos ejes es adecuado, y el eje que cumple con esta condición debe ubicarse entre ambos. Geométricamente, la disposición espacial de las UE para dos variables correlacionadas es la de una nube elíptica (Fig. 6.1C). El eje de mayor variación coincide con el eje mayor de esa elipse y corresponde al primer componente principal (Fig. 6.1C). Si queremos determinar la máxima variación entre las UE en una segunda dimensión perpendicular a la primera, el eje buscado será coincidente con el eje menor de la elipse. Este eje corresponde al segundo componente principal (Fig. 6.1C). Puede observarse que ambos componentes corresponden a una rotación de la ordenada y la abscisa un cierto ángulo θ (Fig. 6.1C). Los vectores que definen la ubicación y la dirección de los ejes mayor y menor se denominan eigenvectores, autovectores o vectores propios, y tienen longitud igual a 1 (Fig. 6.1C). Los eigenvalores, autovalores o valores propios λ_i, reflejan la variación de cada componente principal (mayor en el primero, menor en el segundo, y así sucesivamente). En nuestro ejemplo, los eigenvalores son 1,7 y 0,3 (PC1 y PC2, respectivamente), los eigenvectores para el PC1 son 0,707 (longitud) y 0,707 (peso); estos valores son calculados mediante el uso de software. La suma de todos los eigenvalores constituye la varianza total de la MBD original, y en el caso de una MBD estandarizada corresponde al número total de variables (dos en este caso).

Si analizáramos tres variables, la elipse de la Figura 6.1C se transformaría en un elipsoide y el tercer componente principal estaría representado por el tercer eje del elipsoide, perpendicular a los dos primeros. Si el estudio incluye más de tres variables, se necesitarán dimensiones adicionales cuya representación geométrica no puede ser visualizada, pero de igual forma podrá aplicarse el tratamiento matemático.

Cada componente es una nueva variable hipotética que se construye utilizando todas las variables de la MBD, dado que representa una rotación de los ejes originales. En nuestro ejemplo, cada componente está constituido por la combinación de la longitud y el peso (ambas variables estandarizadas).

$$\text{PC1} = a_{11} \times \text{longitud} + a_{12} \times \text{peso}$$
$$\text{PC2} = a_{21} \times \text{longitud} + a_{22} \times \text{peso}$$

Los coeficientes a_{ij} (i = número de componente, j = número de variable) corresponden a los eigenvectores (Fig. 6.1C) y están sujetos a la condición $a^2_{11} + a^2_{12} = 1$ y $a^2_{21} + a^2_{22} = 1$ (longitud unitaria; Jolliffe y Cadima 2016). Al calcular la suma del producto de los eigenvectores por las variables estandarizadas obtenemos las coordenadas de las UE en el nuevo espacio de ordenación, denominadas *scores*.

Un concepto sumamente importante en el PCA es el concepto de *loading*, definido como el producto de un eigenvector por la raíz cuadrada de su eigenvalor (ver Box 6.1, Fig. 6.1D). Lo que debe recordarse es que el *loading* corresponde al coeficiente de correlación de Pearson (*r*) entre una variable y un componente principal (Abdi y Williams 2010). De esta forma, los *loadings* se expresan como una contribución relativa de cada variable a cada componente. Así, todas las variables contribuyen a todos los componentes pero de manera diferencial; es decir, la variable 1 puede ser un importante aporte para el PC1, pero pobre para el PC2. En nuestro ejemplo, los *loadings* para el PC1 son las correlaciones entre la longitud y el PC1 ($r_{\text{longitud, PC1}}$), y entre el peso y el PC1 ($r_{\text{peso, PC1}}$):

$$r_{\text{longitud, PC1}} = 0,707 \times \sqrt{1,7} = 0,92$$
$$r_{\text{peso, PC1}} = 0,707 \times \sqrt{1,7} = 0,92$$

Box 6.1. Terminología estándar del análisis de componentes principales

Componente principal: variable hipotética que se construye a partir de las variables originales, no se encuentra correlacionada con otros componentes principales.
Eigenvalor: varianza explicada por un componente principal.
Eigenvector: vector que define la dirección de un componente principal.
Loading: se define como $loading = eigenvector\sqrt{eigenvalor}$. El uso del término suele generar confusión, porque algunos autores lo emplean como sinónimo de eigenvector (Abdi y Williams 2010). Esto es una mala práctica, porque el significado de cada uno es diferente. Los eigenvectores definen las direcciones de los componentes, mientras que los *loadings* son eigenvectores que incorporan la información sobre la variabilidad de los datos una vez rotados (eigenvalores). Teniendo en cuenta esto, en el libro adoptamos la primera definición, debido a su interpretación más útil. En este sentido, los *loadings* corresponden al coeficiente de correlación entre una variable y un componente principal y definen las coordenadas en el círculo de correlación unitario. Un *loading* también corresponde al coseno del ángulo entre una variable y un componente. En la práctica, el lector deberá tener cuidado con la definición a la que se hace referencia para una correcta interpretación del análisis.
Score: coordenadas de las UE en el espacio de ordenación.
Círculo de correlación: gráfico en el cual se representan las variables junto con una circunferencia de radio igual a 1 y que permite visualizar la calidad de la representación de las variables. Las coordenadas de las variables corresponden a las correlaciones entre esa variable y un PC (*loadings*). Una variable bien representada se encuentra cercana a la circunferencia. La relación entre variables puede ser interpretada del siguiente modo: (1) las variables correlacionadas positivamente se encuentran cercanas entre sí (ángulo cercano a 0°); (2) las variables correlacionadas negativamente presentan sentidos opuestos (ángulo cercano a 180°); (3) las variables no correlacionadas se encuentran a 90°.
Biplot: gráfico en el cual se representan dos conjuntos de objetos de diferentes formas (UE y variables). Esta superposición es ficticia, dado que las nubes de las UE y las variables no se encuentran en el mismo espacio.
Calidad de la representación: medida que indica qué tan bien proyectada está una UE o variable en el espacio de ordenación. La calidad corresponde al coseno cuadrado del ángulo entre una UE o variable y un componente, lo cual equivale al coeficiente de correlación o *loading* al cuadrado. Varía entre 0 y 1. Para una UE o variable, la suma de todos los valores de calidad a lo largo de todos los PCs es igual a 1.
Contribución: importancia relativa de una UE o variable en la construcción de un determinado PC. La suma de todas las contribuciones de las UE o variables para un mismo PC es igual a 100.

En este caso, al haber sólo dos variables y un único valor de correlación para las dos variables, ambos *loadings* coinciden.

Como se mencionó anteriormente, cada componente corresponde a un eje original (ordenada y abscisa) rotado un cierto ángulo θ, por lo que los componentes también pueden expresarse de forma geométrica como (Sharma 1996):

$$PC1 = \cos\theta \times longitud + sen\theta \times peso$$
$$PC2 = -sen\theta \times longitud + \cos\theta \times peso$$

Para nuestro ejemplo, los eigenvectores toman los siguientes valores (con las variables estandarizadas):

$$PC1 = 0{,}707 \times longitud + 0{,}707 \times peso$$

Al tener sólo dos variables en consideración, ambas contribuyen de igual forma, ya que hay un solo coeficiente de correlación.

Una vez identificados los componentes principales, se debe decidir cuántos deben retenerse para su interpretación, pero al mismo tiempo sin pérdida relevante de información. Debido a que la suma de eigenvalores constituye la varianza total de la MBD original, puede calcularse el porcentaje de variación contenido en cada componente principal según su aporte a esa suma, como el cociente entre el eigenvalor de un componente y la suma total de eigenvalores (en términos matemáticos $100 \times \lambda_i / \Sigma\lambda_i$). Los eigenvalores son 1,7 y 0,3 (PC1 y PC2, respectivamente), por lo que su contribución relativa es $100\% \times 1{,}7 / (1{,}7 + 0{,}3) = 85\%$. Esto significa que el PC1 representa el 85% de la variación total de la MBD, y se considera suficiente retener e interpretar sólo este componente. Esto se debe a que ambas variables están muy correlacionadas y contienen información redundante. Por otra parte, sería trivial mantener ambos componentes, ya que equivaldría a interpretar las dos variables de la MBD original (por definición, utilizar todos los componentes equivale a utilizar todas las variables de la MBD y por lo tanto, acumulan el 100% de la variación total). Por este motivo, en el ejemplo de la Fig. 6.1 proyectamos las UE sólo sobre el PC1 (Fig. 6.1E-F). Como se mencionó anteriormente, las coordenadas de las UE en este espacio reducido se denominan *scores* (que en este caso sólo tienen coordenadas sobre el eje *X*). Así podemos observar en la Figura 6.1F que las UE con valores positivos corresponden a individuos de mayor longitud y peso que las UE con valores negativos. Por lo tanto el PC1 representa una nueva variable que denominamos “tamaño corporal”.

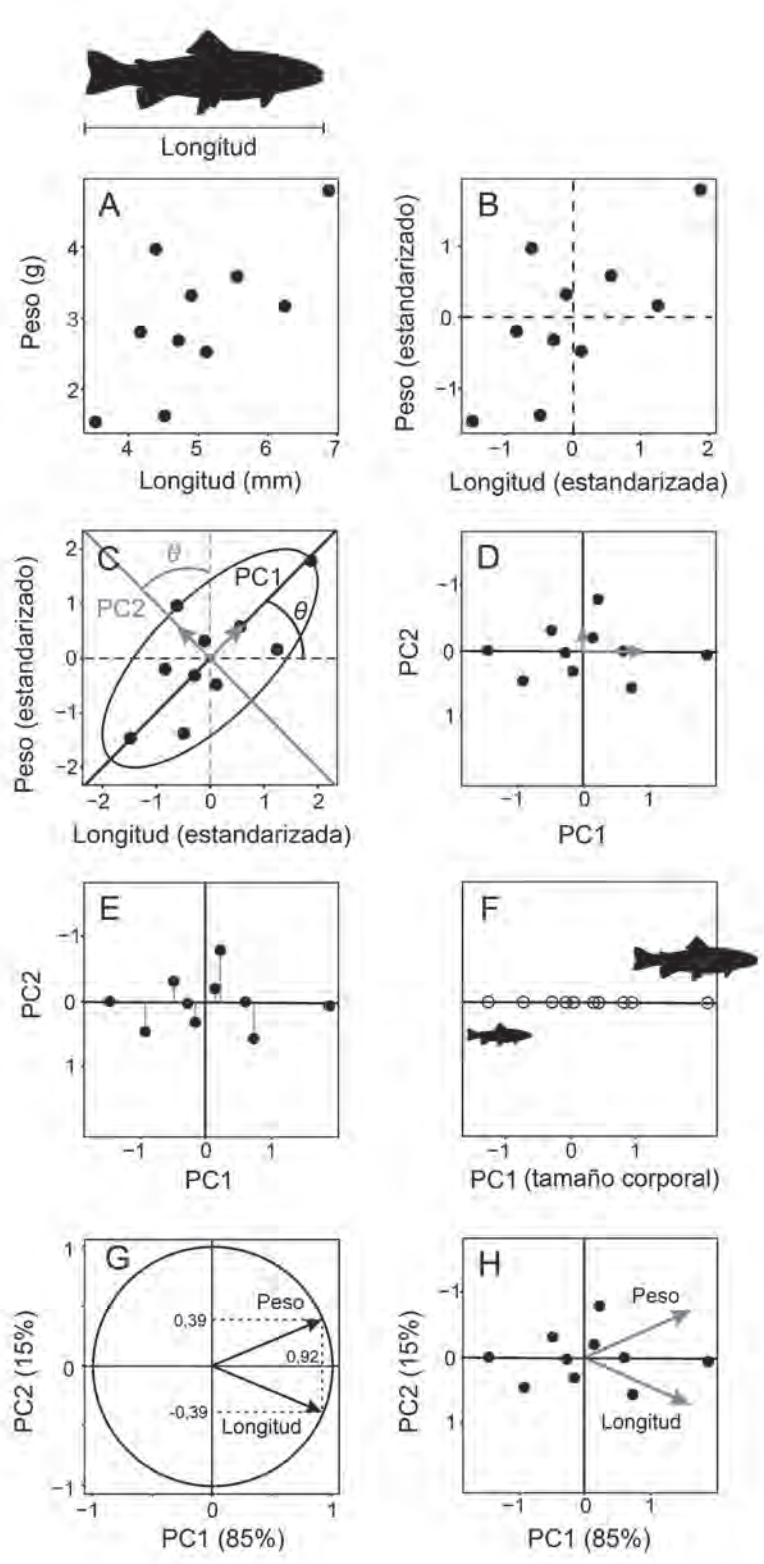


Fig. 6.1. PCA. (A) Gráfico de dispersión de 10 individuos de una especie de pez y dos variables (longitud y peso); (B) estandarización de las variables a media 0 y varianza 1; (C) elipse que engloba las UE y componentes principales (PC) 1 y 2, estos ejes corresponden a la rotación de la ordenada y la abscisa un cierto ángulo θ, los vectores (flechas grises) representan los eigenvectores del PCA cuyas longitudes son 1; (D) resultado del PCA en dos dimensiones, se muestran los eigenvectores multiplicados por la raíz cuadrada de sus eigenvalores (*loadings*); (E) las UE se proyectan sobre el PC1 (líneas perpendiculares); (F) las nuevas coordenadas de las UE proyectadas sobre el PC1 (círculos blancos) se denominan *scores*; (G) círculo de correlación con radio igual a 1, donde se muestran las variables (vectores) y sus coordenadas (*loadings*) que definen la calidad de la representación; (H) *biplot* de UE *vs.* variables.

Representación gráfica e interpretación del análisis de componentes principales

Los resultados del PCA se grafican sobre ejes cartesianos ortogonales que representan los componentes principales, y delimitan un espacio bi- o tridimensional según se utilicen dos o tres ejes, respectivamente. Las UE se sitúan dentro del espacio delimitado por los componentes según los valores de sus coordenadas (*scores*). La selección del número de componentes para representar e interpretar la información contenida en la MBD no es un problema trivial, y existen diversos métodos para seleccionar el número óptimo de componentes (Jackson 1993, Peres-Neto *et al.* 2005). El más simple consiste en tomar un porcentaje acumulado de corte arbitrario. Uno de los más frecuentemente utilizados consiste en retener aquellos componentes que explican más del 70% de la variabilidad total (Jolliffe y Cadima 2016). El criterio de Kaiser-Guttman, por el contrario, considera componentes importantes cuando los eigenvalores son mayores a uno (Guttman 1954, Kaiser 1961, Cliff 1988). Cuando el PCA está estandarizado, la suma de los eigenvalores es igual al número de variables de la MBD, por lo que un eigenvalor > 1 indica que el componente representa más de una variable (Husson *et al*. 2017). Por último, un complemento útil es el gráfico de sedimentación (*screeplot*), que representa cada componente *vs.* sus eigenvalores (Cattell 1966, Cattell y Vogelmann 1977, Jackson 1993). El punto donde se observa una estabilización de la varianza explicada ("codo") indica el número de componentes que se deberían retener (ver en este capítulo "*Técnicas de ordenación en R*").

Los *loadings* (correlación entre variables y componentes) permiten visualizar e interpretar las variables en relación con los componentes, en un círculo de correlación unitario, que muestra la relación entre todas las variables (Fig. 6.1G, Box 6.1). Esta relación puede ser interpretada del siguiente modo: (1) las variables correlacionadas positivamente se encuentran cercanas entre sí (ángulo cercano a 0°); (2) las variables correlacionadas negativamente presentan sentidos opuestos (ángulo cercano a 180°); y (3) las variables no correlacionadas se encuentran a 90° (Kassambara 2017b). En nuestro caso, la longitud y el peso corporal se encuentran correlacionados positivamente entre sí y con el PC1. El peso corporal presenta un valor de correlación de 0,92 con el PC1 y de 0,39 con el PC2, mientras que la longitud presenta un valor de correlación de 0,92 con el PC1 y de -0,39 con el PC2 (Fig. 6.1G). Estos valores corresponden a las coordenadas de las variables y son calculados mediante el uso de software.

Otro aspecto importante para la correcta interpretación de las variables, es analizar en qué medida las mismas están bien proyectadas en el espacio de ordenación, aspecto denominado calidad de la representación (Box 6.1). La calidad corresponde al coseno cuadrado del ángulo entre una variable y un componente, lo cual equivale al coeficiente de correlación o *loading* al cuadrado (para nuestro caso la calidad de la longitud y el peso para el PC1 es $0{,}92^2 = 0{,}85$). Los *loadings* al cuadrado también corresponden al porcentaje de variación explicada de una variable por un determinado componente (Abdi y Williams 2010). Por lo tanto, si es cercano a 1, el componente está bien definido por esa única variable.

En la práctica, la calidad se puede visualizar en el círculo de correlación. Variables con un alto valor de coseno cuadrado están bien representadas sobre el componente principal y se encuentran cercanas a la circunferencia, como en nuestro ejemplo (Fig. 6.1G). Para una variable dada, la suma de todos los cosenos cuadrados para todos los componentes es igual a 1. Esto evidencia que los componentes son ortogonales (independientes en el sentido de que su correlación es 0) y que una variable no puede estar relacionada fuertemente con dos componentes de forma simultánea (Husson *et al.* 2017).

La contribución de una variable a un PC dado se expresa como un porcentaje y se define como el coseno cuadrado de una variable con respecto a la suma total de los cosenos cuadrados de todas las variables para ese componente (Box 6.1); también es equivalente a un eigenvector al cuadrado. Si es cercano al 100%, el componente está bien definido por esa única variable. En otras palabras, las variables que están más correlacionadas con los primeros componentes son las más importantes para explicar la variabilidad total de la MBD. Variables que no se correlacionan con ningún componente o se correlacionan con los últimos componentes, son variables con poca contribución y eventualmente, pueden ser eliminadas para simplificar el análisis (Kassambara 2017b). En nuestro caso, al haber solamente dos variables que comparten una única información dada por el coeficiente de correlación, su contribución a ambos PCs es del 50% ($100\% \times 0{,}707^2 = 50\%$).

Al igual que las variables, las UE tienen valores asociados de calidad de la representación (qué tan bien las UE están proyectadas en el espacio de ordenación) y contribución (en qué porcentaje contribuyen a la ubicación de los componentes). Si bien en Biología se les suelen dar poca importancia, son útiles para identificar posibles valores atípicos que pueden influenciar fuertemente la ubicación de los componentes (Husson *et al.* 2017).

Por último, las UE y las variables pueden visualizarse simultáneamente en un gráfico denominado *biplot* (Fig. 6.1H; Box 6.1). La posición de cada UE en el espacio de componentes principales está dada por sus coordenadas, denominadas *scores*. Sin embargo, las coordenadas de las UE y de las variables no están construidas sobre el mismo espacio (Kassambara 2017b), por lo tanto debemos enfocarnos en la dirección y en el sentido de las variables y no en sus posiciones absolutas sobre el gráfico. Las reglas de interpretación de un *biplot* son las siguientes: (1) las UE cercanas en el espacio tienen características similares en cuanto a sus variables; (2) una UE que está cercana a una variable tiene un alto valor para esa variable; y (3) una UE que se encuentra opuesta a una variable tiene un bajo valor para esa variable. Por ejemplo en la Figura 6.1H se observa que las UE del lado derecho tienen valores altos de longitud y peso corporal.

Las combinaciones más comunes para la elaboración de los gráficos bidimensionales son: PC1 *vs.* PC2, PC1 *vs.* PC3 y PC2 *vs.* PC3 (Fig. 6.2A-C). La posición de uno u otro componente en la abscisa o en la ordenada es indistinta. Los gráficos tridimensionales se basan en los tres primeros componentes (Fig. 6.2D). Las representaciones gráficas deben ir acompañadas de una tabla que contenga la siguiente información acerca del PCA (Abdi y Williams 2010): eigenvalores, porcentaje de variación explicada por cada componente, acumulación de dicho porcentaje y *loadings*. La suma de los porcentajes contenidos en los ejes seleccionados da una idea de la cantidad de variación expresada por la ordenación. Por ejemplo, si el primer componente contiene el 50% de la variación total y el segundo componente 20%, el gráfico bidimensional de estos componentes expresará el 70% de la variación total de la MBD. Una alta variación en los primeros componentes indica que las variables están muy correlacionadas o son redundantes, porque comparten información similar.

Aplicación

A modo de ejemplo, realizamos un PCA aplicado a la MBD de *Bulnesia.* En primer lugar se eliminaron del análisis las variables con valores no disponibles (C13 y C35, ver Tabla 2.8 del Cap. 2). Luego se estandarizó la matriz y se calculó el coeficiente de correlación de Pearson entre variables. De la matriz de correlación resultante (41 variables × 41 variables), se extrajeron los componentes principales y los eigenvalores para los primeros siete componentes. Los tres primeros componentes acumulan más del 80% de la variación total de la MBD. La Tabla 6.1 muestra el eigenvalor de cada componente, el porcentaje de la variación total expresado por los eigenvalores y el porcentaje acumulado a medida que extraemos los componentes.

Tabla 6.1. Tabla resumen del PCA para la MBD de especies de *Bulnesia*. Se muestran los primeros siete componentes principales, sus eigenvalores (λ_i), el porcentaje de variación explicada por cada componente y el porcentaje acumulado.

Componente principal	Eigenvalor	Porcentaje de variación explicada	Porcentaje acumulado
PC1	18,731	45,685	45,685
PC2	8,620	21,024	66,719
PC3	5,556	13,536	80,246
PC4	3,353	8,189	88,414
PC5	2,393	5,847	94,251
PC6	1,455	3,535	97,786
PC7	0,912	2,224	99,999

Dado que los tres primeros componentes expresan más del 80% de la variación, se seleccionan estos para las representaciones gráficas. La Figura 6.2A corresponde al gráfico bidimensional del PC1 *vs.* PC2, y expresa un 45,68% en el primer componente y un 21,02% en el segundo componente, o sea un total del 66,71% de la variación total. La Figura 6.2B muestra el gráfico bidimensional que representa la combinación del PC1 *vs.* PC3, y expresa un 59,21% de la variación total (45,68% del primer componente y 13,53% del tercero). La Figura 6.2C corresponde al gráfico bidimensional del PC2 *vs.* PC3, con un 34,55% de la variación total (21,02% del segundo componente y 13,53% del tercero). La Figura 6.2D corresponde al gráfico tridimensional del PC1, PC2 y PC3, el porcentaje de variación expresado en el gráfico es del 80,24%.

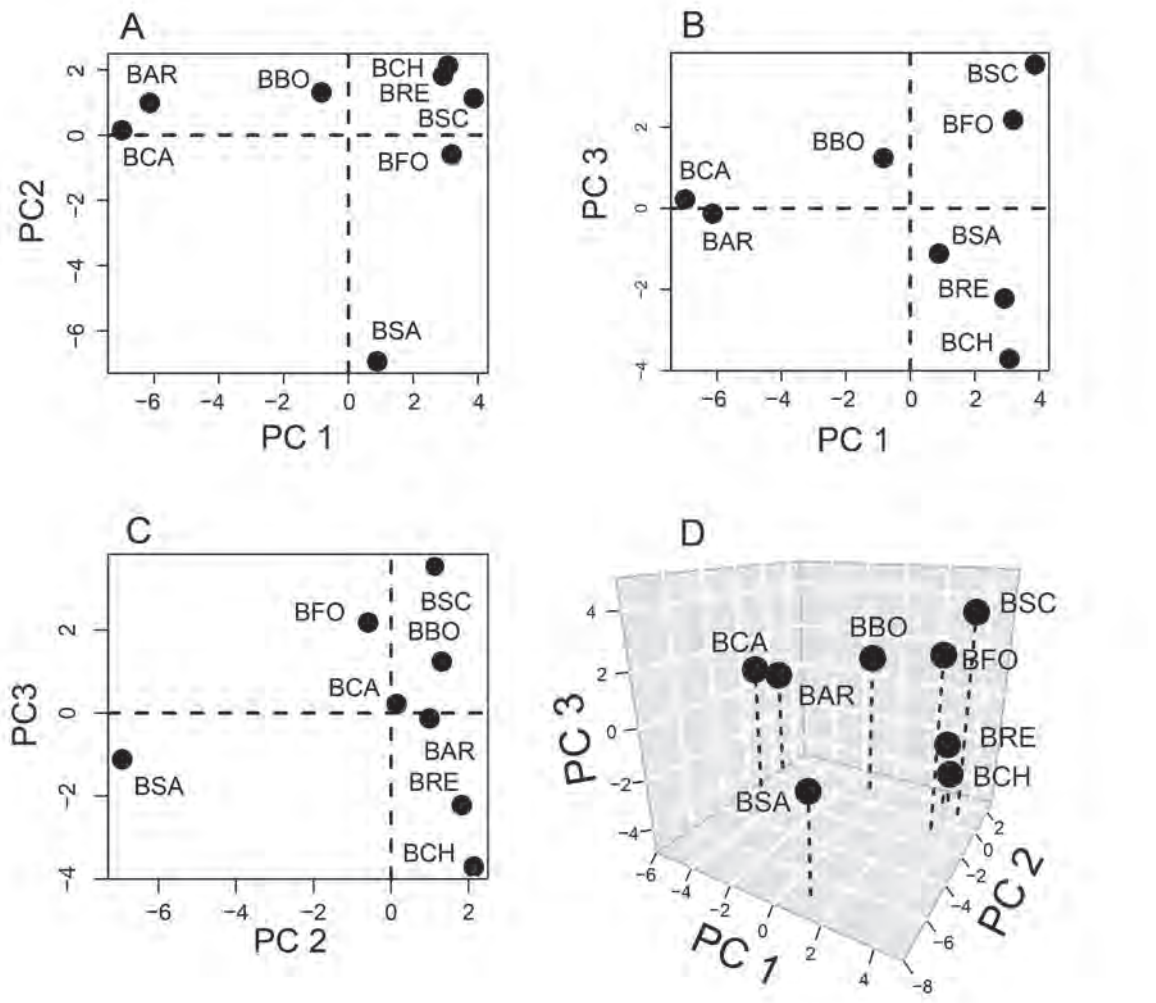


Fig. 6.2. Visualización del PCA aplicado a la MBD de especies de *Bulnesia*. Se grafican los espacios de ordenación bidimensionales del (A) PC1 *vs.* PC2; (B) PC1 *vs.* PC3; (C) PC2 *vs.* PC3; (D) espacio de ordenación tridimensional para el PC1, PC2 y PC3. BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*.

La Tabla 6.2 muestra las correlaciones entre las 41 variables y los tres primeros PCs, dadas por los *loadings*. Cuanto más alto es ese valor (sin importar el signo), mayor es la asociación de esa variable con el componente. El signo indica si la relación es negativa o positiva. En la práctica suelen considerarse variables importantes para un determinado componente cuando los *loadings* son mayores a 0,5, aunque este criterio es completamente arbitrario (Richman 1988, Peres-Neto *et al.* 2003).

Tabla 6.2. *Loadings* de los tres primeros componentes principales (PCs) para cada una de las variables de la MBD de especies de *Bulnesia*. Las variables 13 y 35 fueron excluidas del análisis por presentar valores no disponibles.

Variable	PC1	PC2	PC3
1. Hábito	0,717	–0,48	0,262
2. Longitud del internodio	0,551	0,195	0,562
3. Diámetro del internodio	0,007	0,462	0,759
4. Longitud de la hoja	0,947	0,059	–0,132
5. Ancho de la hoja	0,933	–0,128	–0,099
6. Longitud del pecíolo	0,840	–0,246	–0,193
7. Número de folíolos	0,371	0,611	–0,261
8. Presencia de peciólulos	0,578	–0,314	0,675
9. Disposición de los folíolos en el raquis	–0,487	–0,687	0,488
10. Pubescencia de la hoja	0,313	0,175	–0,696
11. Longitud del folíolo	0,917	–0,186	–0,062
12. Ancho del folíolo	0,674	–0,557	–0,007
14. Posición de los folíolos terminales	–0,713	–0,254	–0,181
15. Presencia de mucrón en folíolos	–0,273	0,251	0,381
16. Tipo de inflorescencia	0,936	0,119	–0,012
17. Longitud del pedúnculo	0,713	0,522	–0,287
18. Longitud del sépalo	0,182	0,955	0,192
19. Ancho del sépalo	0,468	0,547	0,269
20. Color de los pétalos	0,083	0,955	–0,191
21. Longitud del pétalo	0,977	0,071	–0,088
22. Ancho del pétalo	0,990	0,068	0,036
23. Número de nervaduras del pétalo	0,897	0,273	0,133
24. Tipo de estambres	–0,557	–0,665	–0,236
25. Modificación de los estambres	0,914	0,167	–0,124
26. Presencia de gran escama junto al estambre	0,975	0,19	–0,094
27. Presencia de pelos en la base del filamento estaminal	–0,403	0,398	0,786
28. Presencia de una escama suplementaria junto al estambre	–0,361	0,155	–0,607
29. Agrupación de los estambres	0,642	0,508	0,539
30. Longitud del filamento	0,676	0,638	–0,198
31. Longitud de la antera	–0,476	0,768	0,343
32. Longitud de la escama	0,679	0,516	–0,248
33. Presencia de ápice laciniado en la escama estaminal	0,334	–0,391	–0,104
34. Número de carpelos	0,216	0,768	–0,485
36. Número de óvulos por carpelo	–0,620	0,408	0,516
37. Pubescencia del fruto	–0,504	0,056	–0,749
38. Longitud del fruto	0,828	–0,497	0,094
39. Ancho del fruto	0,823	–0,51	0,072
40. Desarrollo del carpóforo	0,637	–0,222	0,41
41. Longitud del carpóforo	0,868	–0,324	0,016
42. Forma de la semilla	–0,807	0,409	0,024
43. Longitud de la semilla	0,677	–0,408	0,083

Por ejemplo, las variables C21 (longitud del pétalo), C22 (ancho del pétalo) y C26 (presencia de gran escama junto al estambre) se encuentran muy correlacionadas positivamente con el PC1. Las variables C18 (longitud del sépalo) y C20 (color de los pétalos) se encuentran muy correlacionadas positivamente con el PC2, mientras que para el PC3, las variables correlacionadas son la C27 (presencia de pelos en la base del filamento estaminal; positivamente) y C37 (pubescencia del fruto; negativamente).

La relación entre variables se puede ver en el círculo de correlación (Fig. 6.3). Por ejemplo, las variables C24 (tipo de estambres) y C9 (disposición de los folíolos en el raquis) se encuentran muy correlacionadas entre sí y de forma positiva, mientras que las variables C24 (tipo de estambres) y C7 (número de folíolos) también se encuentran muy correlacionadas, pero de forma negativa. En cambio, las variables C14 (posición de los folíolos terminales) y C31 (longitud de la antera) están muy poco correlacionadas. El círculo de correlación también muestra la calidad de la representación de las variables (aquellos vectores más cercanos a la circunferencia están mejor representados en el espacio de ordenación; Tabla 6.3). Por ejemplo, las variables C23 (número de nervaduras del pétalo) y C26 (presencia de gran escama junto al estambre) se encuentran bien representadas sobre el PC1, mientras que las variables C18 (longitud del sépalo) y C20 (color de los pétalos) se encuentran bien representadas sobre el PC2. Las variables C15 (presencia de mucrón en folíolos) y C28 (presencia de una escama suplementaria junto al estambre) están pobremente representadas en ambos PCs.

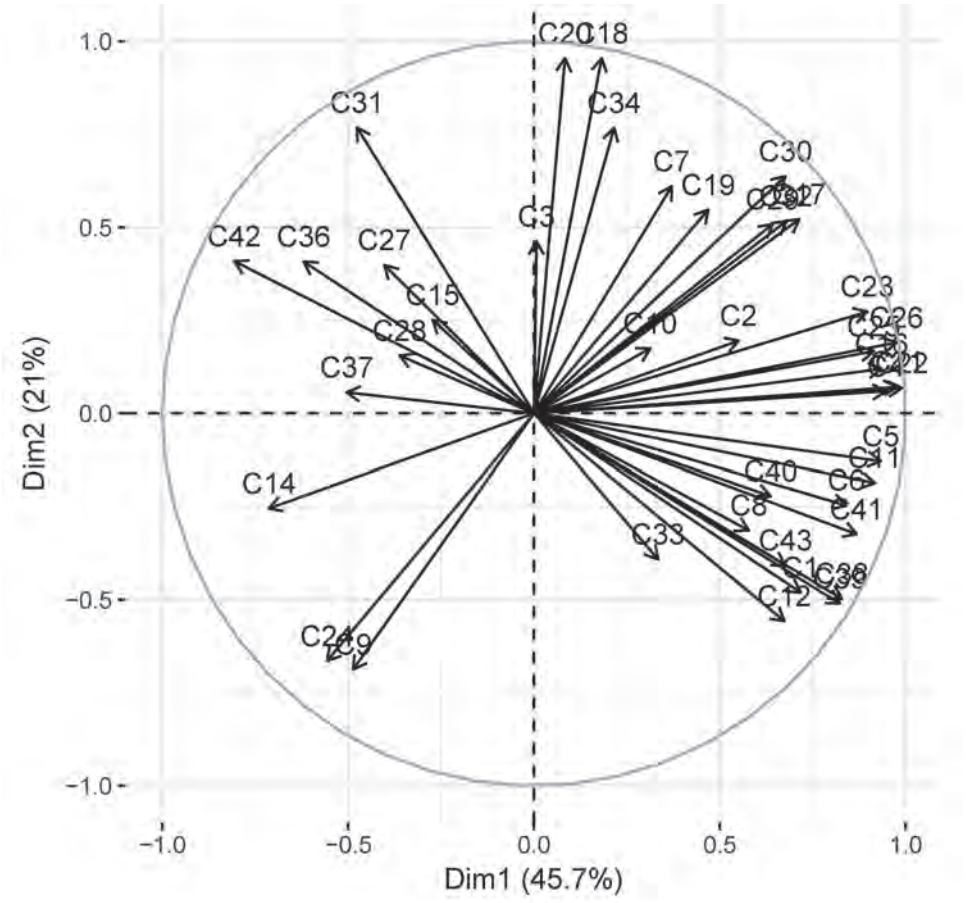


Fig. 6.3. Círculo de correlación unitario aplicado a la MBD de *Bulnesia* que muestra la calidad de la representación de las variables. Los códigos corresponden a los caracteres de la Tabla 2.8.

Tabla 6.3. Calidad de la representación y contribución de las variables de la MBD de *Bulnesia.*

	Calidad			Contribución		
Variable	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
C1	0,515	0,230	0,068	2,748	2,674	1,234
C2	0,303	0,038	0,315	1,620	0,442	5,685
C3	0,000	0,214	0,576	0,000	2,477	10,383
C4	0,896	0,004	0,017	4,785	0,041	0,312
C5	0,870	0,017	0,010	4,647	0,191	0,177
C6	0,706	0,061	0,037	3,771	0,704	0,675
C7	0,137	0,373	0,068	0,733	4,327	1,225
C8	0,334	0,098	0,455	1,782	1,141	8,213
C9	0,237	0,472	0,238	1,265	5,480	4,293
C10	0,098	0,030	0,484	0,522	0,354	8,736
C11	0,842	0,035	0,004	4,494	0,403	0,070
C12	0,454	0,311	0,000	2,423	3,602	0,001
C14	0,509	0,065	0,033	2,715	0,750	0,594
C15	0,074	0,063	0,145	0,397	0,731	2,617
C16	0,875	0,014	0,000	4,674	0,165	0,002
C17	0,509	0,272	0,082	2,717	3,158	1,482
C18	0,033	0,911	0,037	0,177	10,574	0,668
C19	0,219	0,299	0,072	1,172	3,473	1,301
C20	0,007	0,911	0,037	0,037	10,573	0,661
C21	0,954	0,005	0,008	5,094	0,058	0,140
C22	0,979	0,005	0,001	5,228	0,053	0,024
C23	0,804	0,075	0,018	4,292	0,867	0,321
C24	0,310	0,442	0,056	1,657	5,125	1,002
C25	0,836	0,028	0,015	4,465	0,322	0,278
C26	0,952	0,036	0,009	5,080	0,420	0,159
C27	0,162	0,159	0,618	0,865	1,841	11,142
C28	0,131	0,024	0,369	0,698	0,280	6,653
C29	0,412	0,258	0,290	2,199	2,997	5,236
C30	0,458	0,406	0,039	2,443	4,715	0,709
C31	0,227	0,590	0,118	1,211	6,850	2,124
C32	0,461	0,267	0,061	2,460	3,095	1,109
C33	0,111	0,153	0,011	0,595	1,775	0,195
C34	0,046	0,590	0,235	0,248	6,839	4,233
C36	0,384	0,166	0,266	2,050	1,930	4,797
C37	0,254	0,003	0,562	1,354	0,037	10,126
C38	0,686	0,247	0,009	3,663	2,865	0,161
C39	0,677	0,260	0,005	3,613	3,016	0,093
C40	0,405	0,049	0,168	2,165	0,571	3,032
C41	0,754	0,105	0,000	4,023	1,214	0,004
C42	0,651	0,167	0,001	3,473	1,940	0,010
C43	0,458	0,166	0,007	2,445	1,928	0,123

Las relaciones entre las UE se establecen por su proximidad en el espacio delimitado por los componentes: cuanto más próximas se encuentran, más relacionadas están. Esto es válido siempre y cuando la representación en el espacio de menor dimensión capture relativamente bien la información contenida en la MBD original (alto porcentaje de variación explicada). Cabe recordar que el PCA es una reducción de la MBD a un número menor de dimensiones, por lo que las distancias en dos dimensiones son una aproximación a *p* dimensiones. Por lo tanto, hay que tener cuidado con las relaciones entre las UE, ya que sólo serán válidas si los primeros componentes explican mucha variación, indicando que las relaciones se conservan bastante bien en dos dimensiones (lo cual se vincula a la calidad de la representación; Box 6.1). Esta es una diferencia importante con el análisis de agrupamientos, donde las UE se relacionan mediante la distancia en la MBD (aunque con cierta distorsión).

El análisis de la Figura 6.2A muestra estrecha relación entre *B. arborea-B. carrapo*, al igual que el par *B. chilensis-B. retama*. Cercana a este último par se encuentra *B. schickendantzii*. Las demás UE están ligeramente aisladas, con excepción de *B. sarmientoi* que se encuentra sustancialmente alejada de las demás. En la Figura 6.2B se mantienen algunas relaciones con respecto a la Figura 6.2A, pero se observa que aparecen los siguientes cambios: una mayor relación entre *B. schickendantzii* y *B. foliosa* y un acercamiento de *B. sarmientoi* a las UE restantes. En la Figura 6.2D se han integrado las relaciones de las Figuras 6.2A-C en una sola representación.

Al igual que en el caso de las variables, se pueden calcular la calidad de la representación y la contribución de una UE a un determinado componente mediante el uso de software (Tabla 6.4). Por ejemplo, *B. arborea* y *B. carrapo* están bien representadas y son las que más contribuyen al PC1, mientras que *B. sarmientoi* es la única que está bien representada y contribuye más al PC2.

Tabla 6.4. Calidad de la representación y contribución de las UE de la MBD de *Bulnesia.* BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*.

	Calidad			**Contribución**		
UE	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
BAR	0,868	0,023	0,000	28,676	1,625	0,043
BCA	0,904	0,000	0,001	37,261	0,035	0,123
BCH	0,269	0,129	0,391	7,220	7,543	35,467
BBO	0,036	0,090	0,081	0,527	2,838	3,956
BRE	0,286	0,112	0,165	6,506	5,516	12,697
BFO	0,462	0,016	0,215	7,772	0,583	12,223
BSC	0,456	0,039	0,381	11,432	2,108	32,284
BSA	0,016	0,950	0,025	0,607	79,751	3,208

Por último, se puede graficar el *biplot* para evaluar la relación entre las variables y las UE de forma simultánea (Fig. 6.4A). Así, por ejemplo, *B. retama* y *B. chilensis* tienen un alto valor para la variable C36 (número de óvulos por carpelo). *B. arborea* y *B. carrapo* tienen un alto valor para la variable C23 (número de nervaduras del pétalo). *B. sarmientoi* tiene un bajo valor para las variables C3 (diámetro del internodio) y C18 (longitud del sépalo). Para las variables categóricas como la variable C20 (color de los pétalos) un valor alto o bajo depende de la codificación utilizada, en este caso hay dos valores: 1 (blanco) y 2 (amarillo). Al tener un valor bajo significa que la flor de *B. sarmientoi* es de color blanco (Tabla 2.8). Si la variable C20 hubiese sido codificada al revés (1: amarillo y 2: blanco), *B. sarmientoi* tendría un valor alto para esa variable y el vector apuntaría en el sentido de *B. sarmientoi* (Fig. 6.4B).

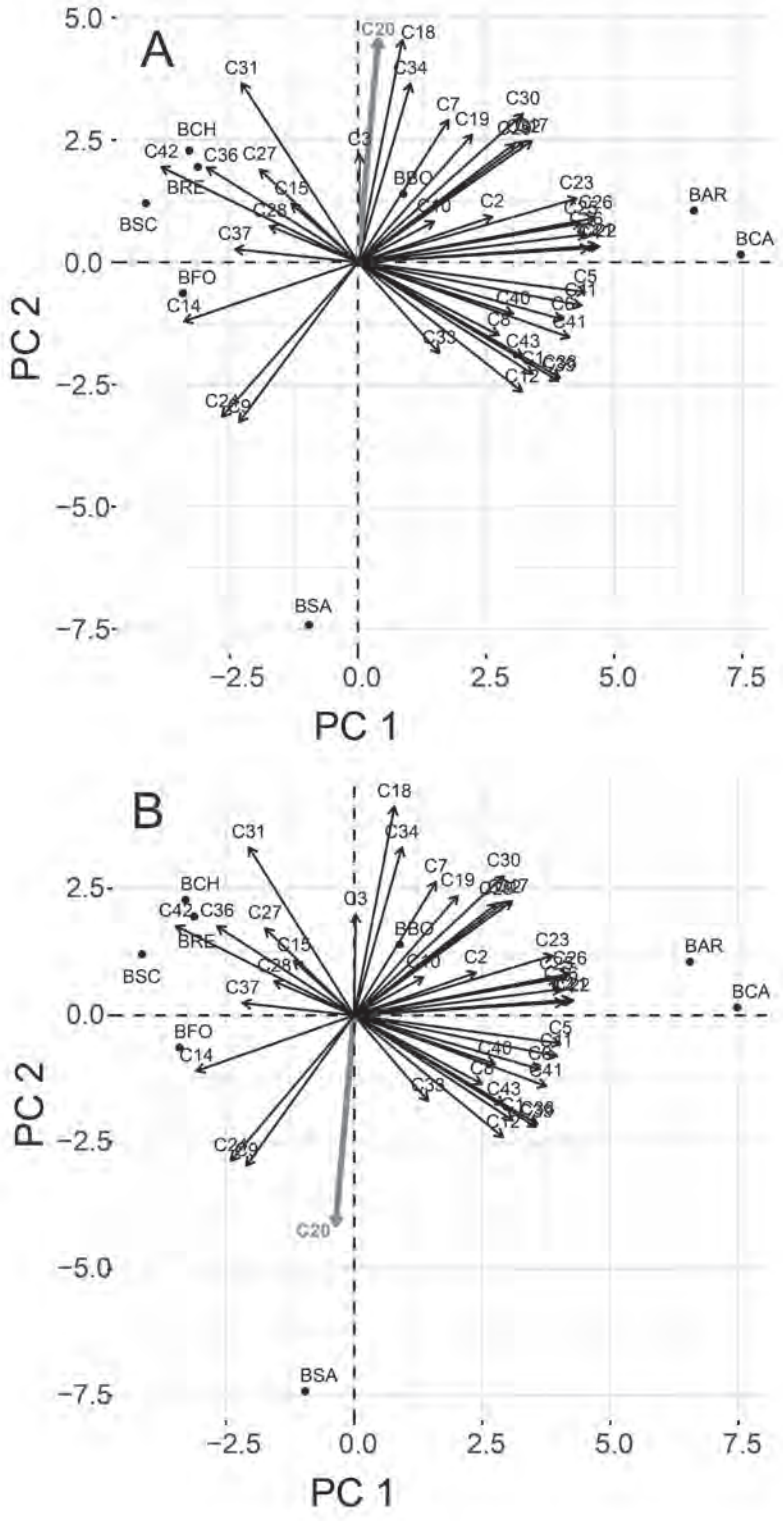


Fig. 6.4. (A) *Biplot* aplicado a la MBD de *Bulnesia,* observe que la ubicación de las UE en el espacio es una imagen especular de la expuesta en la Figura 6.2*;* (B) *biplot* resultado de la recodificación de la variable categórica C20 (color de los pétalos). Los códigos de las variables corresponden a los caracteres de la Tabla 2.8, BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*.

Observando en conjunto las Figuras 6.2 a 6.4 y las Tablas 6.2 a 6.4 podemos concluir que el PC1 es útil para distinguir los grupos conformados por *B. arborea-B. carrapo* de *B. chilensis-B. retama-B. foliosa-B. schickendantzii*, y las variables que más contribuyen a este componente son: C4 (longitud de la hoja), C5 (ancho de la hoja), C21 (longitud del pétalo), C22 (ancho del pétalo) y C26 (presencia de gran escama junto al estambre). El PC2 es útil para distinguir entre *B. sarmientoi* y las restantes UE y las variables que más contribuyen a este componente son: C18 (longitud del sépalo), C20 (color de los pétalos), C34 (número de carpelos) y C9 (disposición de los folíolos en el raquis). El PC3 permite distinguir entre *B. chilensis*, *B. retama* y *B. schickendantzii*, y las variables que más contribuyen a este componente son: C3 (diámetro del internodio), C27 (presencia de pelos en la base del filamento estaminal), C37 (pubescencia del fruto), C10 (pubescencia de la hoja) y C8 (presencia de peciólulos).

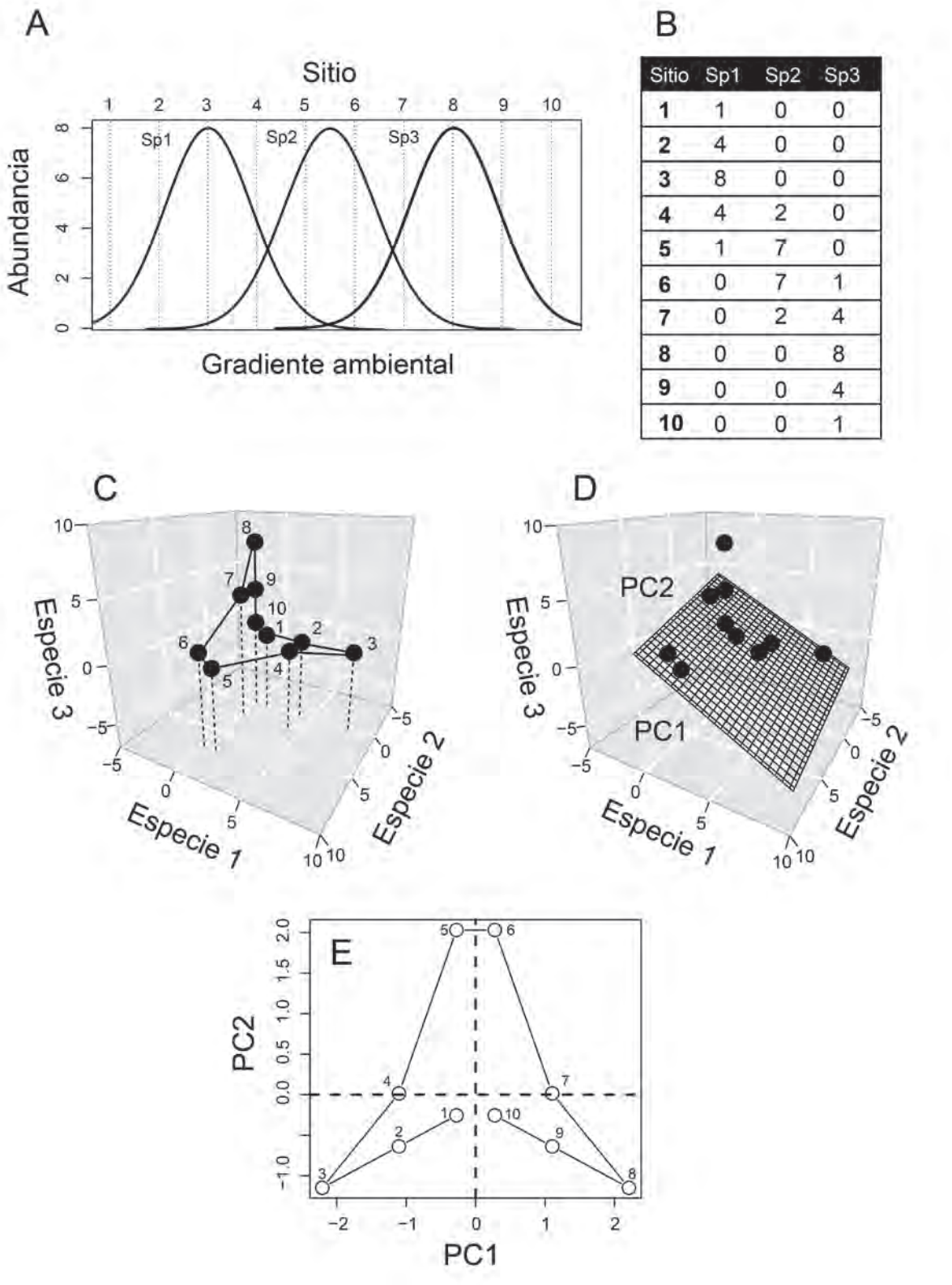
Relación entre el número de variables, el número de unidades de estudio y la reducción de dimensiones

Cuando hay más UE (n) que variables (p), el número máximo de componentes es igual a p; cuando hay igual o menor número de UE que de variables, el número máximo de componentes es $n - 1$ (Legendre y Legendre 1998, Jolliffe y Cadima 2016), considere por ejemplo 10 UE y dos variables. Las UE pueden representarse en un espacio bidimensional, sin importar cuántas sean. Considere ahora el caso de dos UE y dos variables, las dos UE sólo requieren una dimensión para ser representadas (un componente principal que cruza a ambas UE). Un cociente p / n grande causa un aumento exagerado de los eigenvalores de los primeros PCs en comparación con los eigenvalores de los últimos PCs (teorema de Marchenko-Pastur, Marchenko y Pastur 1967), dando como resultado una impresión errónea de la importancia relativa de los PCs (Bookstein 2017).

Efecto arco

Un artefacto (distorsión causada por el modelo subyacente) característico del PCA (y también del análisis de correspondencias y coordenadas principales) aplicado a datos de comunidades (sitios × especies) es el denominado "efecto arco" (Gauch 1982) o "efecto herradura" (Kendall 1971), particularmente visible cuando la composición de especies cambia progresivamente a lo largo de un gradiente ambiental (Gauch 1982, Legendre y Legendre 1998). Esto se produce porque las abundancias de la mayoría de las especies alcanzan un máximo en algún punto del gradiente ambiental, y declinan hacia los extremos del gradiente produciendo una curva de respuesta unimodal (modelo Gaussiano de estructura de comunidades; Gauch 1982, ter Braak 1985). En el caso más dramático se observa que los extremos del gradiente involucionan a lo largo del primer componente, como se muestra en la Figura 6.5.

Un ejemplo de estructura de comunidades con curvas de respuesta Gaussianas se muestra en la Figura 6.5A, en la que se grafican tres especies a lo largo de un gradiente ambiental y 10 sitios que se ubican a lo largo de éste. Los datos de abundancia se utilizan para construir la MBD (Fig. 6.5B), ubicar los sitios en el espacio de las especies (Fig. 6.5C) y generar los componentes (Fig. 6.5D). Luego, las UE se proyectan sobre las dos primeras dimensiones del PCA (Fig. 6.5E). Observe que la configuración bidimensional de sitios a lo largo del gradiente ha sido recuperada por el PCA, pero distorsionada en forma de herradura (Fig. 6.5E). Esta distorsión se debe a que la distancia euclideana considera los sitios extremos muy similares entre sí (distancias pequeñas debido a la presencia de dobles ceros, por ejemplo sitios 1 y 10). Algunos autores consideran que este efecto es un resultado esperado del análisis como producto de un alto reemplazo de especies a lo largo del gradiente, no una anomalía, y por lo tanto debe considerarse válido (Wartenberg *et al.* 1987, James y McCulloch 1990).



Sitio	Sp1	Sp2	Sp3
1	1	0	0
2	4	0	0
3	8	0	0
4	4	2	0
5	1	7	0
6	0	7	1
7	0	2	4
8	0	0	8
9	0	0	4
10	0	0	1

Fig. 6.5. (A) Respuesta unimodal de un modelo de comunidades a lo largo de un gradiente ambiental hipotético; (B) MBD; (C) representación de los sitios (1 a 10) en el espacio multivariado de tres especies; (D) representación de los sitios en el espacio multivariado junto con el plano (PC1 y PC2) sobre el que se proyectan las UE; (E) PCA de los sitios en dos dimensiones. Modificada de Legendre y Legendre (1998).

ANÁLISIS DE CORRESPONDENCIAS

Muchos tipos de datos biológicos son recolectados como conteos (por ejemplo, número de individuos), u otras cantidades no negativas como biomasa, cobertura o porcentaje de ítems en la dieta. El análisis de correspondencias (CA) analiza las diferencias entre cantidades relativas: por ejemplo, si en un sitio hay una abundancia total de 232 individuos considerando todas las especies, y una especie particular tiene una abundancia de 46 individuos, entonces lo que es relevante para el análisis es su abundancia relativa 46/232 (19,8%) y no su valor absoluto. Para utilizar el CA todos los datos tienen que estar medidos en la misma escala, por lo que cobra sentido utilizar sumas de filas y columnas.

El CA fue desarrollado por varios autores de forma independiente. Fue propuesto para el análisis de tablas de contingencia por Hirschfeld (1935), Fisher (1940) y Benzécri (1969). La aplicación más común en datos en Biología es el análisis de ocurrencias de especies (presencia-ausencia o abundancia) en diferentes sitios de muestreo (localidades, cuadrículas). Al igual que el PCA, el CA preserva la distancia euclideana entre UE y variables en un espacio de ejes (o componentes) principales. Esto equivale a preservar la distancia chi-cuadrado entre las filas y columnas de una tabla de contingencia. Para medir las diferencias entre las UE, se utiliza la distancia chi-cuadrado en lugar de la distancia euclideana (como en el PCA) con el fin de estandarizar las especies con diferentes abundancias totales (ver Cap. 4). Otra diferencia importante con el PCA es que cada sitio se pondera proporcionalmente a su abundancia total (232 en el ejemplo), de forma que muestras con mayor abundancia total tienen más peso en el análisis (en el PCA los pesos son iguales para todas las UE). El CA tiene la propiedad de que el análisis de las UE (modo Q) es equivalente al análisis de las variables (modo R).

Box 6.2. Terminología estándar del análisis de correspondencias

Inercia total: varianza explicada por la MBD. A menor inercia, menor es la asociación o correlación entre las filas y las columnas, y viceversa. La inercia máxima de una MBD es igual al mínimo del número de filas – 1 o columnas – 1.
Inercia principal: varianza explicada por un eje principal, corresponde al eigenvalor. La suma de las inercias principales es igual a la inercia total. Las inercias principales varían entre 0 y 1. Si es igual a 1, indica una asociación exclusiva entre UE y variables, representando completa dependencia entre ambas. Si es igual a 0, indica completa independencia. La última inercia principal es igual a 0.
Perfil fila: conjunto de frecuencias de las filas dividido por el total de la fila.
Perfil columna: conjunto de frecuencias de las columnas dividido por el total de la columna.
Perfil promedio: perfil fila (o columna), cuyos elementos corresponden al promedio de las columnas (o filas), respectivamente.
Masa: valores que componen el perfil fila (o columna) promedio.
***V* de Cramér:** proporción de la inercia máxima que presenta una MBD. Un valor igual a 0 corresponde a una completa independencia entre filas y columnas, mientras que un valor igual a1 corresponde a una completa dependencia.
***Biplot* o mapa simétrico:** representación gráfica simultánea de UE y variables en un espacio común. Sólo se pueden obtener conclusiones de la posición de una UE con respecto a la posición de todas las variables, pero es imposible sacar conclusiones sobre la distancia entre una UE y una variable específica.
***Biplot* o mapa asimétrico:** representación gráfica simultánea de UE y variables, en la cual se visualizan las UE en el espacio de las variables o viceversa. Permite visualizar la distancia (intensidad de la relación) entre UE y variables.
Tabla de contingencia: Tabla de doble entrada que muestra el número de observaciones para dos o más variables categóricas.

El CA también genera un eigenvalor que mide el porcentaje de varianza explicada por cada eje, como en el PCA. Sin embargo, hay una terminología particular del CA (Box 6.2): la varianza total se denomina "inercia total". Los eigenvalores o "inercias principales" descomponen la inercia total a lo largo de cada eje. La decisión de cuántos ejes retener es similar a la del PCA.

A continuación se resumen los pasos del CA. Tomaremos como ejemplo la MBD5 hipotética dada en la Tabla 6.5. Cada frecuencia absoluta de la MBD es denominada f_{ij}, donde *i* corresponde a la fila y *j* a la columna.

Tabla 6.5. MBD5 de tres sitios (A a C) × tres variables (sp1 a sp3). Cada valor en la MBD es definido como f_{ij}.

Sitio	sp1	sp2	sp3	Suma
A	12	1	2	15
B	14	5	8	27
C	5	21	4	30
Suma	31	27	14	$N = 72$

En primer lugar, se calculan las frecuencias relativas p_{ij} (Legendre y Legendre 1998). Esto se logra dividiendo las frecuencias de cada UE por el total de su fila correspondiente (Tabla 6.6). Las filas de esta nueva matriz se denominan "perfiles fila" *pf* y son los elementos más básicos y fundamentales del CA (Greenacre 2008). Este tipo de dato se denomina composicional (Pawlowsky-Glahn y Buccianti 2011), donde la suma de las proporciones es igual a 1.

Tabla 6.6. Perfiles fila. En esta matriz, cada valor corresponde a una proporción definida como p_{ij}. Las filas se denominan "perfiles fila" *pf*. La última fila corresponde al perfil fila promedio.

Sitio	sp1	sp2	sp3	Suma
A	12/15 = 0,80	1/15 = 0,07	2/15 = 0,13	15/15 = 1,00
B	14/27 = 0,52	5/27 = 0,19	8/27 = 0,30	27/27 = 1,00
C	5/30 = 0,17	21/30 = 0,70	4/30 = 0,13	30/30 = 1,00
Perfil fila promedio	31/72 = 0,43	27/72 = 0,38	14/72 = 0,19	72/72 = 1,00

De esta forma, los perfiles fila representan la proporción relativa de especies en cada sitio, y es esta información la que es relevante para el análisis. Los perfiles fila dan una idea de cómo las especies se distribuyen entre los sitios, independientemente del tamaño de muestra, lo que permite la comparación entre distintos sitios. El último perfil fila de la Tabla 6.6. corresponde al "perfil fila promedio" o centroide, ya que representa la proporción promedio esperada de cada especie en un sitio cualquiera (Greenacre 2008). De esta forma, podemos comparar si la abundancia de una especie en un sitio determinado está por encima o por debajo del promedio. Los valores del perfil promedio se denominan "masas" (Greenacre y Primicerio 2014). Estas masas se utilizan como pesos en el análisis, con el fin de ponderar las especies de acuerdo a su proporción relativa. Las masas del perfil fila promedio se denominan masas de las columnas.

Dado que hay tres variables, podemos representar los perfiles fila (sitios) en un espacio tridimensional (Fig. 6.6A), donde cada proporción para cada especie corresponde a una coordenada. Los tres perfiles se hallan exactamente sobre un plano definido por un triángulo equilátero que une los siguientes tres puntos de la unidad (1, 0, 0), (0, 1, 0) y (0, 0, 1). Estos puntos corresponden a una especie concentrada en un único sitio (100% de abundancia relativa). Debido a que los perfiles en el espacio tridimensional se hallan sobre un triángulo (bidimensional), podemos situar los perfiles sobre un plano (Fig. 6.6B). Esta representación se conoce como diagrama ternario, de uso frecuente en el análisis de datos geológicos y químicos. A partir de los valores de los perfiles se pueden trazar líneas paralelas a los lados del triángulo estas líneas confluyen en un punto que definen la posición de cada sitio.

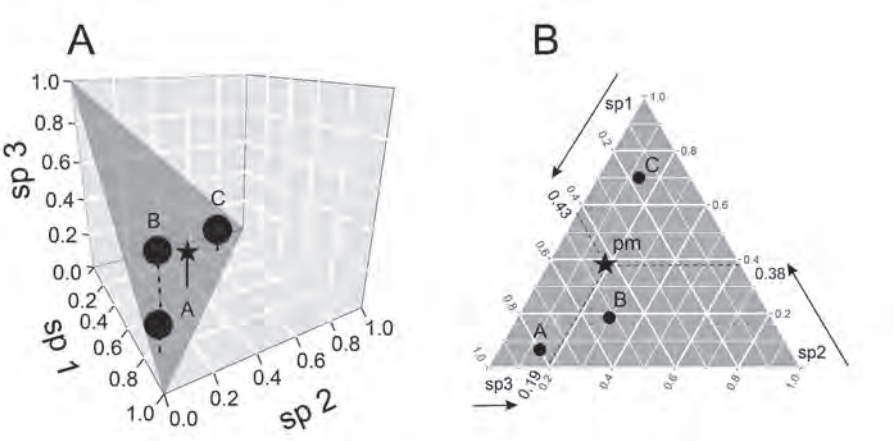


Fig. 6.6. Representación de los perfiles en un: (A) espacio tridimensional de tres especies (sp1 a sp3); (B) espacio bidimensional de tres especies (sp1 a sp3); A, B y C representan los sitios (perfiles fila), mientras que la estrella representa el perfil fila promedio (centroide). En (A) los sitios se encuentran todos sobre un mismo plano que delimita un triángulo equilátero de longitud igual a 1. Por lo tanto, los sitios pueden representarse en un espacio de dos dimensiones, denominado diagrama ternario; en (B) las distancias están representadas por las distancias euclideanas entre los perfiles fila.

Si no hubiera diferencias entre los sitios en cuanto a sus abundancias relativas, esperaríamos que los perfiles fila sean similares al perfil fila promedio. Las diferencias que observamos serían sólo debidas a efectos aleatorios de la muestra. Para establecer si dichas diferencias son lo suficientemente grandes para no ser debidas al azar, calculamos un coeficiente de distancia entre las proporciones observadas O_{ij} y las esperadas E_{ij} bajo el supuesto del azar. Para esto, calculamos la distancia chi-cuadrado (ver Cap. 4) para cada celda de la MBD, que es una distancia euclideana ponderada (Legendre y Legendre 1998).

$$\chi^2 = w_{ij}\left(O_{ij} - E_{ij}\right)^2$$

$$\chi^2 = \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

¿Cuál sería el valor esperado, por ejemplo, para la especie 1 en el sitio C? En el sitio C se detectaron 30 individuos, y dado que la proporción esperada de la especie 1 es 0,43, esperaríamos que haya 30×0,43 = 12,9 individuos. Sin embargo, este valor debemos expresarlo en términos relativos dividiéndolo por el total de la fila, que es 12,9/30 = 0,43. Por lo tanto, bajo el supuesto del azar (no hay asociación entre filas y columnas) el valor esperado de cada celda es el correspondiente al perfil fila promedio. Reemplazando estos valores en la ecuación anterior obtendremos:

$$\chi^2_{ij} = n_i \times \frac{\left(p_{ij} - pf_j\right)^2}{pf_j}$$

$$\chi^2_{C1} = 30 \times \frac{\left(0{,}17 - 0{,}43\right)^2}{0{,}43}$$

$$\chi^2_{C1} = 30 \times \frac{\left(-0{,}26\right)^2}{0{,}43}$$

$$\chi^2_{C1} = 4{,}85$$

Los términos *pf* y n_i representan el perfil fila promedio y el total de la fila *i*, respectivamente. Los valores observados menos los esperados se muestran en la Tabla 6.7. De esta forma, cada celda contiene un valor que representa cuánto difiere la proporción observada del promedio esperado. Por ejemplo, vemos que la especie 1 en el sitio C tiene abundancias relativas bastante por debajo del promedio, mientras que la especie 3 en el sitio A no difiere demasiado del promedio.

Tabla 6.7. Proporciones observadas – esperadas de la MBD5.

Sitio	sp1	sp2	sp3
A	0,80 – 0,43 = 0,37	0,07 – 0,38 = –0,31	0,13 – 0,19 = –0,06
B	0,52 – 0,43 = 0,09	0,19 – 0,38 = –0,19	0,30 – 0,19 = 0,11
C	0,17 – 0,43 = –0,26	0,70 – 0,38 = 0,32	0,13 – 0,19 = –0,06
Perfil fila promedio	31/72 = 0,43	27/72 = 0,38	14/72 = 0,19

Los valores de chi-cuadrado se muestran en la Tabla 6.8. En lugar de utilizar el total de la fila como parte del ponderador, dividimos esta cantidad por el total de datos (*N*), de forma que se multiplique por las masas de las filas. Esto equivale a dividir el valor de chi-cuadrado por *N*, debido a que es una constante.

$$\frac{\chi_{ij}^2}{N} = \frac{n_i}{N} \times \frac{\left(p_{ij} - pf_j\right)^2}{pf_j}$$

$$\frac{\chi_{ij}^2}{N} = masa_i \times \frac{\left(p_{ij} - pf_j\right)^2}{pf_j}$$

Tabla 6.8. Valores de chi-cuadrado para la MBD5.

Sitio	sp1	sp2	sp3
A	$15\times0,37^2/0,43 = 4,76$	$15\times(-0,31)^2/0,38 = 3,80$	$15\times(-0,06)^2/0,19 = 0,29$
B	$27\times0,09^2/0,43 = 0,49$	$27\times(-0,19)^2/0,38 = 2,59$	$27\times0,11^2/0,19 = 1,44$
C	$30\times(-0,26)^2/0,43 = 4,85$	$30\times0,32^2/0,38 = 8,45$	$30\times(-0,06)^2/0,19 = 0,58$

La suma de todos los valores de chi-cuadrado da como resultado una medida que cuantifica las diferencias entre las proporciones observadas y las esperadas de la MBD, independientemente del tamaño muestral (ya que se divide por *N*). A esta medida se la denomina "inercia total" ϕ^2 (Greenacre 2008). Para nuestro ejemplo:

$$\phi^2 = \frac{\sum \chi_{ij}^2}{N}$$

$$\phi^2 = \frac{4,76 + 0,49 + 4,85 + 3,80 + 2,59 + 8,45 + 0,29 + 1,44 + 0,58}{72}$$

$$\phi^2 = \frac{27,24}{72}$$

$$\phi^2 = 0,378$$

La inercia total es una medida de la variabilidad presente en la MBD. Si es alta, significa que las UE se encuentran dispersas y alejadas del perfil fila promedio. Cuando la inercia es baja, los perfiles fila presentan poca variación y se hallan cerca de su perfil fila promedio (Greenacre 2008). En este caso, decimos que hay poca asociación (o correlación) entre las filas y las columnas. En términos biológicos esto implica que en nuestro ejemplo las especies se distribuyen al azar, con poca preferencia por alguno de los sitios. Cuanto mayor sea la inercia, más cerca se encontrarán los perfiles fila de los vértices del triángulo en el diagrama ternario. Es decir, mayor será el grado de asociación entre filas y columnas. Si todos los perfiles fueran idénticos, todos coincidirían con el perfil fila promedio, por lo que las distancias chi-cuadrado serían cero, así como también la inercia total. En contraste, si todos los perfiles se hallan sobre los vértices del triángulo, la inercia será máxima, en cuyo caso se corresponde con la dimensionalidad del espacio (espacio de dos dimensiones en nuestro ejemplo) o en términos matemáticos, al mínimo del número de filas menos 1 o columnas menos 1 (igual a 2 en nuestro ejemplo).

Dado que la distancia chi-cuadrado corresponde a una distancia euclideana ponderada, podemos calcular la distancia entre dos perfiles fila p_{Aj} y p_{Bj}, como el cociente entre la diferencia de los perfiles al cuadrado y el perfil fila promedio (ver *Distancia chi-cuadrado* en el Cap. 4):

$$\chi^2_{AB} = \frac{\left(p_{Aj} - p_{Bj}\right)^2}{pf_j}$$

$$\chi^2_{AB} = \left(\frac{p_{Aj}}{\sqrt{pf_j}} - \frac{p_{Bj}}{\sqrt{pf_j}}\right)^2$$

De esta forma, podemos dividir cada término por la raíz cuadrada del perfil fila promedio y graficarlos en un espacio euclideano (Tabla 6.9, Fig. 6.7). Si el denominador fuera igual a 1, la fórmula correspondería a la distancia euclideana, como se muestra en la Figura 6.4. En la distancia chi-cuadrado los valores de las coordenadas aumentan porque se dividen por valores menores a 1. Si un denominador es menor que otro, las coordenadas aumentarán y viceversa. Por lo tanto, esta transformación afecta más a las frecuencias más bajas. Asimismo, los valores de los vértices ya no tomarán como valor máximo 1, sino $1/\sqrt{pf_j}$.

Tabla 6.9. Coordenadas de los perfiles fila, donde cada perfil se divide por la raíz cuadrada de su perfil fila promedio. Esto permite representar la distancia chi-cuadrado entre perfiles en un espacio euclideano.

Sitio	sp1	sp2	sp3
A	0,80/√0,43 = 1,22	0,07/√0,38 = 0,11	0,13/√0,19 = 0,30
B	0,52/√0,43 = 0,79	0,19/√0,38 = 0,30	0,30/√0,19 = 0,67
C	0,17/√0,43 = 0,25	0,70/√0,38 = 1,14	0,13/√0,19 = 0,30
Perfil fila promedio	0,43/√0,43 = 0,66	0,38/√0,38 = 0,61	0,19/√0,19 = 0,44

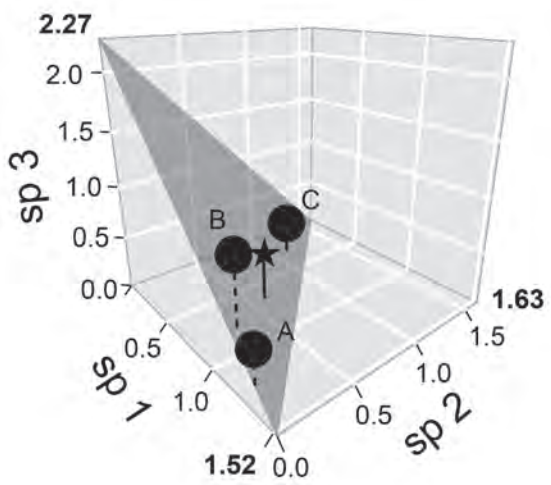


Fig. 6.7. Representación de los perfiles en un espacio tridimensional de tres especies (sp1 a sp3). A, B y C representan los sitios (perfiles fila), mientras que la estrella representa el perfil fila promedio (centroide). Las distancias están representadas por distancias chi-cuadrado entre los perfiles fila, que corresponden a distancias euclideanas ponderadas. Los valores en negrita corresponden a los vértices de un triángulo con valores iguales a 1/√perfil fila promedio.

Al igual que en el PCA, las MBD suelen tener numerosas dimensiones. Así, la idea es representar las UE en un espacio de pocas dimensiones, pero intentando mantener las relaciones de la MBD original de la mejor forma posible. En nuestro ejemplo, el mapa de dos dimensiones representa exactamente las distancias entre las UE, porque hay tres variables solamente. Sin embargo, podemos intentar representar

las UE en un espacio de una sola dimensión. Esto se logra, al igual que en el PCA generando una nube elíptica alrededor de las UE y proyectando las mismas sobre la dimensión que atraviesa el eje mayor de la elipse (Fig. 6.8). Esto significa que también podemos extraer eigenvectores y eigenvalores.

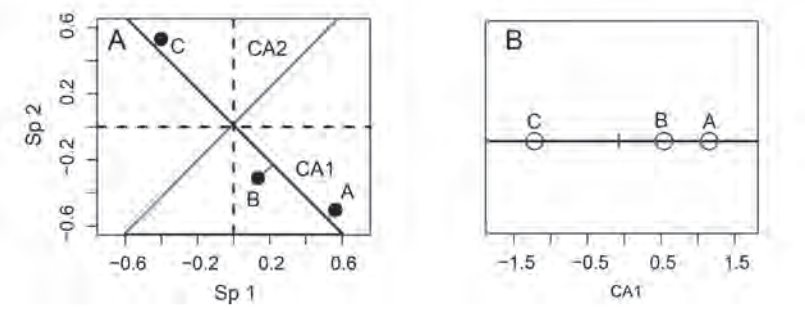


Fig. 6.8. CA de la MBD5 (Tabla 6.5). (A) Gráfico de dispersión de tres sitios (A a C) × dos especies (sp1 y sp2) estandarizado. El CA1 corresponde al eje donde se encuentra la mayor variación en las UE, y es donde se ubicará el primer eje principal; (B) las UE se proyectan sobre el CA1 mediante las líneas perpendiculares y se rota la configuración; estas nuevas coordenadas se denominan *scores*. De esta forma, se ha representado la MBD original de dos variables en un espacio de menor dimensión (un solo eje), sin pérdida de información sustancial (ya que las UE no se alejan demasiado en el CA2).

Los eigenvalores de los ejes se conocen como "inercias principales". En el CA, a diferencia del PCA, los eigenvalores varían entre 0 y 1, por lo que la inercia total nunca puede ser mayor al número de variables menos 1 (cuando todos los ejes tienen eigenvalores igual a 1). Un eigenvalor = 1 indica una asociación exclusiva entre las UE y las variables (cada variable está presente en una sola UE, sin coincidir con las demás variables), representando una completa dependencia entre ambas (Tabla 6.10). Del mismo modo, un eigenvalor = 0 indica una completa independencia. Al igual que en el PCA, la inercia total corresponde a la suma de todos los eigenvalores del análisis. En el CA, a diferencia del PCA, el último eigenvalor siempre es igual a 0, porque el último valor esperado del perfil fila es igual a 100 menos la suma de los restantes porcentajes. Esto significa que la última celda del perfil queda automáticamente definida y es, por lo tanto no informativa. En este ejemplo los eigenvalores (calculados mediante software) son λ_1 = 0,35 y λ_2 = 0,03. La inercia total es igual a 0,35 + 0,03 = 0,38. De igual forma que en el PCA, podemos calcular el porcentaje de variación explicada por cada eje como: 100% × 0,35/0,38 = 91,58% (CA1) y 100% × 0,03/0,38 = 8,42% (CA2). Esto indica que la representación en una sola dimensión de la MBD original es particularmente buena.

Como en nuestro ejemplo la inercia máxima es igual a 2, podemos calcular qué proporción (o porcentaje si se multiplica por 100) de la inercia máxima tiene la MBD analizada como: 0,38/2 = 0,19. Esta medida varía entre 0 y 1, y se conoce como *V* de Cramér (1946). Un valor igual a 0 corresponde a una completa independencia, mientras que un valor igual a 1 corresponde a una completa dependencia. Por lo tanto, esta medida es análoga al coeficiente de correlación para datos categóricos.

Tabla 6.10. Caso hipotético de una MBD de tres sitios (A a C) × tres especies (sp1 a sp3), con una completa dependencia entre filas y columnas. En este caso, los eigenvalores son todos iguales a 1 y la inercia total es igual a 2, por lo que *V* = 2/2 = 1.

Sitio	sp1	sp2	sp3
A	10	0	0
B	0	0	8
C	0	21	0

Tanto las UE como las variables pueden representarse simultáneamente en un *biplot* (Fig. 6.9A). Las variables se representan de la misma forma que las UE, sólo que intercambiando filas por columnas. En este gráfico, sin embargo, hay dos diferencias importantes con respecto al PCA. Una es que se represen-

tan distancias chi-cuadrado entre perfiles en lugar de distancias euclideanas, y la otra es que cada muestra es ponderada diferente según su masa correspondiente. Si se intercambian filas por columnas y se repite el análisis, se llegará exactamente al mismo resultado, por lo que el modo Q y R son equivalentes.

Las reglas de interpretación son similares a las de todos los métodos de ordenación: (1) las UE (perfiles fila) cercanas en el espacio tienen características similares en cuanto a sus variables; (2) las variables (perfiles columna) cercanas en el espacio tienen características similares en cuanto a las UE en las que aparecen. Además, cada fila (o columna) está más cerca de las columnas (o filas) con las cuales está más relacionada y está lejos de las columnas (o filas) menos relacionadas. En la Figura 6.9A se observa que el CA1 separa los sitios A y B del sitio C, y que los sitios A y B son más parecidos entre sí. La especie 2 está más asociada en términos generales al sitio C, mientras que la especie 1 está más asociada a los sitios A y B, y la especie 3 está más asociada al sitio B. Esto concuerda con lo observado en la MBD original (Tabla 6.5). Este tipo de *biplot* se denomina mapa simétrico, y con éste sólo se pueden hacer observaciones de la posición de una fila con respecto a la posición de todas las columnas, pero es imposible sacar conclusiones sobre la distancia entre una fila y una columna específica (Husson *et al*. 2017, Kassambara 2017b).

Para interpretar las distancias (intensidad de la relación) entre una columna y una fila es necesario realizar el *biplot* o mapa asimétrico; esto significa que los perfiles columna deben ser representados en el espacio de las filas o viceversa (Fig. 6.9B-C). El objetivo de estos *biplots* es visualizar la intensidad de la relación expresada por la ordenación. En estos *biplots* es conveniente mostrar las UE y las variables como vectores. Si el ángulo entre dos vectores es agudo, entonces hay una fuerte asociación entre la UE y la variable correspondiente (Husson *et al.* 2017).

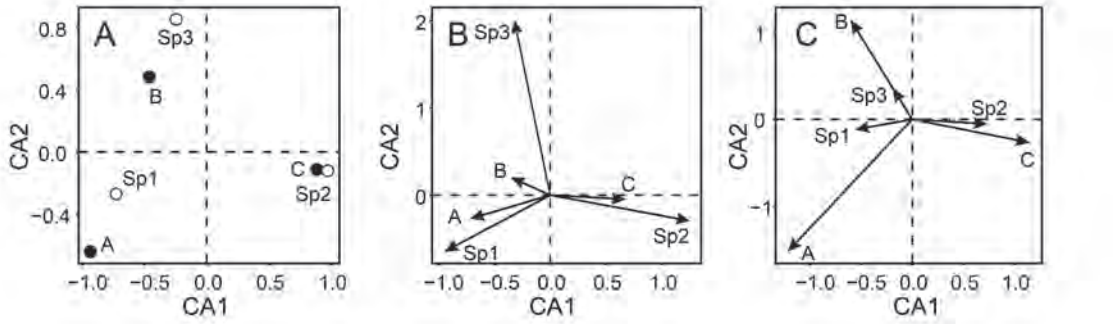


Fig. 6.9. (A) *Biplot* simétrico de la MBD5 (Tabla 6.5). Los círculos negros muestran los sitios (A a C), y los círculos blancos las especies (sp1 a sp3); (B) *biplot* asimétrico de las columnas en el espacio de las filas; (C) *biplot* asimétrico de las filas en el espacio de las columnas.

ANÁLISIS DE COORDENADAS PRINCIPALES

En Biología es común el uso de datos cualitativos o con exceso de dobles ceros (presencia-ausencia, abundancia). En tales casos, la distancia euclideana entre las UE no es apropiada conceptualmente, por lo que el PCA no es adecuado (ver en este capítulo *Efecto arco*). Tanto el PCA como el CA imponen la distancia que se conserva entre las UE: distancia euclideana y chi-cuadrado, respectivamente. Si el investigador está interesado en ordenar las UE de acuerdo a otra medida de similitud, entonces puede utilizar el análisis de coordenadas principales (PCoA). Gower (1966) desarrolló este método para representar un conjunto de UE en un espacio euclideano cuyas relaciones sean cuantificadas por cualquier coeficiente de similitud (Cap. 4). Esto permite utilizar cualquier tipo de dato en la construcción de la MBD. En el caso de utilizar la distancia euclideana para construir una matriz de distancia (MD), los eigenvectores obtenidos por PCoA (coordenadas principales) serán exactamente iguales a los componentes principales obtenidos por PCA sobre la MBD original. A diferencia del PCA, el PCoA no produce *loadings* de las variables, dado que se calcula a partir de la MD de las UE; sin embargo pueden calcularse posteriormente.

Como el PCA o el CA, el PCoA genera un conjunto de ejes ortogonales (no correlacionados entre sí) cuya contribución a la variación total de la MBD se cuantifica a través de sus eigenvalores. Dado que se

basa en una MD, el PCoA puede representar tanto a las UE (modo Q) como a las variables (modo R). Sólo la distancia euclideana y los coeficientes derivados generan un número de ejes principales igual al número de variables de la MBD (cuando $p > n$). Otros coeficientes pueden producir más o menos ejes. La dimensionalidad del espacio de coordenadas principales depende del número de variables y del coeficiente de distancia utilizado (Legendre y Legendre 1998). Si se requiere proyectar a las variables sobre la ordenación de las UE, éstas pueden relacionarse *a posteriori* con los ejes de ordenación, utilizando correlaciones entre las variables originales y los ejes (Legendre y Legendre 1998) o promedios ponderados, y representarse sobre el gráfico de ordenación. La interpretación de la ordenación es igual a la de otros métodos, una mayor proximidad entre las UE representa una mayor similitud.

Retomaremos como ejemplo la MBD4 (Tabla 6.11) presentada en el Capítulo 4, que contiene una variable continua (longitud de la hoja), una variable categórica (color de la flor) con dos estados (roja y blanca), una variable binaria (presencia-ausencia de raíz secundaria) con un dato faltante (NA), y una variable ordinal (densidad de glándulas). La densidad de glándulas corresponde a una variable cuantitativa (glándulas/cm^2) transformada a una escala ordinal (1 = entre 0 y 10 glándulas/cm^2; 2 = entre 11 y 100 glándulas/cm^2).

Tabla 6.11. MBD4 de tres UE (sp1 a sp3) × caracteres morfológicos.

Especie	Longitud de la hoja (cm)	Flor roja	Flor blanca	Raíz secundaria	Densidad de glándulas
sp1	10,0	0	0	NA	2
sp2	9,7	0	1	0	1
sp3	5,0	1	1	1	2

El procedimiento es el siguiente:

1. La matriz inicial debe ser una matriz de distancia (MD) con valores D_{ij}. También pueden realizarse los cálculos sobre una matriz de asociación, aunque por conveniencia primero se transforma a una MD (distancia = 1 – asociación). Para el ejemplo, aplicaremos la distancia de Gower (Tabla 6.12), que permite analizar matrices con cualquier tipo de dato (Cap. 4).

Tabla 6.12. MD de Gower de la MBD5.

	sp1	sp2	sp3
sp1	0	0,69	0,75
sp2	0,69	0	0,79
sp3	0,75	0,79	0

2. La MD se transforma en una nueva matriz A con valores a_{ij} (Tabla 6.13):

$$a_{ij} = -\frac{1}{2} D_{ij}^{\,2}$$

Tabla 6.13. Matriz A transformada con valores a_{ij}.

	sp1	sp2	sp3	Promedio
sp1	0	$-0,5\times0,69^2 = -0,24$	$-0,5\times0,75^2 = -0,28$	–0,17
sp2	$-0,5\times0,69^2 = -0,24$	0	$-0,5\times0,79^2 = -0,31$	–0,18
sp3	$-0,5\times0,75^2 = -0,28$	$-0,5\times0,79^2 = -0,31$	0	–0,20
Promedio	–0,17	–0,18	–0,20	–0,18

3. Se centran los valores a_{ij} con respecto a la media de filas $\overline{a}_i$, columnas $\overline{a}_j$ y total de datos $\overline{a}_{ij}$ obteniendo una matriz Δ con valores δ_{ij} (Tabla 6.14). La media de las filas y las columnas dan el mismo resultado porque la MD es simétrica:

$$\delta_{ij} = a_{ij} - \overline{a}_i - \overline{a}_j + \overline{a}_{ij}$$

La transformación de A en Δ no es estrictamente necesaria, simplemente se utiliza para eliminar uno de los eigenvalores que podría ser el mayor y sólo representar la distancia entre el origen y el centroide.

Tabla 6.14. Matriz Δ transformada y centrada con valores δ_{ij}.

	sp1	**sp2**	**sp3**
sp1	0 – (–0,17) – (–0,17) – 0,18 = 0,16	–0,24 – (–0,17) – (–0,18) – 0,18 = –0,07	–0,28 – (–0,17) – (–0,20) – 0,18 = –0,10
sp2	–0,24 – (–0,18) – (–0,17) – 0,18 = –0,07	0 – (–0,18) – (–0,18) – 0,18 = 0,18	–0,31 – (–0,18) – (–0,20) – 0,18 = –0,12
sp3	–0,28 – (–0,20) – (–0,17) – 0,18 = –0,10	–0,31 – (–0,20) – (–0,18) – 0,18 = –0,12	0 – (–0,20) – (–0,20) – 0,18 = 0,21

4. Se calculan los eigenvectores y eigenvalores a partir de esta nueva matriz. Las coordenadas principales corresponden a los eigenvectores multiplicados por la raíz cuadrada de los eigenvalores (análogos a los *loadings*). Para nuestro ejemplo obtenemos mediante el uso de software dos eigenvalores diferentes de cero (λ_1 = 0,32, λ_2 = 0,23). Al igual que con los otros métodos de ordenación, podemos calcular la contribución de los ejes principales como: 100% × 0,32/(0,32 + 0,23) = 57,88% (PCoA1) y 100% × 0,23/(0,32 + 0,23) = 42,12% (PCoA2). La UE3 está más asociada con el PCoA1 (coordenada principal 0,45), mientras que la UE1 está más asociada con el PCoA2 (coordenada principal 0,37). La UE2 se asocia de forma similar con los PCoA1 y 2 (coordenadas principales –0,30 y –0,29, respectivamente). Por último, podemos representar gráficamente la ordenación (Fig. 6.10). Las UE1 y 2 son más similares entre sí que con respecto a la UE3. A su vez, el PCoA 1 separa las UE1 y 2 de la UE3.

Gower (1966) demostró que las distancias entre las UE en el espacio de coordenadas principales corresponden a las distancias entre las UE en un espacio euclideano, independientemente de la distancia utilizada, y luego de las transformaciones aplicadas a la MD.

Eventualmente, algunas MD pueden dar eigenvalores negativos, como resultado de no cumplirse con ciertas propiedades euclideanas. Este problema puede resolverse mediante una transformación de la MD (por ejemplo, raíz cuadrada) o mediante la adición de una constante a los valores de esta matriz lo suficientemente grande para eliminar este efecto, pero al mismo tiempo evitando crear nuevas dimensiones (Legendre y Legendre 1998). Para esto último existen dos métodos, que no dan exactamente los mismos resultados:

a) El método de Lingoes (1971) agrega una constante $2c_1$ a las distancias cuadradas (excepto en la diagonal principal que, por definición, son iguales a 0), $D_{ij}' = \sqrt{D_{ij}^2 + 2c_1}$. A continuación se transforma la MD en la matriz A. La constante c_1 es el valor absoluto del eigenvalor negativo más grande, obtenido a partir del PCoA sobre la matriz de distancia original. Luego de esta corrección todos los eigenvalores positivos aumentan un valor igual a c_1, por lo que el eigenvalor negativo más grande se hace 0. Así, la solución tiene al menos dos eigenvalores iguales a 0.

b) El método de Cailliez (1983) agrega una constante c_2 a las distancias de la matriz (excepto en la diagonal principal), $D_{ij}' = D_{ij} + c_2$. A continuación se transforma la MD en la matriz A. La constante c_2 es igual al mayor eigenvalor positivo obtenido a partir de una matriz particular que consta a su vez de cuatro matrices (para más detalles ver Legendre y Legendre 1998).

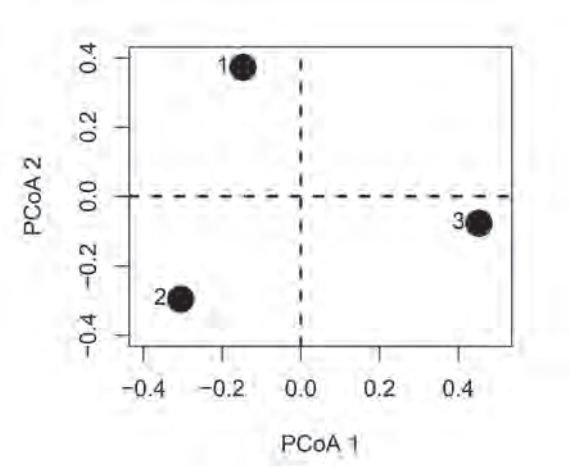


Fig. 6.10. PCoA aplicado a la MBD4. Los círculos indican las UE.

ANÁLISIS DISCRIMINANTE

A diferencia de los métodos de ordenación vistos anteriormente donde los grupos se forman *a posteriori* del análisis (no supervisados), el análisis discriminante (DA) o análisis de funciones discriminantes (FDA) comienza con un número de grupos de UE determinados *a priori* (supervisado). En el DA las UE ya están asignadas a cada uno de los grupos, y el método intenta determinar en qué grado las variables analizadas discriminan dichos agrupamientos (Legendre y Legendre 1998, Quinn y Keough 2002). Además, el DA funciona como herramienta predictiva, dado que permite clasificar nuevas UE. El DA ha sido ampliamente utilizado en sistemática para la identificación de especies, subespecies o poblaciones (Baker *et al.* 2002, Cheng y Han 2004, Tofilski 2008, Piro *et al*. 2018), así como para identificar sexos dentro de una misma especie (Green y Theobald 1989, Dechaume-Moncharmont *et al.* 2011, Montalti *et al.* 2012, Fuchs *et al.* 2017; para una crítica ver Indykiewicz *et al.* 2019). En menor medida, ha sido aplicado al análisis de distribución de especies, utilizando datos de presencia-ausencia (Manel *et al.* 1999, Olden y Jackson 2002). El método fue originalmente propuesto por Fisher (1936) para el caso particular de dos grupos ($K = 2$), denominado análisis discriminante simple, y luego extendido a $K \geq 2$ por Rao (1948, 1952), denominado análisis discriminante múltiple.

Como otros métodos de ordenación, el DA comienza a partir de una MBD de UE × variables. Sin embargo, debe especificarse *a priori* el grupo al que pertenece cada UE. Las variables se estandarizan a unidades de desvío estándar (ver Box 4.1). Cabe mencionar que si cada variable se centra restándole su media el DA arroja valores diferentes, pero tanto la interpretación del análisis como la clasificación de las UE es la misma (Legendre y Legendre 1998). En el caso particular del DA resulta conveniente trabajar con las variables sin estandarizar, ya que de lo contrario se deben estandarizar los valores cada vez que se clasifican nuevas UE.

A diferencia de los otros métodos de ordenación donde se buscan aquellos ejes que maximizan la variación en las UE, el DA busca aquellos ejes que maximizan las diferencias entre los grupos, minimizando al mismo tiempo las diferencias dentro de cada grupo (Legendre y Legendre 1998). Por lo tanto, la comparación se realiza dentro de una misma variable (variación entre grupos *vs.* variación dentro de grupos), y no es necesario estandarizar las variables de la MBD en las mismas unidades. De esta forma, el DA intenta encontrar un eje sobre el cual se puedan discriminar los grupos lo mejor posible (Fig. 6.11). En el Box 6.3 se muestra la terminología estándar del análisis discriminante.

Box 6.3. Terminología estándar del análisis discriminante

Función discriminante: variable hipotética que se construye a partir de las variables originales y permite discriminar o clasificar las UE en base a grupos preexistentes. Es análoga a un componente principal.
Límites de clasificación: valores obtenidos de un AD a partir del cual se pueden clasificar las UE, y determinan áreas que se denominan regiones de clasificación.
Análisis discriminante lineal: análisis discriminante en el cual las regiones de clasificación tienen límites lineales.
Análisis discriminante cuadrático: análisis discriminante en el cual las regiones de clasificación tienen límites no lineales.
Lambda de Wilks (Λ): estadístico que describe la calidad del DA, midiendo en qué grado difieren las posiciones de los centroides de cada grupo.
Matriz de confusión: tabla de doble entrada que permite visualizar el desempeño de la clasificación en una técnica supervisada.

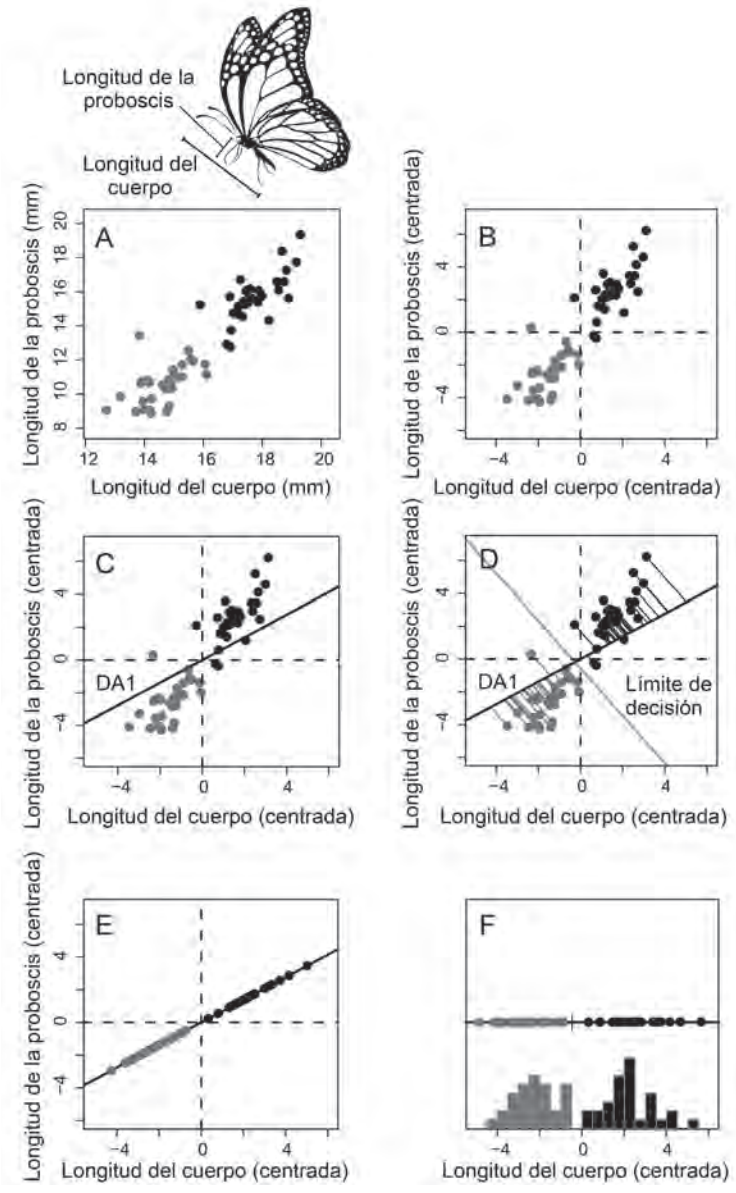


Fig. 6.11. DA. (A) Relación entre dos variables (longitud del cuerpo y longitud de la proboscis) en una especie de mariposa (60 UE) donde se distinguen hembras (círculos negros) y machos (círculos grises); (B) los datos se centran restándole la media de su respectiva variable; (C) el DA busca aquel eje (DA1) que maximiza la variación entre ambos grupos y minimiza la variación dentro de cada grupo. El PCA busca aquel eje que maximiza la variación en la MBD, que para este caso particular tendría una mayor pendiente que el DA1; (D) sobre el DA1 se traza un límite de decisión para clasificar a las UE, aquellas del lado izquierdo se clasifican como machos, mientras que las del lado derecho se clasifican como hembras; (E) las UE se proyectan sobre el DA1, por lo que hay cierta pérdida de información con respecto a la MBD original; (F) se rota la solución final para mejor visualización. También se pueden graficar los histogramas de cada grupo para mostrar en qué medida los ejes discriminan las UE.

Un ejemplo hipotético se muestra en la Figura 6.11, en el cual se midieron dos variables (longitud del cuerpo y de la probosciss) a 60 individuos (30 de cada sexo) de una especie de mariposa. Como se observa, ninguna de las dos variables por sí sola permite discriminar perfectamente ambos sexos, pero se puede encontrar un nuevo eje sobre el cual pueden discriminarse a la perfección ambos grupos. Esto se logra rotando los ejes originales hasta la posición donde se maximiza la diferencia entre los grupos y se minimiza la diferencia dentro de cada grupo. Al igual que en el PCA, estos nuevos ejes representan la combinación de todas las variables, y en la jerga del DA se denominan "funciones o ejes discriminantes" (Fig. 6.11). Así, el DA calcula tres matrices: (1) de variación total **T**, (2) de variación entre grupos **E** y (3) de variación dentro de grupos **D**. La matriz **E** cuantifica las diferencias cuadradas entre las medias de los grupos, mientras que la matriz **D** cuantifica las diferencias cuadradas entre cada UE y su media de grupo. Esta técnica preserva en el espacio multivariado la distancia de Mahalanobis (ver Cap. 4) entre las medias de los grupos.

Sin entrar en el detalle de los cálculos, lo que se busca es maximizar la relación entre **E** y **D** (Legendre y Legendre 1998). De esta forma, tanto el PCA como el DA buscan nuevos ejes que representen la combinación de las variables originales, sólo que difieren en los criterios para identificar esos ejes. El PCA busca ejes que maximicen la variación total en la MBD (matriz **T**), mientras que el DA busca ejes que maximicen el cociente entre variación entre grupos y variación dentro de grupos (Sharma 1996).

En nuestro ejemplo, la función discriminante (variables no estandarizadas) está constituida por la combinación de la longitud del cuerpo y la longitud de la probosciss, multiplicados por los coeficientes a_{ij} de la función discriminante (i = número de eje discriminante, j = número de variable):

$$\text{DA1} = a_{11} \times \text{longitud del cuerpo} + a_{12} \times \text{longitud de la probosciss}$$

Como se requiere tan sólo un eje para discriminar dos grupos, el número de funciones discriminantes es igual al número de grupos – 1. Sobre este eje se proyectan las UE, cuyas coordenadas se denominan *scores*, y se ubica un límite de clasificación que separa las UE de ambos grupos (Fig. 6.11). De esta forma el DA divide el espacio de las variables en regiones. En el análisis discriminante lineal (LDA) estas regiones tienen límites lineales, mientras que en el análisis discriminante cuadrático (QDA) estas regiones tienen límites no lineales (Fig. 6.12). Estos límites se utilizan para predecir a qué grupo pertenece una UE (que puede ser perteneciente a la MBD o una nueva UE), según sobre la región donde se ubique. Por ejemplo, la función discriminante de la Fig. 6.11 cuyos coeficientes son obtenidos mediante software es:

$$\text{DA1} = 0{,}65 \times \text{longitud del cuerpo} + 0{,}45 \times \text{longitud de la probosciss}$$

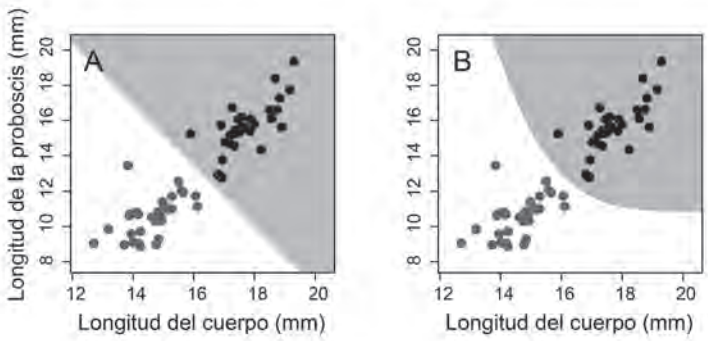


Fig. 6.12. Representaciones del LDA (A) y QDA (B) sobre el ejemplo de la Figura 6.9. Las áreas blancas y grises representan las regiones de decisión para cada grupo (círculos negros: hembras, círculos grises: machos).

Los valores absolutos de los coeficientes indican su importancia para la discriminación de los grupos. Así, la longitud del cuerpo es más importante que la longitud de la probosciss para discriminar entre machos y hembras.

Lo que necesitamos ahora es establecer un valor de corte (c), por encima del cual se asignará una nueva UE a la categoría hembra, y por debajo del cual se asignará una nueva UE a la categoría macho (esto se desprende de la Fig. 6.11F). Para esto, se calcula un promedio de las medias de cada grupo, ponderado por los coeficientes de la función discriminante:

$$c = \frac{1}{2}\left(a_{11}\bar{X}_{11} + a_{21}\bar{X}_{12} + a_{11}\bar{X}_{21} + a_{21}\bar{X}_{22}\right)$$
$$c = \frac{1}{2}\left[a_{11}\left(\bar{X}_{11} + \bar{X}_{21}\right) + a_{21}\left(\bar{X}_{12} + \bar{X}_{22}\right)\right]$$
$$c = 0,5 \times \left[0,65\left(14,61 + 17,73\right) + 0,45\left(10,60 + 15,70\right)\right]$$
$$c = 16,43$$

Supongamos ahora que el objetivo es clasificar una nueva UE con los siguientes valores: longitud del cuerpo = 17,7 mm y longitud de la proboscis = 16,75 mm; reemplazamos estos valores en el DA1:

DA1 = 0,65×17,7 + 0,45×16,75
DA1 = 19,04

En este punto se hace evidente la conveniencia de no estandarizar las variables; si se realizara el DA sobre las variables estandarizadas, sería necesario estandarizar las medidas 17,7 mm y 16,75 mm para clasificar la UE.

Como nuestra UE tiene un valor superior al valor de corte, la clasificamos como hembra. Con esta función de clasificación se suelen clasificar todas las UE de la MBD y analizar qué porcentaje de clasificación es correcto, mediante una matriz de confusión. Como señalan Legendre y Legendre (1998), el término "clasificación" es desafortunado, ya que en Biología, clasificar consiste en establecer grupos (como en el análisis de agrupamientos), mientras que la asignación de las UE a grupos preestablecidos se denomina "identificación".

Un estadístico muy utilizado para describir la calidad del DA es la lambda de Wilks (1932), que mide cuánto difieren entre sí los centroides de cada grupo. Esta medida (Λ) se calcula como el cociente entre los determinantes (líneas verticales) de las matrices **D** y **T**:

$$\Lambda = \frac{|\mathbf{D}|}{|\mathbf{T}|}$$

Este valor varía entre 0 (**D** = 0, máxima separación de los centroides) y 1 (**D** = **T**, no hay distinción entre los grupos). El valor de Λ (calculado mediante software) para este ejemplo es 0,17, indicando una separación relativamente buena entre machos y hembras en base a las dos variables medidas.

La lógica del DA para más de dos grupos (análisis discriminante múltiple) es la misma, y se van agregando nuevas regiones de decisión a medida que aumenta el número de grupos (Fig. 6.13). A diferencia del PCA, las funciones discriminantes no necesariamente son perpendiculares, ya que buscan diferenciar los grupos lo mejor posible.

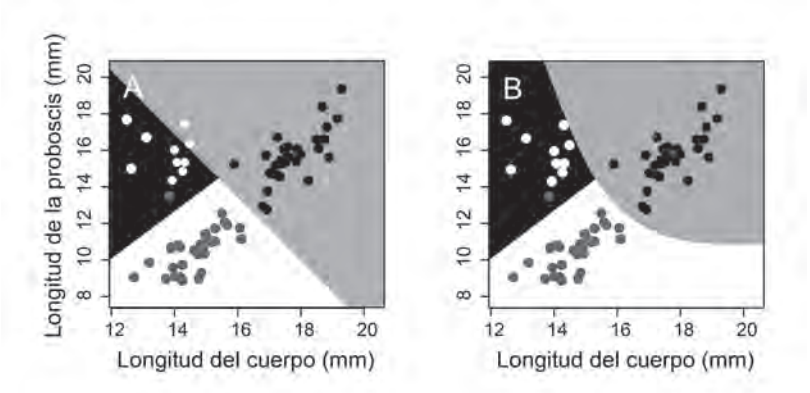


Fig. 6.13. Representaciones del LDA (A) y QDA (B) para un ejemplo hipotético de tres grupos. Las áreas blancas, negras y grises representan las regiones de decisión para cada grupo; los círculos blancos, negros y grises representan las UE de cada grupo.

A modo de ejemplo, tomaremos una MBD de 30 individuos pertenecientes a dos especies de picaflor (Picaflor Común y Picaflor Bronceado) × cuatro caracteres morfológicos (longitud cabeza-cola, cuerda del ala, longitud del culmen y longitud de la cola; Tabla 6.15). Los grupos son dos (15 UE por especie) y corresponden a cada especie de picaflor. Si bien las dos especies pertenecen a géneros distintos, son similares morfológicamente, pero diferenciables en términos de coloración (Fig. 6.14). Antes de realizar el DA, vamos a calcular los valores promedio y desvío estándar de las variables para cada especie (Tabla 6.16), y graficaremos los histogramas de las variables. Esto es con fines exploratorios y puede darnos una idea de qué variables son las que más difieren entre las especies (Tabla 6.16).

Tabla 6.15. MBD de 30 individuos machos pertenecientes a dos especies de picaflor (*Chlorostilbon lucidus* e *Hylocharis chrysura*) × cuatro caracteres morfológicos.

Especie	Longitud cabeza-cola (mm)	Cuerda del ala (mm)	Longitud del culmen (mm)	Longitud de la cola (mm)
Chlorostilbon lucidus	68,90	49,31	18,37	24,44
	78,84	51,97	20,31	30,08
	77,89	49,46	19,39	30,92
	74,92	52,08	18,84	26,20
	73,88	48,59	20,46	27,14
	77,28	52,04	19,41	31,02
	73,91	53,74	20,21	30,07
	75,94	52,26	20,60	33,69
	78,15	50,61	20,11	29,66
	76,46	52,44	19,00	29,94
	84,31	50,56	19,26	34,51
	73,28	51,47	19,7	27,03
	81,59	52,22	18,76	29,13
	75,56	50,16	21,92	25,68
	79,13	50,61	19,89	30,56
Hylocharis chrysura	78,60	53,21	21,39	30,25
	79,76	51,55	20,66	27,19
	79,27	52,05	21,12	29,10
	80,12	52,65	21,27	34,42
	81,68	54,20	20,80	30,15
	76,74	55,65	20,48	30,75
	80,14	50,90	21,07	29,80
	78,91	57,53	21,88	33,92
	79,42	52,14	20,25	26,99
	74,43	52,96	18,73	29,14
	76,84	52,02	21,09	30,07
	83,16	55,08	20,93	32,53
	76,77	51,81	22,51	27,85
	79,69	55,87	21,95	29,45
	78,42	54,86	20,68	28,73

Tabla 6.16. Estadísticos descriptivos (media ± desvío estándar) para cada variable y especie de picaflor.

Especie	Longitud cabeza-cola (mm)	Cuerda del ala (mm)	Longitud del culmen	Longitud de la cola (mm)
Chlorostilbon lucidus	76,77 ± 3,68	51,17 ± 1,41	19,75 ± 0,90	29,34 ± 2,82
Hylocharis chrysura	78,93 ± 2,14	53,50 ± 1,93	20,99 ± 0,86	30,02 ± 2,18

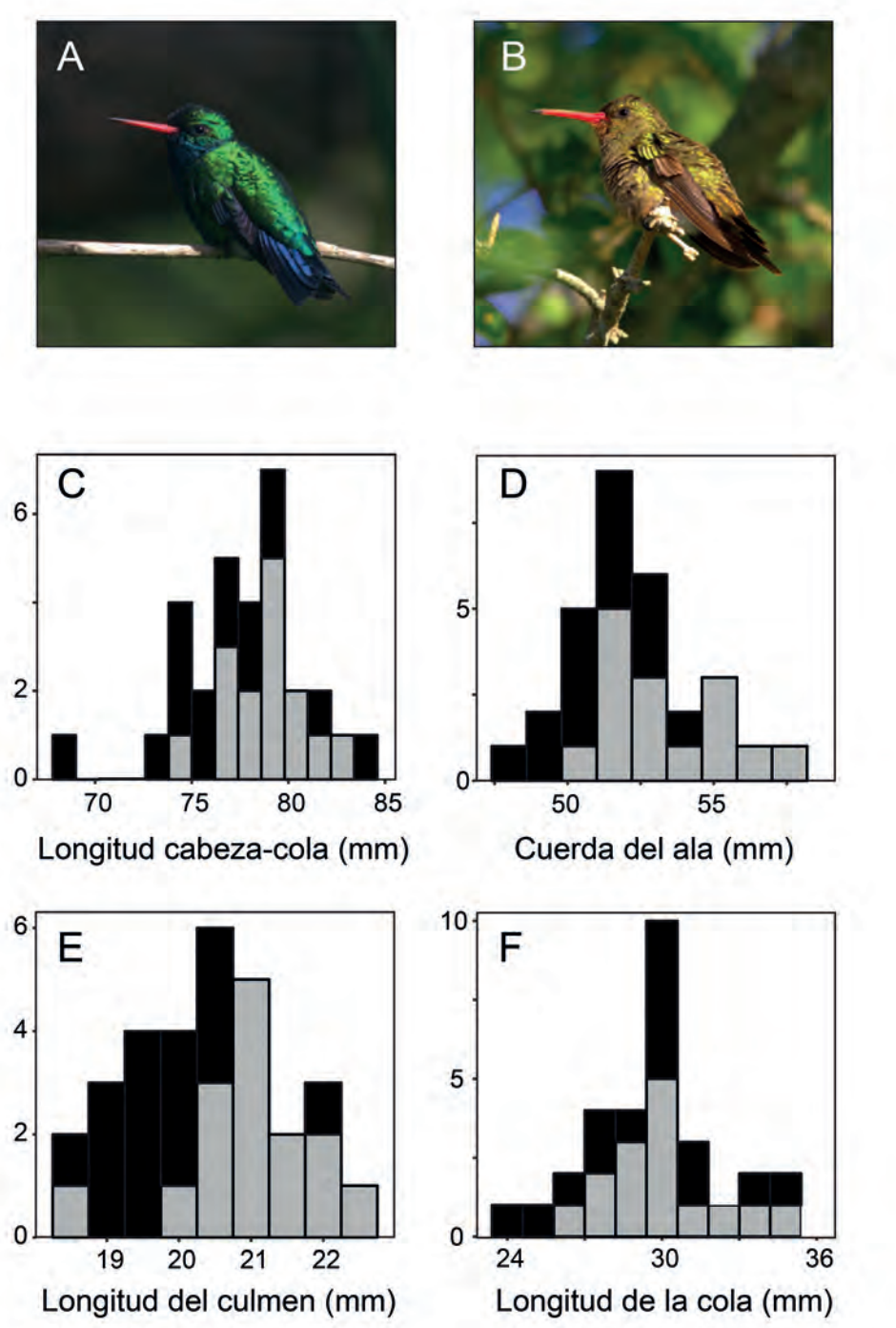


Fig. 6.14. (A) Macho de Picaflor Común (*Chlorostilbon lucidus*) y (B) macho de Picaflor Bronceado (*Hylocharis chrysura*). (C-F) Se muestran los histogramas de cuatro caracteres morfológicos para cada especie (barras negras: *C. lucidus*, barras grises: *H. chrysura*). Fotografías: Palacio, FX.

De la Tabla 6.16 y la Figura 6.14 se desprende que *H. chrysura* es, en promedio, de mayor tamaño que *C. lucidus*. Dado que el DA también funciona como una herramienta predictiva, vamos a utilizar una parte de la MBD original para realizar el análisis, y luego pondremos a prueba la función discriminante para evaluar qué tan bien clasifica nuevas UE. Para ello, seleccionaremos al azar 20 individuos de la MBD (datos de entrenamiento) y dejaremos 10 individuos para fines clasificatorios (datos de prueba). Un criterio frecuente es utilizar el 80% de las UE para ajustar el modelo (entrenamiento), y el 20% restante para evaluar su capacidad predictiva (prueba). En la práctica, muchos estudios utilizan la misma MBD como datos de entrenamiento y de prueba. Sin embargo, esto no es correcto, porque la capacidad predictiva de un modelo debe ponerse a prueba utilizando nuevas observaciones y no la misma matriz que dio origen al modelo (James *et al.* 2013).

La función discriminante generada mediante software por el LDA es la siguiente:

DA1 = 0,26×longitud cabeza-cola + 0,40×cuerda del ala + 0,69×culmen – 0,18×longitud de la cola

Se puede observar que la longitud del culmen es la variable más discriminatoria entre ambas especies, y en menor medida la cuerda del ala. El valor de corte obtenido mediante software este caso es c = 50,26. La función discriminante junto con el valor de corte es lo que debe reportarse en una publicación científica o informe, y es lo que permitirá a futuros investigadores utilizar como herramienta de identificación. Con esta información podemos calcular los porcentajes de clasificación correctos e incorrectos a través de una matriz de confusión (Tabla 6.17). En esta matriz hay cuatro resultados posibles, que determinan aquellas UE clasificadas correcta e incorrectamente. Así, vemos que de los cinco individuos de *C. lucidus*, cuatro fueron clasificados correctamente, mientras que un individuo de esta especie fue clasificado de forma incorrecta. En contraste, los cinco individuos de *H. chrysura* fueron clasificados correctamente. También es importante mostrar cómo se distribuyen las UE una vez proyectadas sobre el eje discriminante (Fig. 6.15). Si bien se trata de pocos individuos, la función discriminante tiene una buena capacidad predictiva.

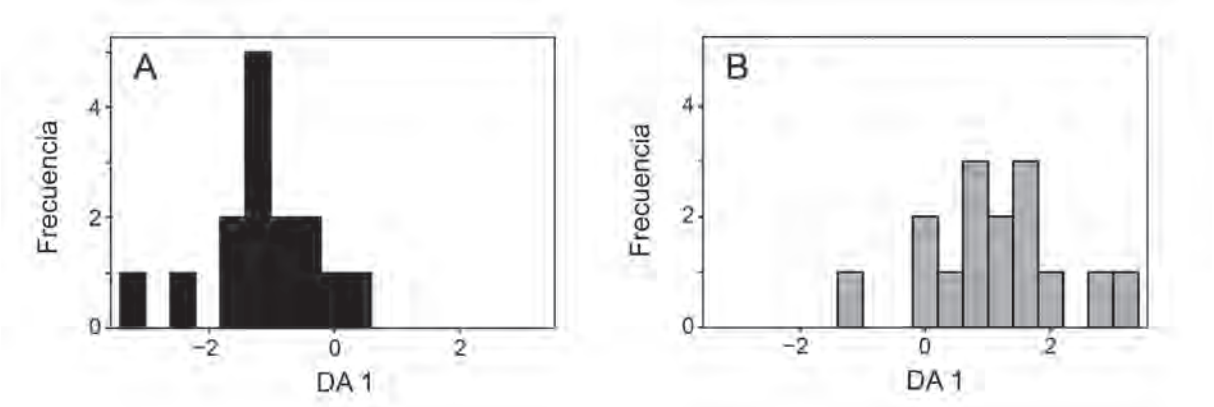


Fig. 6.15. Función discriminante (DA1) e histogramas de las UE proyectadas de (A) Picaflor Común (*Chlorostilbon lucidus*) y (B) Picaflor Bronceado (*Hylocharis chrysura*).

Tabla 6.17. Matriz de confusión de la MBD de especies de picaflores, utilizando el 80% de las UE para el ajuste del LDA y el 20% restante para predecir nuevas UE. Los números entre paréntesis corresponden a los porcentajes de clasificación. Los valores de la diagonal principal (4 y 5) representan las UE clasificadas correctamente, mientras que las restantes son las clasificadas de forma incorrecta.

		Especie predicha	
		Chlorostilbon lucidus	*Hylocharis chrysura*
Especie observada	*Chlorostilbon lucidus*	4 (80%)	1 (20%)
	Hylocharis chrysura	0 (0%)	5 (100%)

Sin embargo, con un único muestreo de un subconjunto de las UE se pueden haber obtenidos resultados muy diferentes. Por lo tanto, un método más apropiado para evaluar la capacidad predictiva del DA es repetir este procedimiento muchas veces, lo que se conoce como validación cruzada. Hay diversas variantes de este método (Wong 2015), aquí mostraremos la validación cruzada "dejando uno fuera" (*leave-one-out cross validation*). Este método consiste en ajustar el modelo con todas las UE – 1 y predecir a qué grupo pertenece la que quedó fuera de este conjunto (Cawley y Talbot 2003, James *et al*. 2013). Luego, se repite este procedimiento tantas veces como UE haya en la MBD, y se construye una matriz de confusión (Tabla 6.18).

Tabla 6.18. Matriz de confusión de la MBD de especies de picaflor, utilizando el método de validación cruzada "dejando uno fuera" resultante del LDA. Los números entre paréntesis muestran los porcentajes de clasificación. Los valores de la diagonal principal (12 y 13) representan las UE clasificadas correctamente, mientras que las restantes son las clasificadas de forma incorrecta.

		Especie predicha	
		Chlorostilbon lucidus	*Hylocharis chrysura*
Especie observada	*Chlorostilbon lucidus*	12 (80%)	3 (20%)
	Hylocharis chrysura	2 (13,3%)	13 (87,7%)

El DA tiene supuestos que en la práctica deben ser corroborados. En el caso del LDA, se asume que las UE dentro de cada grupo tienen una distribución normal, que la varianza de una variable dada es la misma para cada grupo y que la covarianza entre dos variables es la misma para cada grupo (homogeneidad de varianzas-covarianzas; Williams 1983). Un análisis más riguroso (ver en este capítulo *Técnicas de ordenación en R*) muestra que ambos supuestos se cumplen. El QDA, en contraste, es más flexible que el LDA, ya que asume que cada grupo tiene su propia varianza. Sin embargo, requiere de un gran número de parámetros a estimar cuando el número de variables es grande ($g \times p \times (p + 1)/2$ *vs.* $g \times p$), por lo que sólo es apropiado cuando hay un gran número de UE (James *et al.* 2013). Por otra parte, reportar la función discriminante de un QDA resulta más complejo por la gran cantidad de coeficientes generados.

En nuestro ejemplo, las distribuciones de las variables son bastante simétricas (Fig. 6.14) y los desvíos estándares son similares para cada especie (Tabla 6.16). Más aún, calculando la matriz de confusión a partir de la validación cruzada de un QDA, se obtiene una peor capacidad predictiva (Tabla 6.19).

Al igual que otras técnicas multivariadas el DA sufre de la maldición de la dimensionalidad (ver Cap. 1). En el Box 6.4 se brinda un ejemplo de los efectos que tiene utilizar pocas UE y muchas variables.

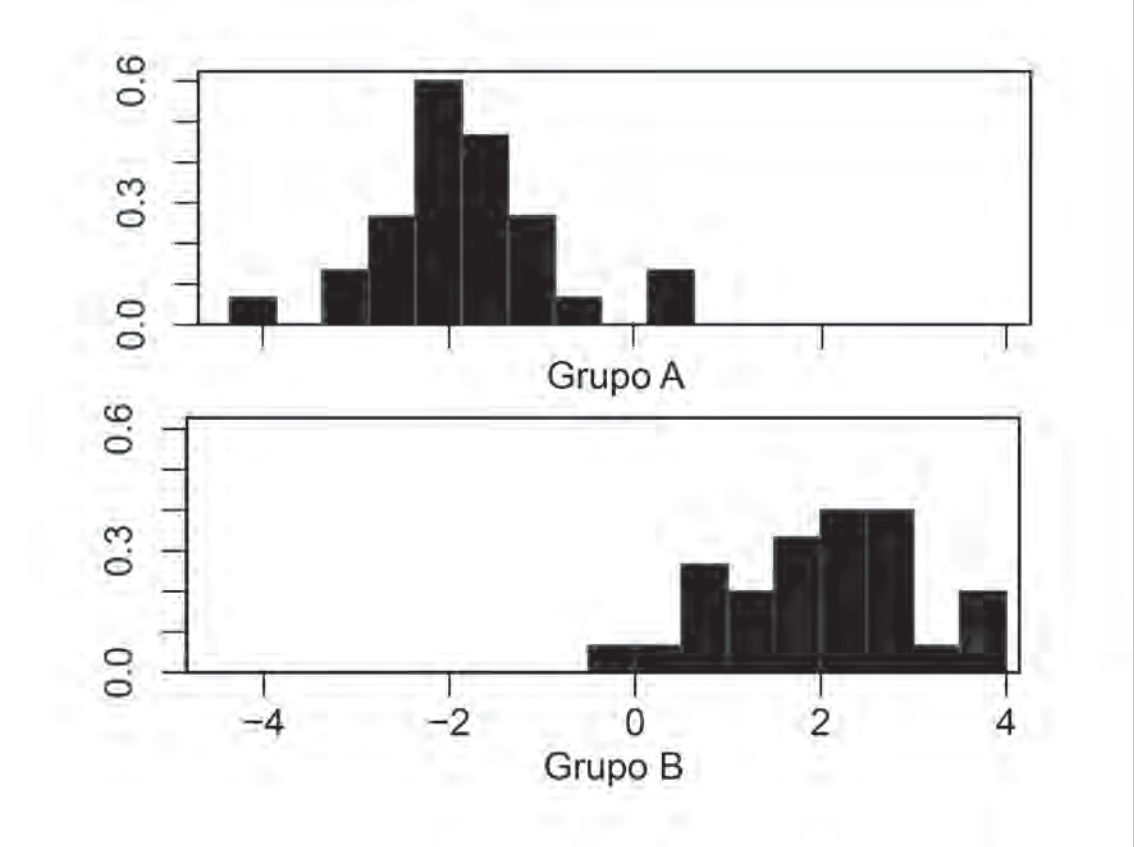
Tabla 6.19. Matriz de confusión de la MBD de especies de picaflores utilizando el método de validación cruzada "dejando uno fuera" resultante del QDA. Los números entre paréntesis muestran los porcentajes de clasificación. Los valores de la diagonal principal (10 y 13) representan las UE clasificadas correctamente, mientras que las restantes, son las clasificadas de forma incorrecta.

		Especie predicha	
		Chlorostilbon lucidus	*Hylocharis chrysura*
Especie observada	*Chlorostilbon lucidus*	10 (66,7%)	5 (33,3%)
	Hylocharis chrysura	2 (13,3%)	13 (87,7%)

Box 6.4. La maldición de la dimensionalidad en el DA y un intento de solución utilizando el PCA entre grupos

Si bien el DA es actualmente uno de los análisis multivariados más utilizados para distinguir grupos, requiere un número de UE que exceda ampliamente el número de variables ($n >> p$) para ser confiable estadísticamente, y de hecho no se puede calcular cuando el número de variables es mayor a $n - g$, donde g es el número de grupos. En la figura se muestran los histogramas pertenecientes a dos grupos con 30 UE cada uno ($n = 60$) y 50 variables. Los valores fueron generados de forma aleatoria, al igual que la asignación de las UE a cada grupo; sin embargo, se observa una clara distinción entre ambos grupos. No todas las muestras aleatorias dan como resultado una separación tan clara entre los grupos, pero la mayoría de ellas presentan una mayor separación que la esperada intuitivamente (Rohlf com. pers.).

Una de las alternativas más utilizadas para enfrentar este problema es el PCA entre grupos (*between-groups* PCA; bgPCA), que consiste en aplicar un PCA sobre las medias o centroides de cada grupo, para luego proyectar las UE originales sobre este nuevo espacio de ordenación (Yendle y MacFie 1989, Boulesteix 2005, Cardini *et al*. 2019). Debido a que este método también es un PCA, no tiene restricciones en cuanto al número de variables que se pueden utilizar. Sin embargo, debido a que las UE están representadas por grupos, el número de componentes no puede ser mayor a $g - 1$ (siempre y cuando el número de variables sea mayor al número de grupos). Por ejemplo, si hay sólo dos grupos y un centroide asociado a cada grupo, se requiere sólo una dimensión para separar ambos centroides. Sin embargo, cuando hay un menor número de grupos que de variables ($g < p$) el bgPCA puede sugerir la presencia de grupos no existentes, incluso para tamaños de muestra grandes (Cardini *et al*. 2019).



ESCALADO MULTIDIMENSIONAL NO MÉTRICO

El escalado multidimensional no métrico (NMDS) es una técnica desarrollada por Shepard (1962, 1966) y Kruskal (1964a, b) en el campo de la Psicología (Kruskal y Wish 1978). Al ser no métrico, tiene la ventaja de poder aplicarse a una MS basada en cualquier tipo de coeficiente. Al igual que el resto de los métodos de ordenación, el NMDS representa a las UE en un espacio de menor dimensión que la de la MBD original, pero utiliza el rango de los valores de la MS (Rohlf 1970, Sneath y Sokal 1973).

Hasta el momento en este capítulo hemos visto técnicas denominadas colectivamente escalado multidimensional métrico, ya que se basan en coeficientes de similitud que conservan las medidas de similitud originales para el análisis. A diferencia del PCA, CA o PCoA que son métodos basados en eigenvalores y eigenvectores, el NMDS no busca maximizar la variabilidad asociada a los ejes principales. Los ejes son arbitrarios, por lo que pueden ser rotados, centrados o invertidos. El objetivo del NMDS es representar la información de la MBD original en unas pocas dimensiones, de manera de poder visualizarla e interpretarla. Por lo tanto, no hay medidas de contribución a la variación de la MBD, sino medidas que comparan las distancias entre las UE de la MBD original *vs.* las distancias entre las UE en el espacio reducido (Legendre y Legendre 1998).

Los pasos del NMDS son los siguientes (Fig. 6.16):

1. A partir de la MBD (Fig. 6.16A) se calcula una MD (Fig. 6.16B) con valores observados D_{ij}. Si se utiliza una matriz con coeficientes de asociación o correlación, primero debe convertirse en una MD.
2. Especificar *a priori* el número de *m* dimensiones (generalmente dos) que se van a utilizar para representar la MBD (Fig. 6.16C). Si se desean varias configuraciones con distinto número de ejes, éstas deben calcularse por separado.
3. Construir una configuración inicial en *m* dimensiones (Fig. 6.16D). Ésta generalmente se construye de forma aleatoria y es crítica, por lo que se recomienda comenzar a partir de varias configuraciones diferentes.
4. Calcular una nueva MD con valores d_{ij} en el espacio de ordenación usando la distancia euclideana. En la primera iteración se calculan las distancias d_{ij} a partir de la configuración aleatoria.
5. Graficar las distancias d_{ij} entre las UE en el espacio de ordenación *vs.* las distancias observadas D_{ij} en la matriz de distancia original (Fig. 6.16F). Este gráfico se conoce como diagrama de Shepard (1962). Ambas distancias se relacionan mediante una regresión no paramétrica monótona (o isótona).
6. Calcular una medida de estrés (*stress*; Kruskal 1964a, b; Fig. 6.16G). Éste mide qué tan bien las distancias en la ordenación reflejan las distancias en la MD original. Para esto se comparan las distancias d_{ij} de la ordenación *vs.* las distancias esperadas por el modelo de regresión $\hat{d}_{ij}$. Las distancias esperadas se denominan disparidades o pseudo-distancias, ya que son valores que se ubican sobre la regresión y pueden considerarse una versión "suavizada" de la distancia en la ordenación (Kruskal 1977). La regresión garantiza que las disparidades tengan el mismo orden de rangos que las distancias observadas. Existen variantes para calcular el estrés, pero todas se basan en la suma de las diferencias al cuadrado entre distancia y disparidad (errores de representación). Por lo tanto, el estrés es la sumatoria de los errores de representación al cuadrado y es una medida del ajuste del NMDS:

$$\text{estrés} = \sum \left(d_{ij} - \hat{d}_{ij} \right)^2$$

Este valor luego se estandariza al rango 0–1. Idealmente, los valores de distancia en la ordenación deberían estar sobre el modelo de regresión. A menor valor de estrés, menor distorsión (Sneath y Sokal 1973). Como regla práctica se considera que valores de estrés menores a 0,05 brindan una excelente representación en el espacio de ordenación, valores entre 0,05 y 0,10 brindan una

muy buena representación, valores entre 0,10 y 0,20 brindan una representación buena, y valores mayores a 0,20 brindan una representación pobre (Clarke 1993).

7. Mejorar la configuración hacia donde disminuye el estrés. Esto se realiza mediante un método de optimización numérica denominado método del "descenso más pronunciado" (Press *et al.* 2007). La dirección del descenso más pronunciado es aquella que disminuye más rápido el estrés.
8. Repetir los pasos 4 a 7 hasta que el estrés ya no disminuya (convergencia).
9. La mayoría de los software rotan la solución final mediante PCA para una mejor interpretación.

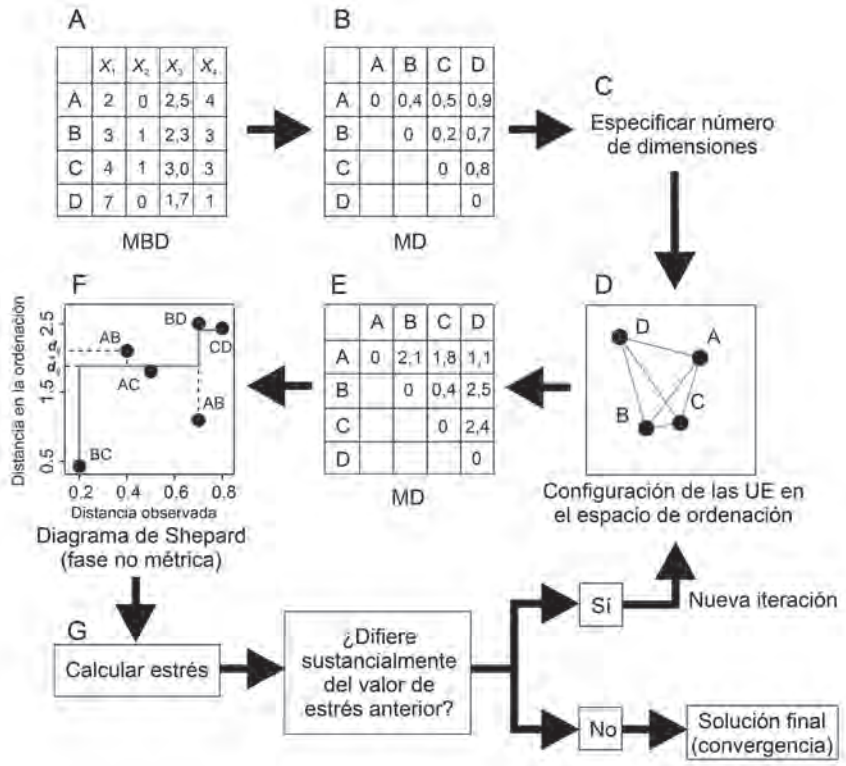


Fig. 6.16. NMDS. (A) MBD hipotética de cuatro UE (A a D) × cuatro variables (X_1 a X_4); (B) MD con distancias observadas D_{ij}. El NMDS toma como punto de partida esta última matriz; (C) se debe especificar de antemano el número de dimensiones, generalmente dos, para representar el espacio de ordenación; (D) las UE se sitúan sobre un espacio de ordenación con coordenadas aleatorias; a partir de esta configuración se calculan las distancias d_{ij} (líneas grises); (E) se construye una nueva MD; (F) se grafican las distancias observadas D_{ij} *vs.* las distancias en el espacio de ordenación d_{ij}, y se ajusta un modelo de regresión no paramétrico (línea escalonada), en un diagrama de Shepard; las líneas verticales punteadas corresponden a los errores de representación entre d_{ij} y disparidad; (G) se calcula una medida de estrés que cuantifica la suma de los errores de representación al cuadrado. Las UE se mueven hacia una nueva posición en el espacio de ordenación donde disminuye el estrés. Se vuelven a repetir los pasos E a G, hasta que el estrés llega a un mínimo.

COMBINANDO MÚLTIPLES TÉCNICAS MULTIVARIADAS: AGRUPAMIENTO JERÁRQUICO SOBRE COMPONENTES PRINCIPALES

Una opción interesante para explorar e identificar grupos, así como relaciones entre las UE y las variables, es combinar los tres métodos estándar del análisis multivariado (Kassambara 2017b): los métodos de ordenación (PCA, CA), los agrupamientos jerárquicos y los agrupamientos no jerárquicos (particularmente *K*-medias). Este enfoque se denomina agrupamiento jerárquico sobre componentes principales (HCPC) y es útil en al menos dos situaciones:

a. Cuando la MBD es grande y contiene múltiples variables con datos continuos (por ejemplo, datos de expresión génica). En esta situación se puede utilizar un PCA para reducir la matriz a un número menor de dimensiones que contenga la mayor parte de la información de la MBD. Luego se puede aplicar un análisis de agrupamientos (Cap. 5) sobre los *scores* de los primeros PCs. En este caso el PCA puede considerarse un método para eliminar el "ruido" de la MBD, y puede dar como resultado una solución más estable del agrupamiento.

b. Cuando la MBD presenta datos categóricos. En este caso se puede utilizar el CA para transformar los datos a unas pocas variables continuas, para luego aplicar un análisis de agrupamientos sobre estos ejes principales. Por lo tanto, el método puede considerarse como un pre-procesamiento de la MBD para aplicar análisis de agrupamientos a datos categóricos.

El método de HCPC puede resumirse en los siguientes pasos:

1. Aplicar un método de ordenación y extraer para cada UE los *scores* de los componentes o ejes principales.
2. Aplicar un análisis de agrupamiento jerárquico sobre la matriz de UE × PCs (Cap. 5).
3. Seleccionar el número de grupos de acuerdo al agrupamiento jerárquico.
4. De manera opcional, aplicar *K*-medias (ver Cap. 5) a la matriz de UE × PCs utilizando el número de grupos seleccionados en el paso 3 para mejorar el agrupamiento inicial (obtenido a partir del agrupamiento jerárquico).

Como ejemplo, aplicaremos el HCPC a la MBD de *Bulnesia*. El primer paso consiste en aplicar un método de ordenación (Fig 6.17A). Como anteriormente aplicamos el PCA a esta misma matriz, podemos considerar resuelto este paso y extraer los dos primeros PCs.

A continuación, debemos aplicar un análisis de agrupamiento jerárquico sobre los PCs, en nuestro caso utilizaremos el método de UPGMA (ver Cap. 5). Podemos graficar el dendrograma resultante aplicado a los PCs para poder identificar visualmente el número óptimo de grupos (Fig. 6.17B). A simple vista pueden identificarse claramente tres grupos. Por otra parte, es posible graficar las UE en el espacio de ordenación, e identificar las UE mediante colores que representen cada grupo (Fig. 6.17C). También se puede graficar este mismo espacio y superponer el agrupamiento jerárquico (Fig. 6.17D).

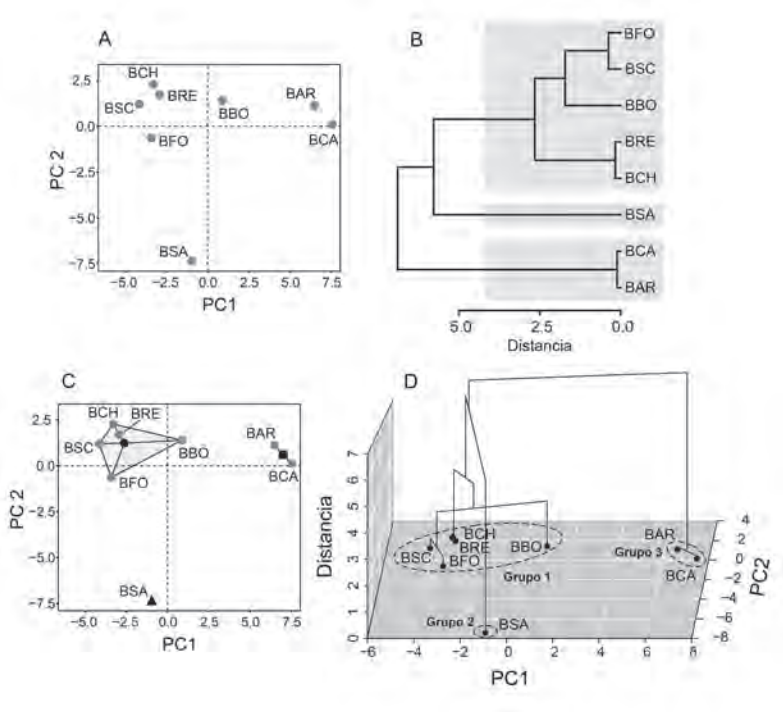


Fig. 6.17. HCPC aplicado a la MBD de *Bulnesia*. (A) PCA sobre la MBD; (B) Agrupamiento jerárquico (UPGMA y distancia euclideana) sobre los PC1 y PC2; (C) PCA donde se pueden distinguir los tres grupos derivados del agrupamiento jerárquico; (D) PC1 y PC2 junto con el agrupamiento jerárquico y los tres grupos de UE. BAR: *B. arborea*, BCA: *B. carrapo*, BCH: *B. chilensis*, BBO: *B. bonariensis*, BRE: *B. retama*, BFO: *B. foliosa*, BSC: *B. schickendantzii*, BSA: *B. sarmientoi*.

Uno de los aspectos interesantes de esta técnica es que también podemos obtener aquellas variables que describen mejor a cada grupo (Tabla 6.20). Para esto, se obtiene el promedio de una variable correspondiente al grupo en consideración, y se compara con la media global de dicha variable en la MBD. Así, por ejemplo, la media de la variable C38 (longitud del fruto) en el grupo 1 se encuentra muy por debajo de la media global (20,18 mm *vs.* 31,82 mm), lo que permite concluir que el grupo 1 se caracteriza por frutos relativamente pequeños. La misma lógica se aplica al resto de las variables. La prueba de *v* es una prueba estadística que, en términos generales, compara la media de un grupo con su respectiva media global y evalúa la probabilidad de obtener la diferencia observada o una mayor por azar (Husson *et al*. 2017). Valores menores a 0,05 corresponden a variables que están más asociadas con un determinado grupo; o sea, a menor valor de probabilidad, mayor asociación entre una variable y un grupo.

También es posible obtener mediante software aquellos componentes más asociados con cada grupo (Tabla 6.21). De esta tabla se deduce que el PC1 se encuentra asociado negativamente con el grupo 1 y positivamente con el grupo 3, mientras que el PC2 se encuentra asociado negativamente con el grupo 2. Observe que la media global es 0 por definición, ya que los componentes se centran en la media. A su vez, es posible extraer aquellas UE representativas o "modelos" de cada grupo. El término "modelo" hace referencia a aquellas UE que presentan valores de las variables cercanos al promedio (centroide) del grupo analizado. En nuestro ejemplo, las especies representativas del grupo 1 son *B. retama* (distancia al centroide = 2,74) y *B. foliosa* (distancia al centroide = 2,93), mientras que para los grupos 2 y 3 este análisis es trivial ya que en cada grupo hay sólo una y dos UE, respectivamente. Para un ejemplo aplicado a datos categóricos, ver en este capítulo *Técnicas de ordenación en R*.

Tabla 6.20. Variables que describen cada grupo, resultante del HCPC aplicado a la MBD de *Bulnesia*.

Grupo	Variable	Media del grupo	Media global	Prueba de *v*	*P*
1	C42	1,8	1,5	2,05	0,040
	C4	20,0	37,9	–1,96	0,050
	C16	1,0	1,3	–1,97	0,049
	C22	5,5	8,8	–1,98	0,047
	C6	3,4	4,8	–2,10	0,036
	C5	14,9	28,6	–2,19	0,029
	C41	1,5	3,2	–2,20	0,028
	C11	8,0	15,8	–2,25	0,025
	C12	3,3	6,7	–2,26	0,024
	C38	20,2	31,8	–2,36	0,018
	C39	18,0	28,8	–2,39	0,017
	C1	0,2	0,9	–2,49	0,013
2	C9	2,0	0,6	1,98	0,048
	C18	2,9	6,1	–2,33	0,020
	C34	0,0	1,6	–2,33	0,020
	C20	1,0	1,9	–2,65	0,008
3	C16	2,0	1,3	2,65	0,008
	C4	90,8	37,9	2,60	0,009
	C5	63,7	28,6	2,50	0,012
	C22	17,9	8,8	2,46	0,014
	C26	2,0	0,6	2,45	0,014
	C11	34,7	15,8	2,44	0,014
	C21	23,4	13,9	2,33	0,020
	C23	14,0	8,8	2,30	0,022

Grupo	Variable	Media del grupo	Media global	Prueba de *v*	*P*
	C6	8,4	4,8	2,26	0,024
	C25	1,5	0,5	2,16	0,031
	C17	17,7	11,7	2,14	0,032
	C32	5,3	4,2	1,97	0,049
	C14	0,5	1,5	–2,16	0,031

Tabla 6.21. PCs que describen cada grupo, resultante del HCPC aplicado a la MBD de *Bulnesia*.

Grupo	PC	Media del grupo	Media global	Prueba de *v*	*P*
1	1	–2,61	0,00	–2,06	0,039
2	2	–7,42	0,00	–2,53	0,012
3	1	7,01	0,00	2,48	0,013

Por último, es posible aplicar el método de *K*-medias al PCA, con el fin de evaluar si la pertenencia de las UE a los grupos establecidos por el agrupamiento jerárquico se sigue sosteniendo. En nuestro caso, si aplicamos *K*-medias (con número de grupos $K = 3$) al PCA el agrupamiento da el mismo resultado (resultados no mostrados en este libro).

RELACIÓN ENTRE LAS TÉCNICAS MULTIVARIADAS Y CRITERIOS PARA SELECCIONARLAS

Las técnicas de análisis de agrupamientos o de ordenación permiten obtener diferentes resultados gráficos y en muchos casos, difieren en las relaciones que sugieren. Por ejemplo, Rohlf (1972) ha demostrado que el PCA refleja con mayor fidelidad relaciones entre grupos formados a bajos niveles de similitud. En cambio es menos fiel en reflejar relaciones muy estrechas. El caso inverso se da con las técnicas del análisis de agrupamientos, donde las relaciones estrechas son distorsionadas en menor medida que las conexiones de baja similitud del dendrograma.

Esto nos permite concluir que en el procesamiento de los datos es aconsejable el uso de más de una técnica. En lo posible se debería utilizar una técnica de análisis de agrupamientos y una de ordenación, con el fin de minimizar los efectos metodológicos. Las conclusiones finales deben surgir de un complemento de diferentes técnicas.

Vale aclarar que una misma pregunta puede responderse utilizando más de una técnica. La gama de técnicas propuestas es tan amplia que el investigador tendrá siempre que afrontar el problema de seleccionar cuáles utilizar. En algunos casos se ha demostrado que algunas técnicas suelen ser más robustas (entendido como aquellas técnicas poco afectadas por valores atípicos) para ciertos tipos de datos. Por ejemplo, el NMDS es considerada una de las más robustas para el análisis de datos de comunidades ecológicas con respecto a las técnicas de escalado multidimensional métrico (Kenkel y Orlóci 1986, Minchin 1987, Legendre y Legendre 1998). En cambio, el escalado multidimensional métrico es muy útil para la identificación de taxones (Marramá y Kriwet 2017).

A modo de guía puede consultarse la Tabla 6.22, donde se resumen las técnicas en base a los coeficientes de similitud y tipo de dato que utilizan. Cabe resaltar que siempre debe haber una coherencia entre el objetivo del estudio y el resultado de las técnicas a aplicar. Por ejemplo, el escalado métrico (PCA, CA o PCoA) permite obtener el porcentaje de variación explicada por cada eje, así como las variables que más contribuyen a cada eje. Por el contrario, el escalado no métrico sólo permite visualizar e interpretar la relación entre las UE y las variables.

Tabla 6.22. Comparación de los métodos de ordenación. PCA: análisis de los componentes principales, CA: análisis de correspondencias, PCoA: análisis de coordenadas principales, DA: análisis discriminante, NMDS: escalado multidimensional no métrico.

Método	Tipo de escalamiento	Coeficiente de similitud	Tipo de técnica	Tipos de datos
PCA	Métrico	Euclideano	No supervisada	Cuantitativos
CA	Métrico	Chi-cuadrado	No supervisada	No negativos, cuantitativos o binarios. Todas las variables de una matriz deben ser del mismo tipo.
PCoA	Métrico	Cualquiera	No supervisada	Cuantitativos, cualitativos o ambos.
DA	Métrico	Mahalanobis	Supervisada	Cuantitativos y cualitativos (para los grupos *a priori*).
NMDS	No métrico	Cualquiera	No supervisada	Cuantitativos, cualitativos o ambos.

TÉCNICAS DE ORDENACIÓN EN R

Algunas de las técnicas de ordenación se encuentran disponibles cuando se instala R (paquete stats). Similar a lo que ocurre con el análisis de agrupamientos, hay varios paquetes que agregan una gran diversidad de métodos para representar y graficar los grupos. Algunos de estos se resumen en la Tabla 6.23. Para el PCA utilizaremos una MBD de caracteres de frutos de Tala (*Celtis tala*; Celtidaceae) de Palacio *et al.* (2014). Para el PCoA utilizaremos la MBD de especies de *Bulnesia* × caracteres. Para el CA y el NMDS utilizaremos una MBD de sitios × especies de aves de Palacio y Montalti (2013). Finalmente, para el HCPC utilizaremos una MBD de áreas de la Cuenca Inferior del Río de La Plata × 84 taxones de plantas y animales de Apodaca *et al*. (2019a).

Tabla 6.23. Algunas funciones y paquetes asociados a técnicas de ordenación en R.

Método de ordenación	Función	Paquete	Referencia
PCA	`princomp`	stats	R Core Team (2018)
	`prcomp`		
	`PCA`	FactoMineR	Lê *et al*. (2008)
	`dudi.pca`	ade4	Chessel *et al*. (2004)
	`pca`	labdsv	Roberts (2016)
CA	`ca`	ca	Nenadic y Greenacre (2007)
	`corresp`	MASS	Venables y Ripley (2002)
	`CA`	FactoMineR	Lê *et al*. (2008)
	`cca`	vegan	Oksanen *et al*. (2018)
	`dudi.coa`	ade4	Chessel *et al*. (2004)
	`epCA`	ExPosition	Beaton *et al.* (2014)
PCoA	`cmdscale`	stats	R Core Team (2018)
	`pcoa`	ape	Paradis y Schliep (2018)
	`dudi.pcoa`	ade4	Chessel *et al*. (2004)
	`wcmdscale`	vegan	Oksanen *et al*. (2018)
	`pco`	ecodist	Goslee y Urban (2007)
	`pco`	labdsv	Roberts (2016)

Método de ordenación	Función	Paquete	Referencia
NMDS	`metaMDS`	vegan	Oksanen *et al*. (2018)
	`isoMDS`	MASS	Venables y Ripley (2002)
	`nmds`	ecodist	Goslee y Urban (2007)
	`nmds`	labdsv	Roberts (2016)
Análisis discriminante	`lda`	MASS	Venables y Ripley (2002)
	`qda`		
	`discrimin`	ade4	Chessel *et al*. (2004)
	`mda`	mda	Leisch *et al.* (2017)
	`partimat`	klaR	Weihs *et al.* 2005
	`linDA`	DiscriMiner	Sánchez (2013)
	`quaDA`		
	`train`	caret	Kuhn (2019)

Análisis de componentes principales

Como ejemplo utilizaremos una MBD que contiene diversos caracteres del fruto (diámetro, concentración de azúcar, peso de la pulpa, peso de la semilla y relación peso de la pulpa-peso de la semilla) de *Celtis tala*. Esta es una MBD de 614 frutos × seis caracteres (Palacio *et al.* 2014). Como primera medida, y con fines exploratorios, calcularemos la matriz de correlación de Pearson entre variables (valores redondeados a tres dígitos).

```
> round(cor(Celtis), 3)
            diam     az  pulpa    sem pulpa.sem
diam       1.000 -0.330  0.993  0.450     0.560
az        -0.330  1.000 -0.373  0.200    -0.494
pulpa      0.993 -0.373  1.000  0.342     0.645
sem        0.450  0.200  0.342  1.000    -0.430
pulpa.sem  0.560 -0.494  0.645 -0.430     1.000
```

En la matriz de correlación puede observarse que el diámetro del fruto está fuertemente correlacionado y de forma positiva con el peso de la pulpa, y en menor medida, con la relación peso de la pulpa-peso de la semilla. Esta asociación implica que las variables contienen información redundante, y debería expresarse en alguno de los componentes principales como asociadas en signo y magnitud.

Para realizar el PCA usaremos la función `PCA()` del paquete FactoMineR (Lê *et al.* 2008). Si bien el paquete stats ofrece funciones para realizar un PCA –`princomp()` y `prcomp()`–, el paquete FactoMineR ofrece una gran cantidad de funciones adicionales. Hay que tener en cuenta que para realizar el PCA aplicado a las variables estandarizadas debemos trabajar con la matriz de correlación, que se logra con el argumento `scale.unit = TRUE`. Esto es útil cuando las variables están medidas en diferentes escalas, lo que hace posible su comparación.

```
> library(FactoMineR)
> Celtis <- read.table("C:/R datos/Celtis.txt", header = TRUE)
> pca <- PCA(Celtis, scale.unit = TRUE)
```

Como se puede observar, el PCA es muy simple de realizar. A continuación, calcularemos los eigenvalores (variación explicada) de cada componente principal, así como sus porcentajes acumulados.

```
> round(pca$eig,2)
    eigenvalue percentage of variance cumulative percentage of variance
comp 1    2.76                 55.30                             55.30
comp 2    1.58                 31.54                             86.84
comp 3    0.61                 12.19                             99.02
comp 4    0.05                  0.98                            100.00
comp 5    0.00                  0.00                            100.00
```

Si un eigenvalor es mayor a 1, puede considerarse que el componente principal explica mayor variación que una única variable. Por este motivo, uno de los criterios de retención del número de componentes es tomar eigenvalores mayores a 1 (criterio de Kaiser-Guttman). Así, los dos primeros componentes representan casi el 87% de la variación total en la MBD. Observe que hay tantos componentes como variables, que en conjunto representan el 100% de la variación en los datos. Dicho de otra forma, utilizar los cinco componentes equivale a utilizar la MBD original. Así, el objetivo del PCA es representar este conjunto de información (cinco variables en el conjunto de datos) con un número menor de variables, pero que expliquen el mayor porcentaje de variación total de los datos (en este caso, tomando sólo dos componentes representamos una buena cantidad de la variación total de la MBD).

El siguiente paso consiste en analizar los denominados *loadings*, que corresponden al coeficiente de correlación entre una variable y su respectivo componente principal. Para esto accedemos a la información de las variables y luego a sus coordenadas.

```
> round(pca$var$coord,3)
           Dim.1  Dim.2 Dim.3  Dim.4  Dim.5
diam       0.939  0.328 0.085 -0.063 -0.002
az        -0.575  0.465 0.673 -0.004  0.000
pulpa      0.965  0.220 0.117 -0.082  0.002
sem        0.174  0.953 -0.213 0.127  0.000
pulpa.sem  0.769 -0.544  0.300 0.148  0.000
```

En este caso, vemos que el PC1 tiene una asociación fuerte y positiva con el diámetro, el peso de la pulpa y la relación peso de la pulpa-peso de la semilla, y, en menor medida, una asociación negativa con la concentración de azúcar. Esto es consistente con el análisis de la matriz de correlación. Por lo tanto, el PC1 puede interpretarse como un eje de variación en el tamaño y concentración de azúcar del fruto. El PC2 tiene un *loading* alto y positivo para el peso de la semilla.

La decisión de cuántos componentes a utilizar no es trivial y queda a juicio del investigador. Además del criterio de Kaiser-Guttman, otro método consiste en utilizar el gráfico de sedimentación (*scree plot*), que muestra el porcentaje de la varianza total (eigvenvalores) de los datos explicado por cada componente (Fig. 6.18). Para gráficos vinculados al PCA utilizaremos el paquete factoextra (Kassambara y Mundt 2017).

```
> library(factoextra)
> fviz_eig(pca, barfill = "gray70", barcolor = "black")
```

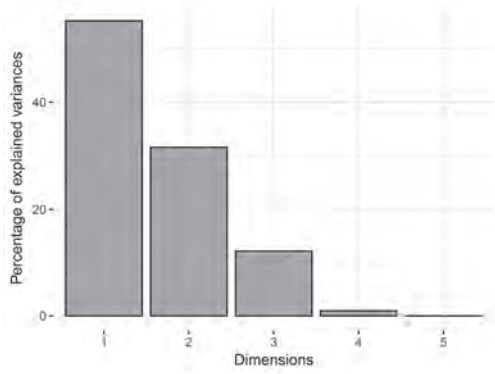


Fig. 6.18. *Scree plot* del PCA correspondiente a la matriz de *C. tala*.

El punto donde se produce una estabilización de la varianza explicada (“codo”) indica el número de componentes que se deben retener. En nuestro ejemplo los valores de la varianza explicada no cambian demasiado del PC4 al PC5, por lo que utilizando cuatro componentes sería suficiente para su interpretación. Este criterio junto con el anterior es arbitrario, y a veces suele arrojar un número irracionalmente alto de componentes.

En cuanto a las variables, podemos analizar la calidad de la representación y su contribución a los componentes, y graficar el círculo de correlación.

```
> round(pca$var$cos2, 3)
          Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
diam      0.881 0.108 0.007 0.004     0
az        0.331 0.217 0.453 0.000     0
pulpa     0.931 0.049 0.014 0.007     0
sem       0.030 0.908 0.046 0.016     0
pulpa.sem 0.592 0.296 0.090 0.022     0

> round(pca$var$contrib, 3)
           Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
diam      31.874  6.823  1.184  7.998 52.120
az        11.966 13.736 74.271  0.027  0.000
pulpa     33.667  3.080  2.258 13.924 47.071
sem        1.093 57.587  7.471 33.040  0.809
pulpa.sem 21.400 18.774 14.817 45.010  0.000

> fviz_pca_var(pca, col.var = "black")
```

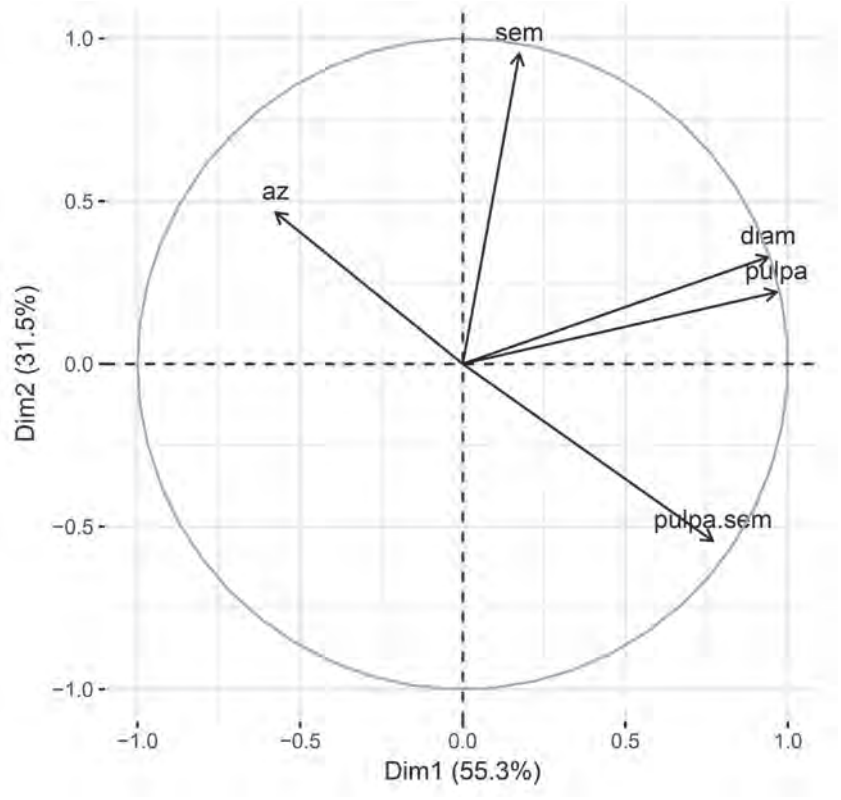


Fig. 6.19. Círculo de correlación aplicado a la MBD de *C.* tala. az: concentración de azúcar, diam: diámetro del fruto, pulpa: peso de la pulpa, pulpa.sem: relación peso de la pulpa-peso de la semilla, sem: peso de la semilla.

Las variables diámetro y peso de la pulpa son las mejor representadas y que más contribuyen al PC1, mientras que el peso de la semilla es la más pobremente representada en el PC1 y contribuye más al PC2 (Fig. 6.19). Para obtener los valores de calidad de la representación y de la contribución de las UE (de posible interés para el lector, pero no mostrados aquí debido a su extensión), debemos extraer el objeto `ind` del objeto `pca` (`pca$ind$cos2` y `pca$ind$contrib`, respectivamente).

Por último realizaremos el *biplot* con la `función fviz_pca_biplot()`, donde se visualizan simultáneamente las UE (frutos) y las variables (caracteres) sobre el espacio de los componentes principales (PC1 y PC2 en este caso; Fig. 6.20).

```
> fviz_pca_biplot(pca, col.ind = "gray70", col.var = "black")
```

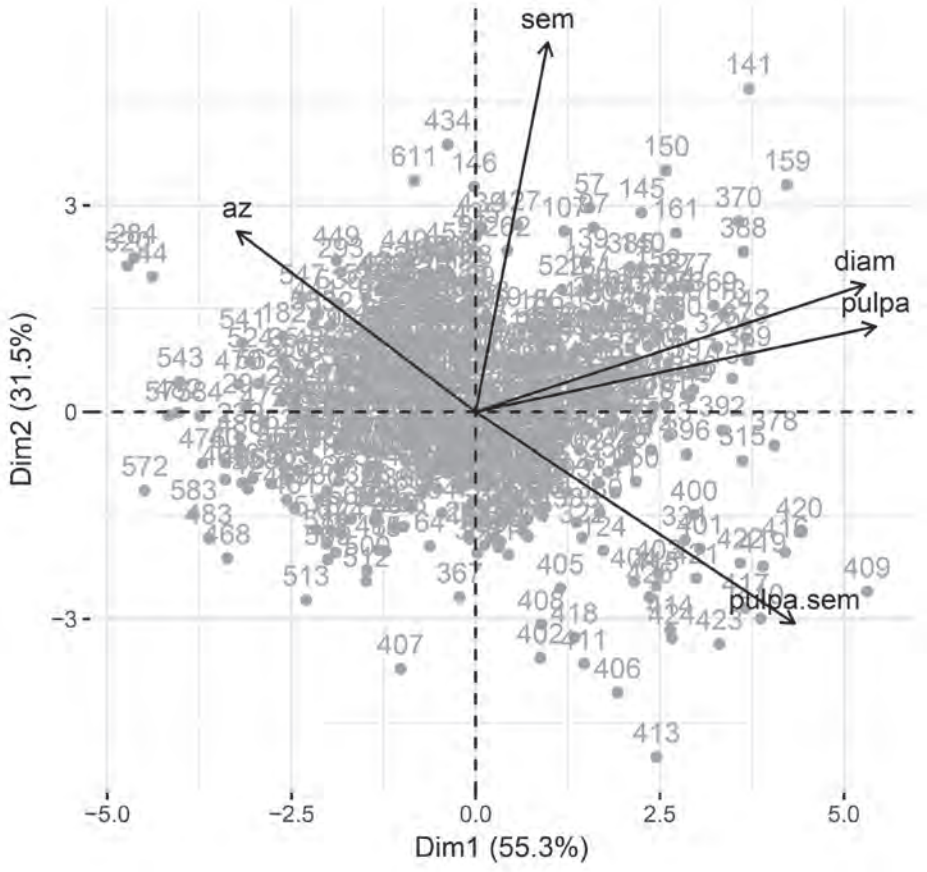


Fig. 6.20. *Biplot* de los PC1 y PC2 correspondiente a la MBD de *C. tala*. az: concentración de azúcar, diam: diámetro del fruto, pulpa: peso de la pulpa, pulpa.sem: relación peso de la pulpa-peso de la semilla, sem: peso de la semilla.

Las UE se muestran en gris para una mejor visualización. Los números son etiquetas que coinciden con el orden dado en la MBD, mientras que las flechas corresponden a las variables (representadas como vectores). La posición de cada UE (fruto) en el espacio de componentes principales está dada por sus coordenadas, denominadas *scores*.

A modo de recordatorio, las reglas de interpretación de un *biplot* son las siguientes: (1) las UE cercanas en el espacio tienen características similares en cuanto a sus variables; (2) el coseno del ángulo entre dos vectores corresponde a su correlación. Por lo tanto, vectores cercanos indican una alta correlación positiva, vectores diametralmente opuestos indican una alta correlación negativa, y vectores perpendiculares indican una correlación nula; (3) el coseno del ángulo entre un vector y un componente principal corresponde a su correlación; (4) una UE en la dirección y sentido de una variable y alejada del centro de origen tiene un valor alto para esa variable. Así, por ejemplo, el diámetro del fruto y el peso de la pulpa se encuentran muy correlacionados entre sí y de forma positiva, mientras que ambos se encuentran levemente correlacionados

y de forma negativa con la concentración de azúcar. Asimismo, el diámetro del fruto y el peso de la pulpa son las variables que se encuentran más correlacionadas con el PC1 y también las que más contribuyen a este componente, mientras que el peso de la semilla se encuentra casi paralelo al PC2, y es también la que más contribuye a este último. De esta manera, el *biplot* nos permite interpretar la relación entre las variables y las UE de forma simultánea. Por ejemplo, aquellos frutos que se encuentran hacia el extremo derecho de la ordenación son de gran tamaño y bajo contenido de azúcar (por ejemplo, etiquetas 378, 370 y 159), mientras que los frutos que se encuentran del lado izquierdo son pequeños con alta concentración de azúcar (por ejemplo, etiquetas 541, 543 y 572). Por otra parte, los frutos que se encuentran sobre el extremo superior tienen semillas relativamente grandes (por ejemplo, etiquetas 146, 434 y 611), contrario a lo que sucede con los frutos que se encuentran en el extremo inferior (por ejemplo, etiquetas 406, 402 y 367). Los frutos que se ubican en el centro de la ordenación tienen valores promedios de las variables analizadas. Si bien estos casos son extremos, cada componente define un gradiente de variación del tamaño del fruto/concentración de azúcar (PC1) y del tamaño de la semilla (PC2).

Vale aclarar que aunque usemos la misma matriz el usuario puede obtener un gráfico diferente, pero en el que se mantienen las mismas relaciones entre las variables y las UE, resultado de la rotación arbitraria de sus ejes. Por último, si se quieren extraer los *scores* de las UE (por ejemplo, para personalizar los gráficos), éstos pueden obtenerse a partir del objeto creado. A modo de ejemplo, se extraerán los *scores* de los primeros tres componentes principales y luego se visualizarán los primeros 10 *scores*.

```
> scores <- pca$ind$coord[, 1:3]
> scores[1:10, ]
         Dim.1      Dim.2       Dim.3
1   1.43563251  0.2043206 -0.63443773
2  -0.43269278 -0.7970755 -0.11120705
3   0.93469498  0.4557096 -0.94186752
4  -0.15783387  0.5338784 -0.01066021
5  -1.31755766 -0.8240601 -1.85633229
6  -0.64607196  0.6999834 -0.55621604
7  -0.27892382 -0.1241703  0.01219352
8  -0.06820373  1.0126248 -0.31893027
9  -0.97288626 -0.6536408 -0.66999537
10 -0.26468680 -0.7732862 -0.68525030
```

Con esta información podemos realizar un gráfico en tres dimensiones. Hay varios paquetes para realizar gráficos 3D, en este caso utilizaremos el paquete plot3D (Soetaert 2017). Definiremos tres ejes (*x*, *y*, *z*) que representan los tres primeros componentes y graficaremos las UE con la función `points3D()`.

```
> library(plot3D)
> x <- scores[, 1]
> y <- scores[, 2]
> z <- scores[, 3]
> points3D(x , y, z, pch = 19, cex = 1, alpha = 0.2, bty = "g",
+          colkey = FALSE, theta = -60, phi = 20, col = "darkorange",
+          xlab = "PC 1", ylab = "PC 2", zlab = "PC 3", ticktype = "detailed")
```

Hemos agregado transparencia a las UE (argumento `alpha`) debido al gran número de observaciones que dificultaría su visualización. Los ángulos de perspectiva pueden variarse con los argumentos `theta` (plano horizontal) y `phi` (plano vertical). A continuación, extraemos los *loadings* para agregar las variables al espacio de ordenación (recuerde que estos no están ubicados en el mismo espacio de las UE).

```
> loadings <- pca$var$coord[, 1:3]
```

Agregamos las etiquetas de las variables –función `text3D()`–, extraemos las coordenadas de los *loadings*, y luego graficamos los vectores con la función `arrows3D()`. Para utilizar esta función es necesario especificar sus coordenadas de inicio (x_0, y_0, z_0) y terminación (x_1, y_1, z_1). En el PCA las coordenadas de inicio corresponden al origen (0, 0, 0), mientras que las coordenadas de terminación se encuentran en el objeto `loadings`. Por último, multiplicamos los *loadings* por una constante para mejorar la visualización de los vectores (en este caso elegiremos el valor de 4; Fig. 6.21).

```
> text3D(x = 4*loadings[, 1] + 0.4, y = 4*loadings[, 2] + 0.4,
+        z = 4*loadings[, 3] + 0.4, labels = rownames(loadings),
+        col = "blue", cex = 0.8, add = TRUE)
> arrows3D(x0 = rep(0, nrow(loadings)), y0 = rep(0, nrow(loadings)), z0 =
+          rep(0, nrow(loadings)), x1 = 4*loadings[, 1],
+          y1 = 4*loadings[,2], z1 = 4*loadings[, 3],
+          col = "blue", lwd = 1, add = TRUE)
```

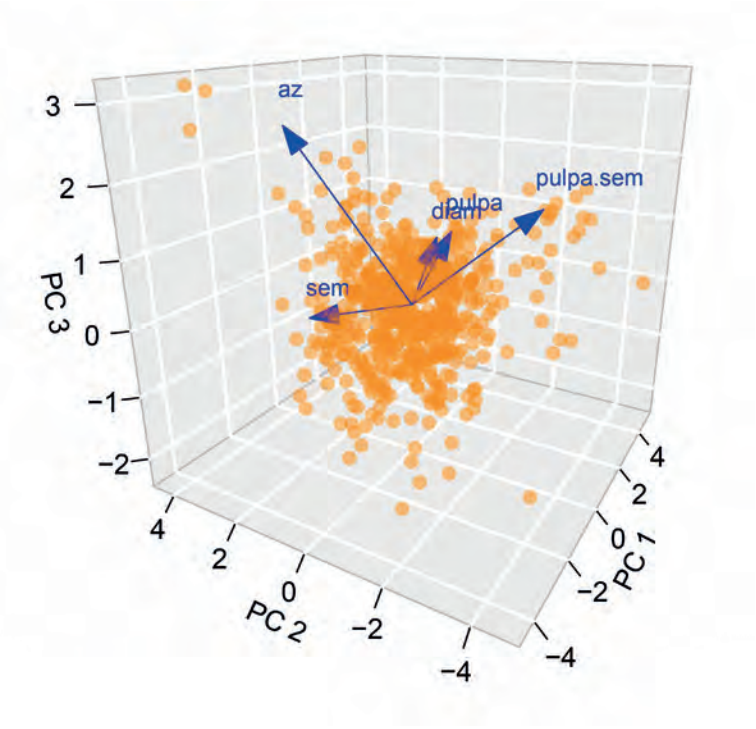


Fig. 6.21. *Biplot* en 3D correspondiente al PC1, PC2 y PC3 aplicado a la MBD de *C. tala*. az: concentración de azúcar, diam: diámetro del fruto, pulpa: peso de la pulpa, pulpa.sem: relación peso de la pulpa-peso de la semilla, sem: peso de la semilla.

Análisis de correspondencias

Utilizaremos una MBD que consiste en datos de abundancia de aves, obtenidos en nueve puntos de conteo de 30 metros de radio (sitios) a lo largo de un año en la provincia de Buenos Aires (Palacio y Montalti 2013). La matriz también contiene datos de la estación del año (otoño, invierno, primavera y verano) y del tipo de ambiente (arbustal y bosque) para cada sitio.

```
> Aves <- read.table("C:/R datos/Aves.txt", header = TRUE)
```

En esta MBD las especies son las variables (columnas), mientras que las UE son los sitios (filas). Así, la MBD contiene datos tomados en la misma escala, es decir, conteos de individuos. También con-

tiene datos no negativos, por lo que es apropiado utilizar un CA en lugar de un PCA. Para realizar el CA vamos a utilizar el paquete FactoMineR (Lê *et al.* 2008) junto con el paquete factoextra (Kassambara y Mundt 2017). Para esto, necesitamos que la MBD contenga solamente los datos de los conteos, excluyendo cualquier otra variable (36 UE × 32 variables).

```
> library(FactoMineR)
> library(factoextra)
> comm <- Aves[, -c(1:3)]
```

Previamente, etiquetaremos las filas de la MBD para que contengan el sitio y la estación del año.

```
> rownames(comm) <- paste(Aves$sitio, Aves$estacion, sep = ".")
```

Por el momento no graficaremos la ordenación (argumento `graph = FALSE`).

```
> ca <- CA(comm, graph = FALSE)
```

La función `get_eigenvalue()` arroja los eigenvalores (inercias principales) asociados a cada eje, la proporción de variación explicada por cada eje y su acumulado.

```
> round(get_eigenvalue(ca), 3)
       eigenvalue variance.percent cumulative.variance.percent
Dim.1       0.432           13.967                      13.967
Dim.2       0.404           13.076                      27.043
Dim.3       0.320           10.363                      37.406
Dim.4       0.295            9.553                      46.959
Dim.5       0.214            6.923                      53.882
Dim.6       0.188            6.066                      59.948
Dim.7       0.160            5.159                      65.107
Dim.8       0.139            4.489                      69.596
Dim.9       0.127            4.122                      73.718
Dim.10      0.121            3.900                      77.618
Dim.11      0.091            2.934                      80.552
Dim.12      0.086            2.766                      83.318
Dim.13      0.082            2.644                      85.961
Dim.14      0.072            2.330                      88.291
Dim.15      0.063            2.047                      90.338
Dim.16      0.059            1.904                      92.242
Dim.17      0.046            1.484                      93.725
Dim.18      0.039            1.265                      94.990
Dim.19      0.037            1.185                      96.175
Dim.20      0.026            0.853                      97.029
Dim.21      0.022            0.696                      97.725
Dim.22      0.017            0.544                      98.269
Dim.23      0.013            0.432                      98.701
Dim.24      0.011            0.367                      99.068
Dim.25      0.008            0.268                      99.337
Dim.26      0.007            0.240                      99.577
Dim.27      0.005            0.173                      99.750
Dim.28      0.004            0.142                      99.892
```

```
Dim.29   0.002          0.055                    99.947
Dim.30   0.001          0.044                    99.991
Dim.31   0.000          0.009                   100.000
```

Como es común en el estudio de comunidades biológicas (MBD de sitios × especies), los primeros ejes suelen explicar poca variación de la MBD, debido a la heterogeneidad inherente a dichas comunidades (Gauch 1982). De hecho, los primeros tres ejes acumulan sólo el 37% de la variación total de la MBD.

Recuerde que en el CA la suma de todos los eigenvalores es igual a la inercia total.

```
> inercia <- sum(get_eigenvalue(ca)[, 1])
> inercia
[1] 3.092511
```

Este valor representa la asociación o grado de dependencia entre las UE y las variables, o dicho de otra forma, cuánto se desvía la MBD de la independencia. La inercia total máxima corresponde al mínimo del número de filas – 1 o columnas – 1 (en nuestro ejemplo, 31). Dado que la inercia total es mucho menor que la inercia máxima, podemos decir que hay poca dependencia entre las UE y las variables. Calculando la *V* de Cramér, el CA captura el 100%×3,093/31 = 9,97% de la inercia máxima.

```
> V <- 100*inercia/min(dim(comm) - 1)
> V
[1] 9.975843
```

Valores de *V* entre 0 y 50% se consideran como independencia entre las filas y las columnas, mientras que valores mayores a 50% se consideran como dependencia entre las mismas. Mediante una prueba de chi-cuadrado podemos calcular estadísticamente si la asociación entre las filas y las columnas es significativa. Debido a que la inercia total es el cociente entre el valor de chi-cuadrado y el número de observaciones *N*, podemos obtener el valor de chi-cuadrado como el producto de la inercia por *N*. Con esta información calculamos la probabilidad de obtener un valor de inercia mayor o igual al esperado por el azar.

```
> chi <- inercia*sum(comm)
> gl <- (nrow(comm) - 1)*(ncol(comm) - 1)
> pchisq(chi, df = gl, lower.tail = FALSE)
[1] 9.927702e-197
```

Si bien la inercia es baja, obtener este valor o uno mayor por azar es altamente improbable (influenciado por el alto tamaño muestral), lo que indica que hay una asociación entre filas y columnas.

A continuación, podemos realizar un gráfico de sedimentación para determinar el número de ejes a analizar (Fig. 6.22). El gráfico sugiere que se requieren al menos cinco o seis ejes para obtener un buen porcentaje de variación presente en la MBD.

```
> fviz_screeplot(ca, barcolor = "black", barfill = "gray70")
```

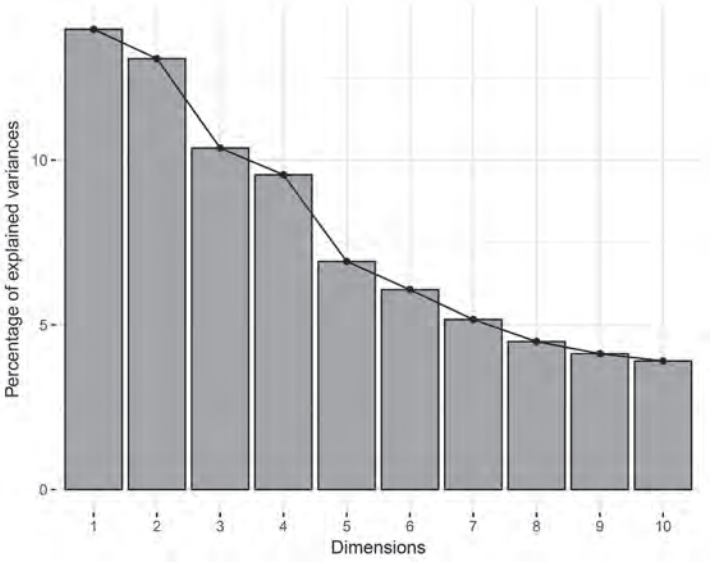


Fig. 6.22. Gráfico de sedimentación de eigenvalores *vs.* los componentes de la MBD de sitios × especies de aves.

Al igual que en el PCA, podemos obtener las coordenadas de las especies y de los sitios (*scores*), así como la calidad de la representación (cos^2), y la contribución de las UE (filas) y de las variables (columnas). Para esto hay que extraer la información contenida en los objetos `ca$row` (filas) `ca$col` (columnas). Debido a su gran extensión sólo se mostrarán los primeros valores.

```
> head(ca$row$coord)
           Dim 1       Dim 2       Dim 3       Dim 4        Dim 5
2.inv   0.1175565 -0.18615177   0.4514406 -0.41914090   0.27690123
2.oto   0.5242250 -0.28445679   0.7139074 -0.66749868   0.38867840
2.pri   0.2682785 -0.04193224  -0.3950545  0.17500578  -0.02476545
2.ver   0.3852398 -0.21980493   0.2946362 -0.01422871   0.32870536
3.inv  -0.2098920 -0.16063202   0.3267361 -0.31661466   0.46862395
3.oto   0.1648216 -0.33535092   0.6417482 -0.45139885   0.48768866

> head(ca$row$cos2)
           Dim 1        Dim 2       Dim 3        Dim 4        Dim 5
2.inv  0.01693534 0.0424653683  0.24974785 0.2152883930 0.0939615757
2.oto  0.06380838 0.0187877573  0.11833850 0.1034530164 0.0350770569
2.pri  0.03014364 0.0007364105  0.06536386 0.0128271019 0.0002568717
2.ver  0.08364288 0.0272296125  0.04892590 0.0001141032 0.0608947929
3.inv  0.02914090 0.0170677167  0.07061639 0.0663091115 0.1452647112
3.oto  0.01036928 0.0429259314  0.15719908 0.0777752862 0.0907833138

> head(ca$row$contrib)
          Dim 1       Dim 2       Dim 3      Dim 4     Dim 5
2.inv 0.1075956 0.288198678   2.1385563 1.99993187 1.2043419
2.oto 1.7620396 0.554203598   4.4043550 4.17710173 1.9541594
2.pri 0.2637024 0.006881687   0.7706807 0.16407446 0.0045335
2.ver 1.1894677 0.413639134   0.9377365 0.00237255 1.7470396
3.inv 0.3732637 0.233531079   1.2190923 1.24188095 3.7538059
3.oto 0.1741840 0.770257153   3.5589989 1.91027164 3.0765535
```

```
> head(ca$col$coord)
            Dim 1        Dim 2       Dim 3      Dim 4       Dim 5
agebad  0.7189047   1.40247784  -1.1616143 -0.7630765 -0.03032420
amabra  1.8440297  -0.92450284  -1.0724535  3.0311286  2.14953908
rupmag  1.1737891  -0.27873241  -0.3706709  0.4953773  0.09343289
spimag -0.5800426   2.33336739   1.1798562  1.4288668 -0.39723112
chlluc  0.6442459  -0.42117464  -0.2510214  0.5310065  0.38343096
colmel  0.3274249  -0.03443434   0.2815567 -0.3197943  0.38333092

> head(ca$col$cos2)
            Dim 1        Dim 2       Dim 3      Dim 4       Dim 5
agebad 0.11520443 0.4384484510 0.300781021 0.12979640 0.000204977
amabra 0.15373466 0.0386413674 0.051998756 0.41537917 0.208894344
rupmag 0.31551972 0.0177918577 0.031464684 0.05619770 0.001999153
spimag 0.03209235 0.5193356103 0.132782241 0.19474456 0.015051105
chlluc 0.03831201 0.0163740484 0.005816386 0.02602743 0.013570817
colmel 0.06714785 0.0007426634 0.049652410 0.06405456 0.092035689

> head(ca$col$contrib)
           Dim 1        Dim 2       Dim 3      Dim 4       Dim 5
agebad 8.5211337 34.642020803 29.98450969 14.0373397  0.03058667
amabra 3.8933911  1.045358385  1.77487395 15.3813668 10.67288507
rupmag 6.6255461  0.399091926  0.89050602  1.7254712  0.08469154
spimag 1.0015792 17.313636896  5.58524259  8.8867513  0.94765666
chlluc 0.4752209  0.216956732  0.09723712  0.4720477  0.33959829
colmel 0.7119406  0.008411242  0.70952925  0.9930148  1.96864236
```

Esta información es más fácil de visualizar en un gráfico *biplot* con las UE y las variables (Fig. 6.23).

```
> fviz_ca_biplot(ca, map = "symbiplot", col.row = "black",
+                col.col = "black", repel = TRUE)
```

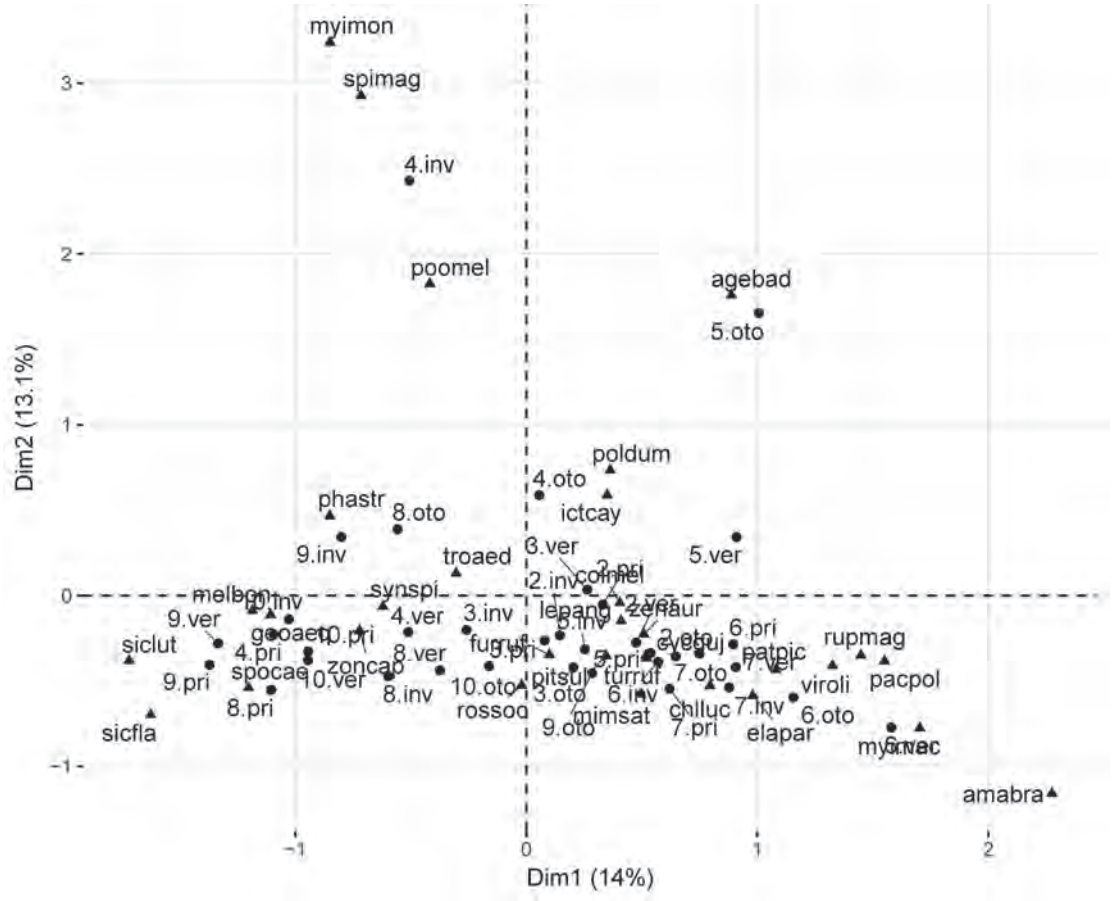


Fig. 6.23. *Biplot* simétrico de la MBD de sitios (círculos) × especies de aves (triángulos). Agebad: *Agelaioides badius*, amabra: *Amazonetta brasiliensis*, rupmag: *Rupornis magnirostris*, spimag: *Spinus magellanicus*, chlluc: *Chlorostilbon lucidus*, colmel: *Colaptes melanochloros*, patpic: *Patagioenas picazuro*, cycguj: *Cyclarhis gujanensis*, elapar: *Elaenia parvirostris*, furruf: *Furnarius rufus*, geoaeq: *Geothlypis aequinoctialis*, ictcay: *Icterus pyrrhopterus*, lepang: *Lepidocolaptes angustirostris*, mimsat: *Mimus saturninus*, molbon: *Molothrus bonariensis*, myimac: *Myiodynastes maculatus*, myimon: *Myiopsitta monachus*, pacpol: *Pachyramphus polychopterus*, phastr: *Phacellodomus striaticollis*, pitsul: *Pitangus sulphuratus*, poldum: *Polioptila dumicola*, poomel: *Poospiza melanoleuca*, rossoc: *Rostrhamus sociabilis*, sicfla: *Sicalis flaveola*, siclut: *S. luteola*, spocae: *Sporophila caerulescens*, synspi: *Synallaxis spixi*, troaed: *Troglodytes aedon*, turruf: *Turdus rufiventris*, viroli: *Vireo olivaceus*, zenaur: *Zenaida auriculata*, zoncap: *Zonotrichia capensis*.

El gráfico por defecto en el CA es el *biplot* simétrico (argumento `map`) donde las relaciones entre las UE y entre las variables se deben interpretar en forma separada (entre una UE y otra UE, o entre una variable y otra variable, no entre una UE y una variable). Las reglas de interpretación son similares a las de otros métodos de ordenación: (1) las UE (perfiles fila) cercanas en el espacio tienen características similares en cuanto a sus variables. Para el ejemplo, aquellos sitios cercanos en el espacio de los ejes tienen una proporción de especies (composición de especies) similar; y (2) las variables (perfiles columna) cercanas en el espacio tienen características similares en cuanto a las UE en las que aparecen. Para el ejemplo, aquellas especies cercanas en el espacio de los ejes tienden a aparecer en conjunto.

Alternativamente, podemos graficar *biplots* asimétricos en los cuales los perfiles columna (variables) pueden representarse en el espacio de las UE, o viceversa (Fig. 6.24). En este sentido es conveniente mostrar las UE y las variables como vectores (argumento `arrow`). Si el ángulo entre dos vectores es agudo, significa que hay una fuerte asociación entre la UE y la variable correspondiente.

```
> fviz_ca_biplot(ca, map = "rowprincipal", col.row = "black",
+                col.col = "black", repel = TRUE, arrow = c(TRUE, TRUE))
```

```
> fviz_ca_biplot(ca, map = "colprincipal", col.row = "black",
+                col.col = "black", repel = TRUE, arrow = c(TRUE, TRUE))
```

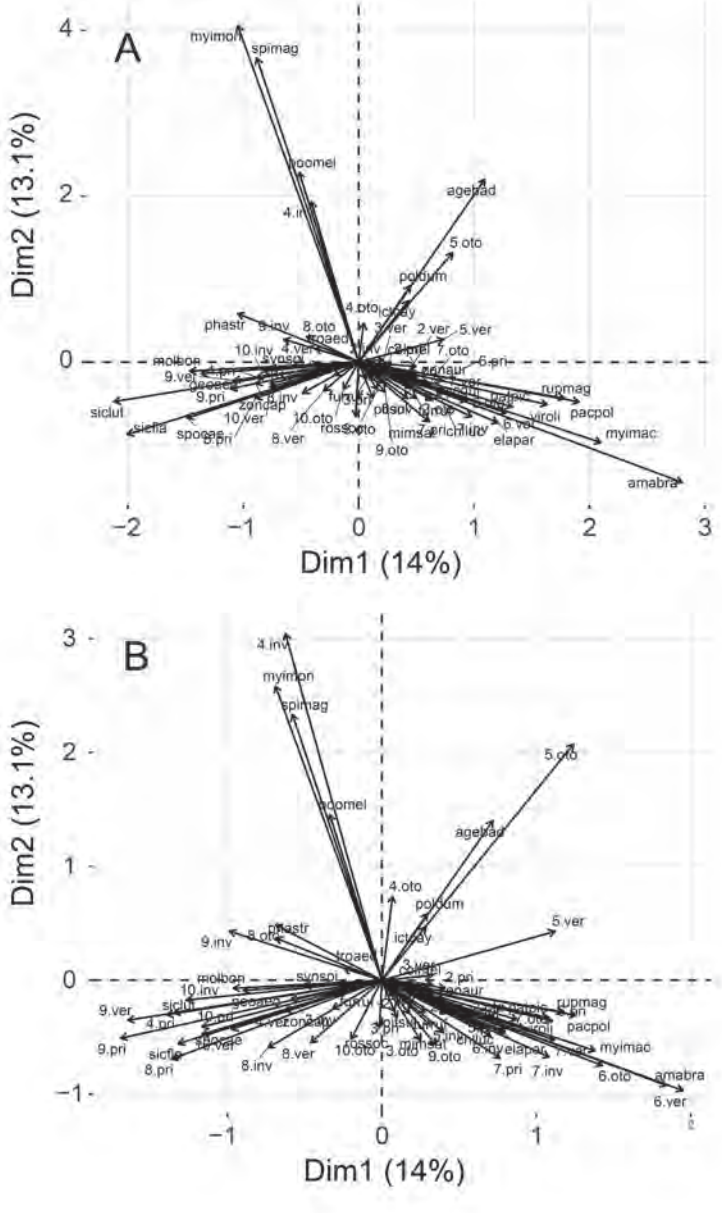


Fig. 6.24. *Biplots* asimétricos de la MBD de sitios × especies de aves. (A) *Biplot* asimétrico de las especies en el espacio de las UE; (B) *biplot* asimétrico de las UE en el espacio de las especies. Agebad: *Agelaioides badius*, amabra: *Amazonetta brasiliensis*, rupmag: *Rupornis magnirostris*, spimag: *Spinus magellanicus*, chlluc: *Chlorostilbon lucidus*, colmel: *Colaptes melanochloros*, patpic: *Patagioenas* picazuro, cycguj: *Cyclarhis gujanensis*, elapar: *Elaenia parvirostris*, furruf: *Furnarius rufus*, geoaeq: *Geothlypis aequinoctialis*, ictcay: *Icterus pyrrhopterus*, lepang: *Lepidocolaptes angustirostris*, mimsat: *Mimus saturninus*, molbon: *Molothrus bonariensis*, myimac: *Myiodynastes maculatus*, myimon: *Myiopsitta monachus*, pacpol: *Pachyramphus polychopterus*, phastr: *Phacellodomus striaticollis*, pitsul: *Pitangus sulphuratus*, poldum: *Polioptila dumicola*, poomel: *Poospiza melanoleuca*, rossoc: *Rostrhamus sociabilis*, sicfla: *Sicalis flaveola*, siclut: *S. luteola*, spocae: *Sporophila caerulescens*, synspi: *Synallaxis spixi*, troaed: *Troglodytes aedon*, turruf: *Turdus rufiventris*, viroli: *Vireo olivaceus*, zenaur: *Zenaida auriculata*, zoncap: *Zonotrichia capensis*.

Por ejemplo, el punto 4 en invierno mostró una alta predominancia de Cotorra Común (*Myiopsitta monachus*), Cabecitanegra Común (*Spinus magellanicus*) y Monterita Cabeza Negra (*Poospiza melanoleuca*) (cuadrante superior izquierdo, etiquetas myimon, spimag y poomel, respectivamente).

Un aspecto interesante para analizar cuando se tienen factores adicionales que no intervienen en la configuración del análisis (variables suplementarias), es explorar si se observa algún patrón de agrupamiento en cuanto a las UE. Por ejemplo, podemos explorar si las UE se agrupan de acuerdo a la estación del año o al tipo de ambiente.

```
> fviz_ca_row(ca, col.row = "black", geom.row = "point", pointsize = 3,
+             shape.row = Aves$estacion)
> fviz_ca_row(ca, col.row = "black", geom.row = "point", pointsize = 3,
+             shape.row = Aves$ambiente)
```

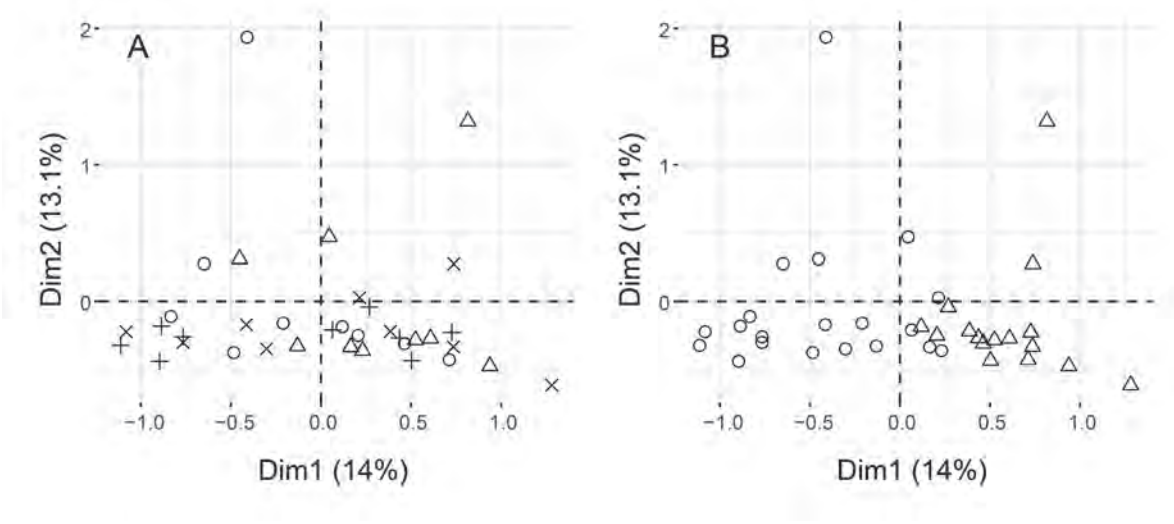


Fig. 6.25. Gráfico de correspondencias de la MBD de sitios × especies de aves. (A) Sitios agrupados por estación del año (signos de suma = verano, triángulos = otoño, círculos = invierno, signos de multiplicación = primavera); (B) sitios agrupados por tipo de ambiente (círculos: bosque, triángulos: arbustal).

Los gráficos resultantes muestran que los sitios no se distinguen de forma clara en cuanto a la estación del año, pero sí se observa una separación de sitios en cuanto al tipo de ambiente, dada por el primer eje (los valores negativos corresponden al bosque, mientras que los valores positivos corresponden en su mayoría al arbustal).

Análisis de coordenadas principales

Como ejemplo utilizaremos la MBD de *Bulnesia* (Tabla 2.11) y la función `pcoa()` del paquete ape (Paradis y Schliep 2018). A diferencia del PCA y del CA, el PCoA requiere una MD en lugar de la MBD. Debido a la naturaleza de las variables presentes en la MBD (cuantitativas y cualitativas) utilizaremos la distancia de Gower (ver Cap. 4) con el paquete FD (Laliberté *et al.* 2014).

```
> library(FD)
> library(ape)
> Bulnesia <- read.table("C:/R datos/Bulnesia.txt", header = TRUE)
> rownames(Bulnesia) <- Bulnesia$species
> D <- gowdis(Bulnesia)
> pcoa <- pcoa(D)
```

Como en los métodos anteriores, nos interesa obtener los eigenvalores y eigenvectores (coordenadas principales) del análisis, así como sus contribuciones relativas a las coordenadas principales.

```
> pcoa$values
  Eigenvalues Relative_eig Broken_stick Cumul_eig Cumul_br_stick
1 0.408264482  0.578247074   0.37040816 0.5782471      0.3704082
2 0.140531218  0.199041968   0.22755102 0.7772890      0.5979592
3 0.069989104  0.099129355   0.15612245 0.8764184      0.7540816
4 0.044808150  0.063464206   0.10850340 0.9398826      0.8625850
```

```
5 0.022395930  0.031720568   0.07278912 0.9716032      0.9353741
6 0.013202115  0.018698869   0.04421769 0.9903020      0.9795918
7 0.006847129  0.009697959   0.02040816 1.0000000      1.0000000
```

Las columnas 1, 2 y 4 muestran los eigenvalores, la proporción explicada por cada coordenada principal y su acumulado, respectivamente. Así, las tres primeras coordenadas principales acumulan el 87% de la variación total de la MBD. Eventualmente, si llegaran a aparecer eigenvalores negativos, se pueden utilizar las transformaciones vistas anteriormente (Lingoes 1971, Cailliez 1983) con el argumento `correction = "lingoes"` o `"cailliez"`. Por defecto, no se aplica ninguna corrección (`correction = "none"`).

Las columnas restantes (3 y 5) corresponden a valores del modelo de vara quebrada (*broken stick model*) desarrollado por MacArthur (1957) para el estudio de la estructura de comunidades biológicas, y aplicado por Frontier (1976) a métodos de ordenación. Este método se utiliza para seleccionar el número de ejes a utilizar. La idea básica es retener aquellos componentes cuyos eigenvalores sean mayores a los propuestos por el modelo de vara quebrada. En nuestro caso el único componente que reúne este requisito es el PCoA1. A continuación obtendremos las coordenadas principales.

```
> pcoa$vectors
                       Axis.1       Axis.2        Axis.3      Axis.4
B_arborea          -0.32794654  -0.06039308 -0.0004064549 -0.02516102
B_carrapo          -0.37565518  -0.03502267 -0.0026790217 -0.03148966
B_chilensis         0.17440642  -0.08139430  0.1674660734  0.10012632
B_bonariensis      -0.04932985  -0.09037579 -0.0719531604  0.10624921
B_retama            0.18109410  -0.07890608  0.0920899945 -0.13786808
B_foliosa           0.19670483   0.06420631 -0.0870762065 -0.04667171
B_schickendantzii   0.23286656  -0.04718764 -0.1380217596  0.01118640
B_sarmientoi       -0.03214035   0.32907324  0.0405805352  0.02362852

                       Axis.5       Axis.6        Axis.7
B_arborea          -0.04446358   0.07619560   0.017248347
B_carrapo          -0.01357022  -0.07595889  -0.019073929
B_chilensis        -0.05043586  -0.01282170   0.004438234
B_bonariensis       0.10091786   0.00354298   0.014167270
B_retama            0.06355754   0.01076104  -0.010419674
B_foliosa          -0.03299752  -0.03038187   0.057054223
B_schickendantzii  -0.04401175   0.01438992  -0.048896592
B_sarmientoi        0.02100352   0.01427292  -0.014517878
```

La función `biplot()` realiza el *biplot* de especies y variables (Fig. 6.26). Para esto, especificamos el argumento `X`, con el resultado del análisis, y el argumento `Y` que contiene la MBD original con las variables a graficar. Para una mejor visualización, se transformarán las variables a logaritmo. Si se quieren graficar sólo las UE, se especifica `Y = NULL`.

```
> biplot(x = pcoa, Y = log(bulnesia[, -1] + 1), plot.axes = c(1, 2))
```

Debido a que el PCoA se calcula a partir de una MD, no se pueden obtener los *loadings* de las variables a partir del algoritmo. Sin embargo, se pueden calcular *a posteriori* como la correlación entre las variables originales y los ejes principales.

```
loadings <- cor(Bulnesia[, -1], pcoa$vectors, method = "pearson",
+               use = "complete.obs")
```

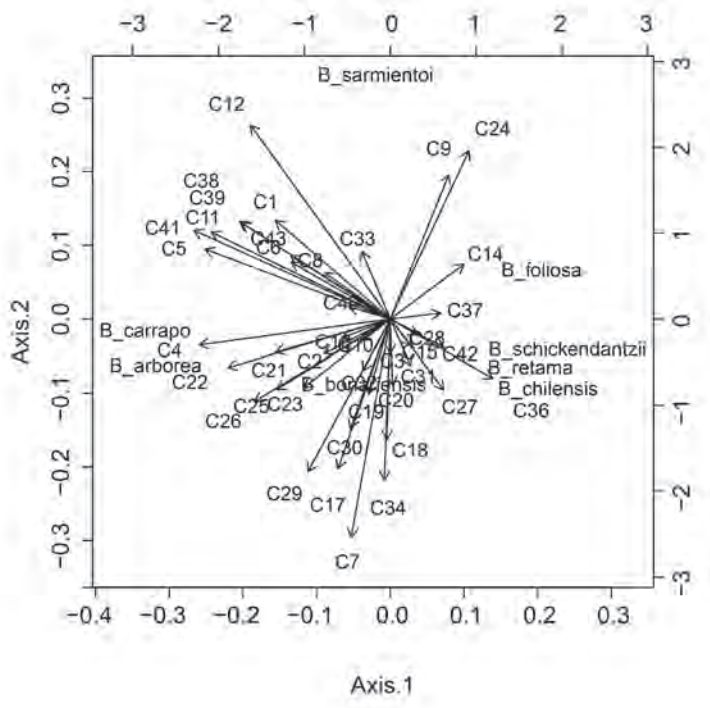


Fig. 6.26. *Biplot* resultante del PCoA (distancia de Gower) aplicado a la MBD de *Bulnesia*.

La interpretación del PCoA es similar a la explicada en el PCA y CA. Así, *B. sarmientoi* queda alejada del resto de las especies, mientras que *B. carrapo* y *B. arborea* forman un grupo, y el resto de especies conforman otro grupo.

Análisis discriminante

Como ejemplo tomaremos la MBD de las dos especies de picaflor (*Chlorostilbon lucidus* e *Hylocharis chrysura*) × cuatro caracteres morfológicos (Tabla 6.15). Los grupos son dos y corresponden a cada especie de picaflor. Antes de realizar el análisis, calcularemos la media y el desvío estándar de las variables para cada especie con el paquete doBy (Højsgaard y Halekoh 2018). Esto es con fines exploratorios, y puede darnos una idea de aquellas variables que más difieren entre las especies. En la función `summaryBy()` debemos especificar aquellas variables para las cuales queremos el resumen (longitud cabeza-cola, cuerda del ala, longitud de la cola y longitud del culmen) como función de (símbolo ~) otra variable (especie). El argumento `FUN` indica la función que queremos calcular (en este caso, la media y el desvío estándar).

```
> library(doBy)
> spp <- read.table("C:/Picaflores.txt", header = TRUE)
> summaryBy(longitud.total + ala + culmen + cola ~ especie,
+           FUN = c(mean, sd), data = spp)
            especie longitud.total.mean ala.mean culmen.mean
1 Chlorostilbon_lucidus      76.66933   51.16800    19.74867
2 Hylocharis_chrysura        78.93000   53.49867    20.98733
```

```
 cola.mean longitud.total.sd   ala.sd culmen.sd  cola.sd
1 29.33800          3.679341 1.411252 0.9012996 2.821940
2 30.02267          2.144878 1.926796 0.8641963 2.188151
```

Antes de realizar el análisis, debemos verificar si los supuestos se cumplen para que los resultados sean considerados válidos. Para esto, aplicaremos dos pruebas estadísticas del paquete MVTests (Bulut 2019). La primera (prueba de normalidad multivariada de Shapiro) evalúa la normalidad de las variables de cada grupo –función `mvShapiro()`–, en la cual debemos indicar si hay grupos (argumento `group = TRUE`) y cuáles son (argumento `G`). Un valor de probabilidad mayor a 0,05 es evidencia a favor de que se cumple este supuesto.

```
> mvShapiro(data = spp[, -1], group = TRUE, G = spp$especie)
$Stat
                 GROUPS Statistic
1 Chlorostilbon_lucidus 0.9657659
2   Hylocharis_chrysura 0.9467283

$p.value
                 GROUPS  P.Values
1 Chlorostilbon_lucidus 0.5009329
2   Hylocharis_chrysura 0.0402061

$Test
[1] "mvShapiro"

$group
[1] "TRUE"

attr(,"class")
[1] "MVTests" "list"
```

Para *H. chrysura* rechazaríamos la hipótesis de que las variables se distribuyen normalmente. Sin embargo, este valor es cercano a 0,05 y el DA es robusto a las violaciones de este supuesto (Quinn y Keough 2002). Por lo tanto, podemos considerar que esto no representa un problema grave para el análisis. La segunda prueba (prueba M de Box) analiza la homogeneidad de varianzas-covarianzas de la MBD –`función BoxM()`–. Nuevamente, un valor de probabilidad mayor a 0,05 es evidencia a favor de que se cumple este supuesto.

```
> BoxM(data = spp[, -1], group = spp$especie)
$Chisq
[1] 6.692906

$df
[1] 10

$p.value
[1] 0.7540835

$Test
```

```
[1] "BoxM"

attr(,"class")
[1] "MVTests" "list"
```

Debido a que no hay evidencia fuerte para rechazar ambos supuestos, podemos aplicar a continuación el DA.

El DA puede realizarse de forma muy sencilla con las funciones `lda()` (DA lineal) y `qda()` (DA cuadrático) del paquete MASS (Venables y Ripley 2002). En este caso, el objetivo es identificar aquellas variables que discriminen más entre las especies de picaflor.

```
> library(MASS)
> ADlineal <- lda(especie ~ longitud.total + ala + culmen + cola, data = spp)
> ADcuad <- qda(especie ~ longitud.total + ala + culmen + cola, data = spp)
```

En los objetos `ADlineal` y `ADcuad` tenemos información sobre las medias de cada grupo y variable, así como los coeficientes de la función discriminante. También podemos graficar los histogramas de las observaciones sobre el eje discriminante (Fig. 6.27).

```
> ADlineal$means
                       longitud.total       ala   culmen      cola
Chlorostilbon_lucidus        76.66933  51.16800 19.74867  29.33800
Hylocharis_chrysura          78.93000  53.49867 20.98733  30.02267

> ADlineal$scaling
                      LD1
longitud.total   0.1658483
ala              0.4483341
culmen           0.7085883
cola            -0.2096638

> ADcuad$scaling
, , Chlorostilbon_lucidus

                        1           2           3            4
longitud.total  0.2717878  0.03961928  0.00168546  -0.27365040
ala             0.0000000 -0.71607991  0.05510573  -0.23167310
culmen          0.0000000  0.00000000  1.11295723  -0.09649744
cola            0.0000000  0.00000000  0.00000000   0.52747993

, , Hylocharis_chrysura

                        1           2           3           4
longitud.total  0.4662271 -0.05897429  0.13657342   0.1132551
ala             0.0000000  0.52313203  0.06345135   0.2951931
culmen          0.0000000  0.00000000 -1.21898762   0.1269198
cola            0.0000000  0.00000000  0.00000000  -0.5528938

> plot(ADlineal)
```

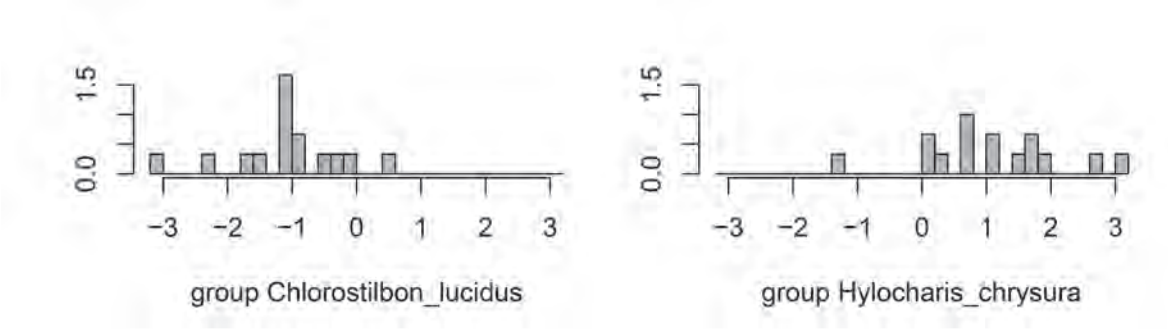


Fig. 6.27. Distribución de las observaciones sobre el primer eje del DA.

Observe que en el QDA la cantidad de coeficientes de la función discriminante es mucho mayor.

Con este análisis también podemos predecir a qué especie pertenece cada individuo en base a las variables medidas. Para esto, realizaremos de nuevo el análisis pero utilizando el argumento `CV = TRUE`, con el objetivo de estimar qué tan bien las especies son clasificadas, mediante validación cruzada "dejando uno fuera".

```
> ADlinealcv <- lda(especie ~ longitud.total + ala + culmen + cola,
+                   CV = TRUE, data = spp)
> ADcuadcv <- qda(especie ~ longitud.total + ala + culmen + cola,
+                 CV = TRUE, data = spp)
> ADlinealcv$class
 [1] Chlorostilbon_lucidus Chlorostilbon_lucidus Chlorostilbon_lucidus
 [4] Chlorostilbon_lucidus Chlorostilbon_lucidus Chlorostilbon_lucidus
 [7] Hylocharis_chrysura Chlorostilbon_lucidus Chlorostilbon_lucidus
[10] Chlorostilbon_lucidus Chlorostilbon_lucidus Chlorostilbon_lucidus
[13] Hylocharis_chrysura Hylocharis_chrysura Chlorostilbon_lucidus
[16] Hylocharis_chrysura Hylocharis_chrysura Hylocharis_chrysura
[19] Chlorostilbon_lucidus Hylocharis_chrysura Hylocharis_chrysura
[22] Hylocharis_chrysura Hylocharis_chrysura Hylocharis_chrysura
[25] Chlorostilbon_lucidus Hylocharis_chrysura Hylocharis_chrysura
[28] Hylocharis_chrysura Hylocharis_chrysura Hylocharis_chrysura
Levels: Chlorostilbon_lucidus Hylocharis_chrysura

> ADcuadcv$class
 [1] Chlorostilbon_lucidus Hylocharis_chrysura Chlorostilbon_lucidus
 [4] Chlorostilbon_lucidus Chlorostilbon_lucidus Chlorostilbon_lucidus
 [7] Hylocharis_chrysura Hylocharis_chrysura Chlorostilbon_lucidus
[10] Chlorostilbon_lucidus Chlorostilbon_lucidus Chlorostilbon_lucidus
[13] Hylocharis_chrysura Hylocharis_chrysura Chlorostilbon_lucidus
[16] Hylocharis_chrysura Hylocharis_chrysura Hylocharis_chrysura
[19] Chlorostilbon_lucidus Hylocharis_chrysura Hylocharis_chrysura
[22] Hylocharis_chrysura Hylocharis_chrysura Hylocharis_chrysura
[25] Chlorostilbon_lucidus Hylocharis_chrysura Hylocharis_chrysura
[28] Chlorostilbon_lucidus Hylocharis_chrysura Hylocharis_chrysura
Levels: Chlorostilbon_lucidus Hylocharis_chrysura
```

Esta información resulta más valiosa si evaluamos en qué medida el análisis predice de forma correcta e incorrecta la especie asignada a cada individuo, mediante una matriz de confusión. Más aún, podemos comparar la capacidad predictiva de ambos modelos.

```
> table(spp$especie, ADlinealcv$class,
+       dnn = c("Especie observada", "Especie predicha"))
                      Especie predicha
Especie observada       Chlorostilbon_lucidus Hylocharis_chrysura
 Chlorostilbon_lucidus                    12                   3
 Hylocharis_chrysura                       2                  13

> table(spp$especie, ADcuadcv$class,
+       dnn = c("Especie observada", "Especie predicha"))
                      Especie predicha
Especie observada       Chlorostilbon_lucidus Hylocharis_chrysura
 Chlorostilbon_lucidus                    10                   5
 Hylocharis_chrysura                       3                  12
```

O, expresada en términos de porcentajes (dividiendo los valores por la cantidad de observaciones de cada especie).

```
> N.Cluc <- length(spp$especie[spp$especie == "Chlorostilbon_lucidus"])
> N.Hchr <- length(spp$especie[spp$especie == "Hylocharis_chrysura"])

> 100*table(spp$especie, ADlinealcv$class,
+           dnn = c("Especie observada", "Especie predicha"))/c(N.Cluc, N.Hchr)
                       Especie predicha
Especie observada        Chlorostilbon_lucidus Hylocharis_chrysura
 Chlorostilbon_lucidus               80.00000            20.00000
 Hylocharis_chrysura                 13.33333            86.66667

> 100*table(spp$especie, ADcuadcv$class,
+           dnn = c("Especie observada", "Especie predicha"))/c(N.Cluc, N.Hchr)
                       Especie predicha
Especie observada        Chlorostilbon_lucidus Hylocharis_chrysura
 Chlorostilbon_lucidus               66.66667            33.33333
 Hylocharis_chrysura                 20.00000            80.00000
```

En este caso es claro que el QDA clasifica las observaciones con una mayor tasa de error que la del LDA, sumado a que este último tiene un menor número de parámetros. Además, podemos realizar gráficos de dispersión entre cada par de variables, así como la función discriminante y los límites de clasificación (Figs. 6.28 y 6.29), con el paquete klaR (Weihs *et al.* 2005).

```
> library(klaR)
> partimat(especie ~ longitud.total + ala + culmen + cola, data = spp,
+         method = "lda")
> partimat(especie ~ longitud.total + ala + culmen + cola, data = spp,
+          method = "qda")
```

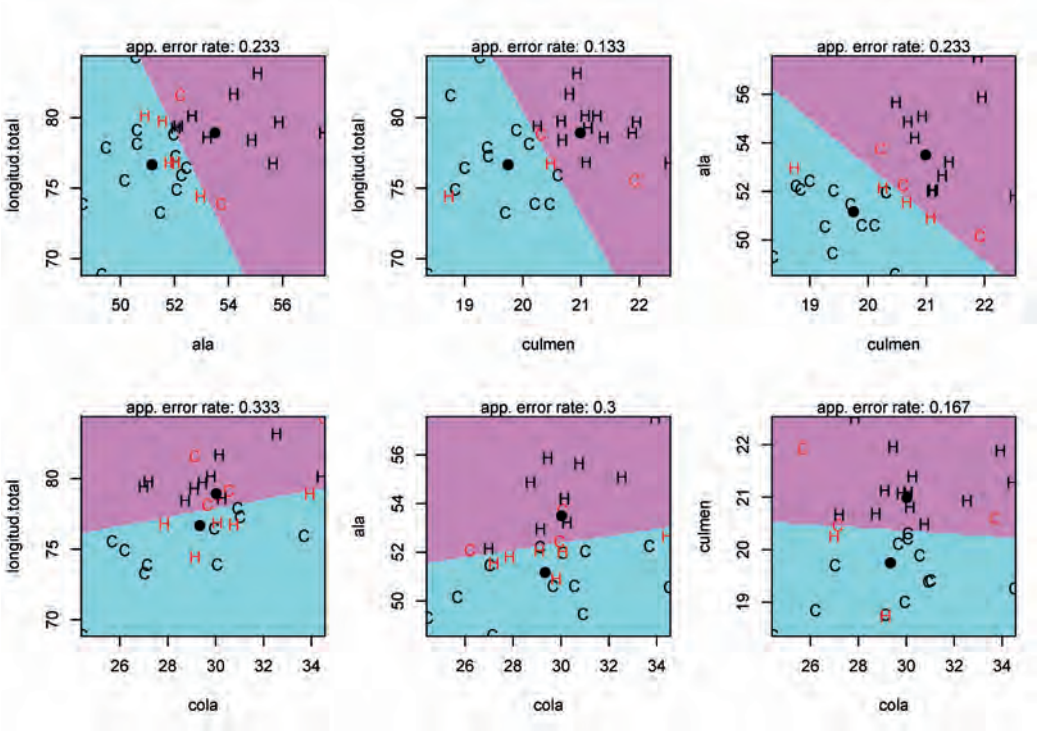


Fig. 6.28. Gráficos de dispersión entre cada par de variables, la función discriminante y los límites de clasificación, resultantes del LDA. La zona rosa representa el área de clasificación para la especie *Hylocharis chrysura,* mientras que la zona celeste representa el área de clasificación para la especie *Chlorostilbon lucidus.* Los círculos negros representan la media de cada especie. Las letras representan las UE (C: *C. lucidus*; H: *H. chrysura*); en negro se muestran las UE bien clasificadas, mientras que en rojo se muestran las UE incorrectamente clasificadas.

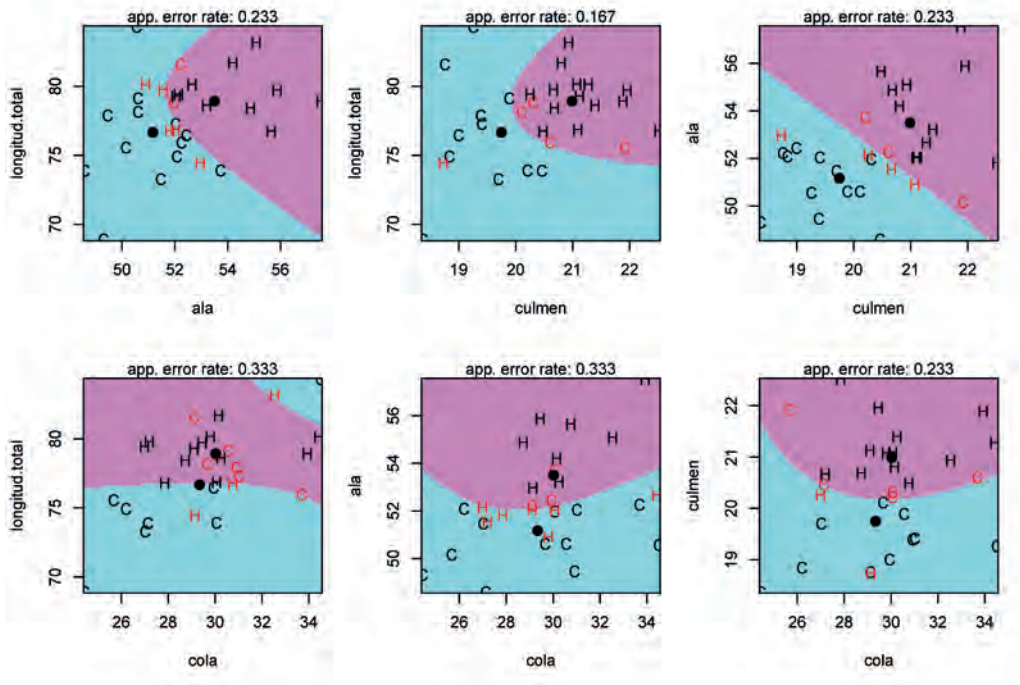


Fig. 6.29. Gráficos de dispersión entre cada par de variables, la función discriminante y los límites de clasificación, resultantes del QDA. La zona rosa representa el área de clasificación para la especie *Hylocharis chrysura,* mientras que la zona celeste representa el área de clasificación para la especie *Chlorostilbon lucidus.* Los círculos negros representan la media de cada especie. Las letras representan las UE (C: *C. lucidus*; H: *H. chrysura*); en negro se muestran las UE bien clasificadas, mientras que en rojo se muestran las UE incorrectamente clasificadas.

Para clasificar una o más UE en particular, podemos utilizar la función `predict()` especificando el análisis utilizado y un marco de datos que contenga las UE de interés. En este último, los nombres de las variables deben ser iguales a los de la MBD utilizada en el DA. A modo de ejemplo, se clasificarán dos individuos utilizando el LDA.

```
> nuevasUE <- data.frame(longitud.total = c(75, 80), ala = c(50, 52),
+                        culmen = c(20, 22), cola = c(28, 33))
> predict(ADlineal, newdata = nuevasUE)
$class
[1] Chlorostilbon_lucidus Hylocharis_chrysura
Levels: Chlorostilbon_lucidus Hylocharis_chrysura

$posterior
  Chlorostilbon_lucidus Hylocharis_chrysura
1             0.9550531          0.04494689
2             0.1891039          0.81089615

$x
         LD1
1 -1.4188883
2  0.6758791
```

La salida muestra que la primera UE es clasificada como *C. lucidus*, mientras que la segunda es clasificada como *H. chrysura* (objeto `class`). También muestra cuál es la probabilidad de que una cada UE corresponda a *C. lucidus* o *H. chrysura* (objeto `posterior`). Así, la probabilidad de que la primera UE corresponda a *C. lucidus* es 0,95, mientras que la probabilidad de que la segunda UE corresponda a esta especie es sólo de 0,19. Como valor de corte para poder clasificar a una UE determinada se suele utilizar el valor de probabilidad de 0,5. Por último, también muestra la coordenada de cada UE sobre el primer eje discriminante (objeto `x`).

A continuación calcularemos la lambda de Wilks, una medida de separación entre los centroides de los grupos que varía entre 0 (máxima separación) y 1 (nula separación). Para esto, debemos combinar en un solo objeto las variables utilizadas en el DA. Luego, utilizamos la función `manova()` para aplicar un análisis multivariado de la varianza (MANOVA) de dichas variables en función de las especies (al revés de lo empleado en las funciones discriminantes). Si bien no se describe aquí, el MANOVA guarda estrecha relación con el análisis discriminante (el lector puede consultar Legendre y Legendre 1998 y Zar 1999 para una descripción del MANOVA). La función `summary()` aplicado al MANOVA con el argumento `test = "Wilks"` arroja el valor de lambda.

```
Y <- cbind(spp$longitud.total, spp$ala, spp$culmen, spp$cola)
manova.res <- manova(Y ~ especie, data = spp)
wilks <- summary(manova.res, test = "Wilks")
wilks
           Df   Wilks approx F num Df den Df    Pr(>F)
especie     1 0.44588 7.7674        4     25 0.0003268 ***
Residuals 28
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Además, la salida arroja una prueba estadística para evaluar si el valor de lambda es significativo o no para la discriminación de los grupos. A pesar de ser un valor relativamente alto, la probabilidad de obtener un valor igual o mayor a 0,446 –columna `Pr(>F)`–, por azar, es menor a 0,001. Por lo tanto, podemos concluir que hay una separación significativa entre ambas especies en base a las variables utilizadas.

Finalmente cabe mencionar que si se analizan tres grupos, el número de ejes discriminantes será igual a dos. Sin embargo, es posible que el primer eje discrimine bien los tres grupos sin la necesidad de recurrir al segundo eje. Esta idea vale igualmente para aquellas situaciones en las cuales hay más de tres grupos. En estos casos y al igual que en el resto de las técnicas de escalado multidimensional métrico, la salida arroja como resultado el porcentaje de variación explicada por cada eje (denominado proporción de traza).

Escalado multidimensional no métrico

El NMDS a diferencia de los otros métodos, es una técnica no métrica y permite utilizar cualquier coeficiente de similitud. Como ejemplo, retomaremos la MBD de sitios × especies de aves. Para esto trabajaremos con la función `metaMDS()` del paquete vegan (Oksanen *et al.* 2018).

```
> library(vegan)
> Aves <- read.table("C:/R datos/Aves.txt", header = TRUE)
> comm <- Aves[, -c(1:3)]
```

Previamente, etiquetaremos las filas de la MBD para identificar el sitio y la estación del año.

```
> rownames(comm) <- paste(Aves$sitio, Aves$estacion, sep = ".")
```

Debido a la naturaleza de la MBD (sitios × especies), es apropiado utilizar algún coeficiente de similitud de los presentados en el Capítulo 4, que no considere las ausencias compartidas en favor de la similitud. En la función `metaMDS()` puede utilizarse cualquier coeficiente disponible en la función `vegdist()` (Cap. 4); en este caso utilizaremos el coeficiente de Bray-Curtis. Recuerde también que la cantidad de ejes (argumento `k`) para representar la MBD se debe determinar *a priori*, dado que el análisis busca aquella configuración que mejor representa a la MBD original, en la cantidad de dimensiones especificadas.

```
> nmds <- metaMDS(comm, distance = "bray", k = 2)
Wisconsin double standardization
Run 0 stress 0.2319559
Run 1 stress 0.2615648
Run 2 stress 0.2297402
... New best solution
... Procrustes: rmse 0.03153464 max resid 0.1564544
Run 3 stress 0.2297441
... Procrustes: rmse 0.00140403 max resid 0.005989011
... Similar to previous best
Run 4 stress 0.2307836
Run 5 stress 0.2297403
... Procrustes: rmse 0.0001443429 max resid 0.0006975807
... Similar to previous best
Run 6 stress 0.2307834
Run 7 stress 0.2297402
... New best solution
```

```
... Procrustes: rmse 8.284434e-05 max resid 0.0002175069
... Similar to previous best
Run 8 stress 0.23083
Run 9 stress 0.2308315
Run 10 stress 0.2308344
Run 11 stress 0.2297403
... Procrustes: rmse 0.0001850787 max resid 0.0008076595
... Similar to previous best
Run 12 stress 0.2297403
... Procrustes: rmse 0.0001578153 max resid 0.0006707185
... Similar to previous best
Run 13 stress 0.2307834
Run 14 stress 0.2297435
... Procrustes: rmse 0.001161384 max resid 0.004953395
... Similar to previous best
Run 15 stress 0.2316109
Run 16 stress 0.2307834
Run 17 stress 0.2297402
... Procrustes: rmse 4.011945e-05 max resid 0.0001479074
... Similar to previous best
Run 18 stress 0.2297403
... Procrustes: rmse 0.0001636615 max resid 0.0006901519
... Similar to previous best
Run 19 stress 0.2313544
Run 20 stress 0.2316112
*** Solution reached
```

Como primera medida, la función aplica una estandarización previa cuando el valor máximo de abundancia de la MBD es mayor a nueve. Ésta corresponde a la doble estandarización de Wisconsin, en la que cada elemento se divide por el máximo de su columna, y luego por el total de su fila correspondiente a esta nueva matriz (Cottam *et al*. 1978). Los valores resultantes varían entre 0 y 1. Si el máximo valor de abundancia en la MBD es mayor a 50, la función además aplica la raíz cuadrada. Estas transformaciones evitan la influencia de abundancias altas sobre el resultado del análisis.

Durante la corrida, la función comienza probando diferentes configuraciones aleatorias con el fin de evitar caer en mínimos locales (Legendre y Legendre 1998). El número de corridas puede controlarse con el argumento `trymax`, cuyo valor por defecto es 20 (puede aumentarse si no se encuentra una solución óptima). Si un valor de estrés es menor que el de la solución anterior, se toma como mejor solución. Si el nuevo valor de estrés es similar al anterior, se considera que el análisis convergió y se toma el valor mínimo de los dos. Una vez realizado el NMDS obtenemos el valor de estrés como:

```
> nmds$stress
[1] 0.2297402
```

Observe que este valor es el mínimo de todas las corridas (corridas 2, 7 y 17). La representación no es del todo satisfactoria (valores de estrés mayor a 0,20 indican una representación pobre), por lo que en este caso sería conveniente agregar otra dimensión. Sin embargo, a los fines prácticos continuaremos con esta configuración para entender la lógica del análisis.

El diagrama de Shepard (Fig. 6.30) muestra qué tan bien las distancias de la MBD (*Observed Dissimilarity*) se preservan en el espacio de ordenación (*Ordination Distance*).

```
> stressplot(nmds, pch = 19, p.col = "gray70", l.col = "black")
```

La Figura 6.30 también muestra el ajuste de una regresión no paramétrica, donde se crean intervalos sobre el eje *x* para los cuales se calcula el promedio de los puntos que caen en dicho intervalo. Luego se unen los promedios mediante líneas verticales. También se reportan dos medidas del ajuste, un R^2 no métrico (*Non-metric fit*) que corresponde a la regresión no paramétrica y un R^2 lineal (*Linear fit*) que corresponde a una recta. Ambas medidas indican la proporción explicada de la variación en los datos por dicho modelo (máximo = 1), y su interpretación es contraria al estrés (un alto valor de R^2 indica un buen ajuste, un alto valor de estrés indica un ajuste pobre). En este caso es preferible tener en cuenta el ajuste no paramétrico, ya que la relación es no lineal.

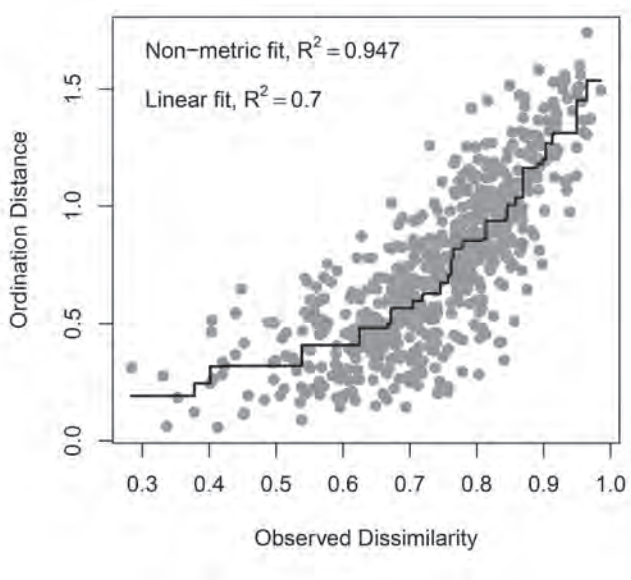


Fig. 6.30. Diagrama de Shepard de la MBD de sitios × especies. Los puntos representan distancias entre pares de UE, la línea muestra el ajuste de una regresión no paramétrica.

Ahora graficaremos los sitios junto con las especies en el espacio de ordenación (Fig. 6.31).

```
> ordiplot(nmds, type = "n")
> orditorp(nmds, display = "species", air = 0.01)
> orditorp(nmds, display = "sites", col = "gray50", air = 0.01)
> abline(h = 0, lty = 2)
> abline(v = 0, lty = 2)
```

El argumento `air` representa la cantidad de espacio entre etiquetas, valores menores a 1 permiten que las etiquetas se superpongan.

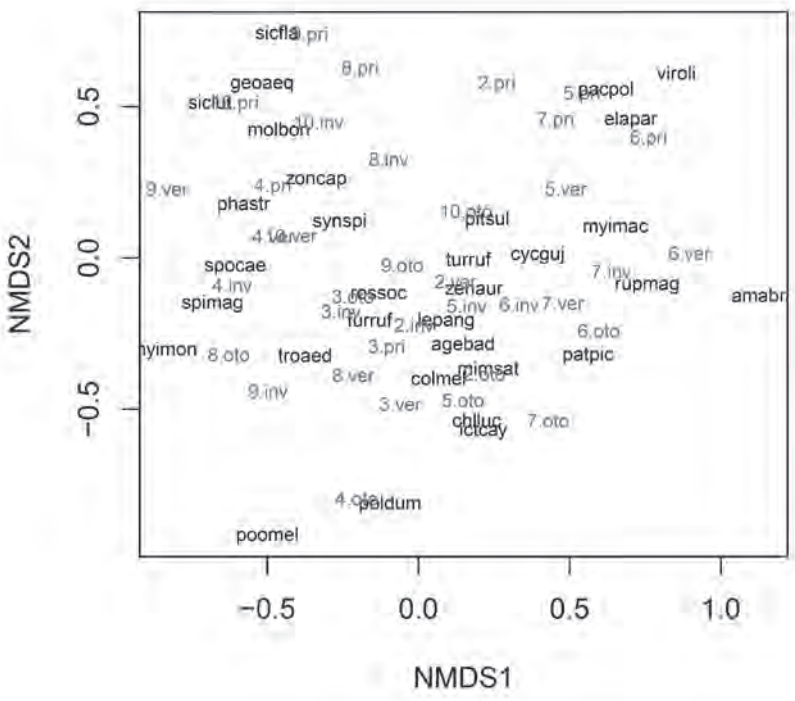


Fig. 6.31. *Biplot* resultante del NMDS aplicado a la MBD de sitios (letras grises) × especies de aves (letras negras). Agebad: *Agelaioides badius*, amabra: *Amazonetta brasiliensis*, rupmag: *Rupornis magnirostris*, spimag: *Spinus magellanicus*, chlluc: *Chlorostilbon lucidus*, colmel: *Colaptes melanochloros*, patpic: *Patagioenas* picazuro, cycguj: *Cyclarhis gujanensis*, elapar: *Elaenia parvirostris*, furruf: *Furnarius rufus*, geoaeq: *Geothlypis aequinoctialis*, ictcay: *Icterus pyrrhopterus*, lepang: *Lepidocolaptes angustirostris*, mimsat: *Mimus saturninus*, molbon: *Molothrus bonariensis*, myimac: *Myiodynastes maculatus*, myimon: *Myiopsitta monachus*, pacpol: *Pachyramphus polychopterus*, phastr: *Phacellodomus striaticollis*, pitsul: *Pitangus sulphuratus*, poldum: *Polioptila dumicola*, poomel: *Poospiza melanoleuca*, rossoc: *Rostrhamus sociabilis*, sicfla: *Sicalis flaveola*, siclut: *S. luteola*, spocae: *Sporophila caerulescens*, synspi: *Synallaxis spixi*, troaed: *Troglodytes aedon*, turruf: *Turdus rufiventris*, viroli: *Vireo olivaceus*, zenaur: *Zenaida auriculata*, zoncap: *Zonotrichia capensis*.

La interpretación es la misma que para cualquier otro método de ordenación, es decir, las UE cercanas en el espacio son similares en cuanto a sus variables. A diferencia del CA aplicado a la misma MBD, el NMDS brinda una distribución de las UE más homogénea a través de los cuadrantes, eliminando el efecto arco debido a la estandarización y al uso de rangos en lugar de distancias originales. Podemos también acceder a los *scores* de los sitios y de las especies.

```
> scores(nmds, display = "sites")
            NMDS1       NMDS2
2.inv  -0.01471644 -0.22135992
2.oto   0.21709620 -0.38400477
2.pri   0.25960025 0.57328799
2.ver   0.12197818 -0.07943164
3.inv  -0.25615383 -0.17755195
3.oto  -0.21720525 -0.12882998
3.pri  -0.10490042 -0.29821922
```

```
3.ver  -0.06067626 -0.48347319
4.inv  -0.61604477 -0.09098942
4.oto  -0.20399886 -0.79642116
4.pri  -0.48164160  0.23411164
4.ver  -0.47953421  0.07002950
5.inv   0.15921843 -0.15901387
5.oto   0.14735666 -0.47245435
5.pri   0.54031046  0.53963173
5.ver   0.48662588  0.22748027
6.inv   0.33321644 -0.15463723
6.oto   0.59627361 -0.24175409
6.pri   0.76133095  0.39281674
6.ver   0.89627041  0.01483790
7.inv   0.63785652 -0.04422817
7.oto   0.42854180 -0.53925354
7.pri   0.45379497  0.45344949
7.ver   0.47707045 -0.15094077
8.inv  -0.09897418  0.32609523
8.oto  -0.62758873 -0.32042881
8.pri  -0.19248219  0.62240824
8.ver  -0.21591409 -0.38933363
9.inv  -0.49718795 -0.44073888
9.oto  -0.05356640 -0.02530201
9.pri  -0.35992143  0.73452043
9.ver  -0.83196975  0.22732688
10.inv -0.33091012  0.44757372
10.oto  0.15674030  0.15338093
10.pri -0.60749194  0.50882132
10.ver -0.42240311  0.07259461

> scores(nmds, display = "species")
            NMDS1        NMDS2
agebad  0.14798613 -0.289312887
amabra  1.14164677 -0.123543387
rupmag  0.75657178 -0.095968774
spimag -0.68096094 -0.152095337
chlluc  0.19281345 -0.536085815
colmel  0.06666481 -0.397438565
patpic  0.56103141 -0.322780309
cycguj  0.39421657  0.009098578
elapar  0.70173329  0.455750148
furruf -0.16121190 -0.204264732
geoaeq -0.51693510  0.568916182
ictcay  0.21477029 -0.572315250
lepang  0.09140585 -0.209281110
mimsat  0.23063490 -0.364318305
molbon -0.46223077  0.425867592
```

```
myimac  0.65155984  0.100421182
myimon -0.84235589 -0.306477282
pacpol  0.61957706  0.553568843
phastr -0.57866707  0.174676021
pitsul  0.22769250  0.125248346
poldum -0.09411887 -0.815603649
poomel -0.49970899 -0.919960425
rossoc -0.13232853 -0.119210924
sicfla -0.46787644  0.745851934
siclut -0.69028181  0.514852005
spocae -0.60612462 -0.033981905
synspi -0.26057178  0.118564077
troaed -0.37473785 -0.321992267
turruf  0.16506554 -0.003761465
viroli  0.85318488  0.611851432
zenaur  0.18298364 -0.105128548
zoncap -0.33874187  0.253722032
```

Como se hizo en el CA, también podemos explorar si se forman grupos de UE según variables que no intervienen en el análisis, como la estación del año o el tipo de ambiente (Fig. 6.32). El paquete vegan brinda facilidades para mostrar agrupamientos, donde podemos utilizar polígonos –función `ordihull()`– o elipses –función `ordiellipse()`– para agrupar las UE con las mismas características. Para esto debemos tener activo el gráfico anterior sobre la consola.

```
> ordiplot(nmds, type = "n")
> orditorp(nmds, display = "species", air = 0.01)
> orditorp(nmds, display = "sites", col = "gray50", air = 0.01)
> abline(h = 0, lty = 2)
> abline(v = 0, lty = 2)
> ordihull(nmds, groups = Aves$estacion, draw = "polygon", label = TRUE)
```

```
> ordiplot(nmds, type = "n")
> orditorp(nmds, display = "species", air = 0.01)
> orditorp(nmds, display = "sites", col = "gray50", air = 0.01)
> abline(h = 0, lty = 2)
> abline(v = 0, lty = 2)
> ordiellipse(nmds, groups = Aves$ambiente, label = TRUE)
```

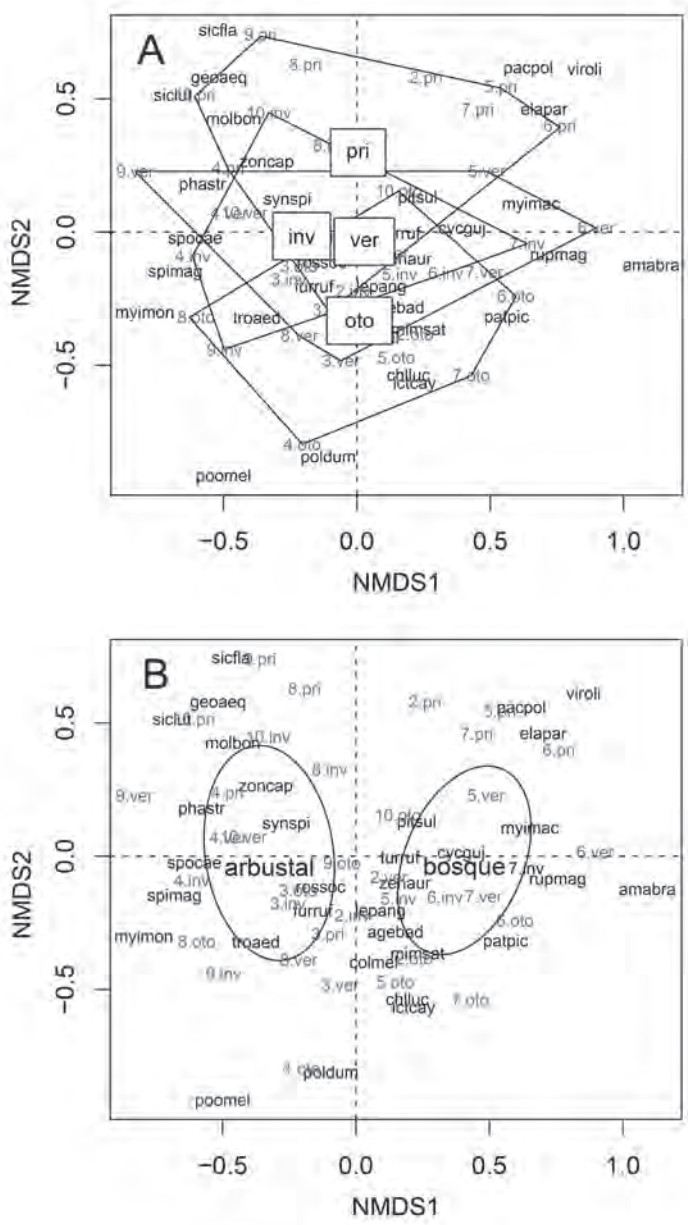


Fig. 6.32. *Biplots* resultantes del NMDS aplicado a la MBD de sitios (letras grises) × especies de aves (letras negras). (A) Se muestran los sitios agrupados por estación del año (polígonos; oto: otoño, inv: invierno, pri: primavera, ver: verano); (B) sitios agrupados por tipo de ambiente (elipses). Agebad: *Agelaioides badius*, amabra: *Amazonetta brasiliensis*, rupmag: *Rupornis magnirostris*, spimag: *Spinus magellanicus*, chlluc: *Chlorostilbon lucidus*, colmel: *Colaptes melanochloros*, patpic: *Patagioenas* picazuro, cycguj: *Cyclarhis gujanensis*, elapar: *Elaenia parvirostris*, furruf: *Furnarius rufus*, geoaeq: *Geothlypis aequinoctialis*, ictcay: *Icterus pyrrhopterus*, lepang: *Lepidocolaptes angustirostris*, mimsat: *Mimus saturninus*, molbon: *Molothrus bonariensis*, myimac: *Myiodynastes maculatus*, myimon: *Myiopsitta monachus*, pacpol: *Pachyramphus polychopterus*, phastr: *Phacellodomus striaticollis*, pitsul: *Pitangus sulphuratus*, poldum: *Polioptila dumicola*, poomel: *Poospiza melanoleuca*, rossoc: *Rostrhamus sociabilis*, sicfla: *Sicalis flaveola*, siclut: *S. luteola*, spocae: *Sporophila caerulescens*, synspi: *Synallaxis spixi*, troaed: *Troglodytes aedon*, turruf: *Turdus rufiventris*, viroli: *Vireo olivaceus*, zenaur: *Zenaida auriculata*, zoncap: *Zonotrichia capensis*.

Al igual que el CA, el NMDS muestra una buena separación entre los sitios según el tipo de ambiente, mientras que la distinción entre las UE según la estación del año no es clara. Cabe mencionar que siempre que se realiza un NMDS es recomendable exportar los resultados del análisis –función `write.table()`–, incluidos los *scores* de las UE y de las variables. Ya que es un método iterativo, obtendremos resultados diferentes en cada corrida.

Agrupamiento jerárquico sobre componentes principales

El agrupamiento jerárquico sobre componentes principales (HCPC) combina técnicas de ordenación con análisis de agrupamientos (Cap. 5), permitiendo explorar y explotar al máximo una MBD. Para realizar este análisis utilizaremos la función `HCPC()` del paquete FactoMineR (Lê *et al.* 2017), aplicado a una MBD que consiste en nueve áreas de la Cuenca Inferior del Río de La Plata × 84 taxones de plantas y animales con datos de presencia-ausencia (Apodaca *et al.* 2019a). El objetivo es encontrar grupos que representen áreas de endemismo. La función `HCPC()` sólo contempla algunos métodos de escalado multidimensional métrico. Sin embargo, el usuario podrá aplicar la misma lógica (ordenación-agrupamiento) utilizando funciones por separado (incluyendo PCoA, NMDS) vistas en los Capítulos 3, 4 y 5.

El primer paso consiste en aplicar un método de ordenación. Debido a que la MBD presenta solamente datos categóricos medidos en las mismas unidades, es apropiado utilizar un método como el CA. Para esto, utilizaremos la función `CA()` y analizaremos la variación explicada por los ejes principales. Observe que el número total de ejes es igual a ocho, debido a que hay menos UE que variables. El argumento `ncp` indica el número de ejes a retener. En nuestro caso, ya que nuestro objetivo es transformar los datos categóricos de la MBD en datos continuos (representados por los *scores* del CA), mantendremos todos los componentes. Previamente añadiremos etiquetas a las filas de la MBD para poder visualizarlas en los gráficos.

```
> library(FactoMineR)
> RLP <- read.csv("C:/R Datos/biogeografia Rio de La Plata.csv")
> rownames(RLP) <- RLP$sitio
> ca <- CA(RLP[, -1], ncp = 8, graph = FALSE)
> round(ca$eig, 3)
   eigenvalue percentage of variance cumulative percentage of variance
dim   1 0.392                 27.949                             27.949
dim 2   0.312                 22.250                             50.198
dim 3   0.192                 13.722                             63.920
dim 4   0.168                 11.991                             75.911
dim 5   0.107                  7.607                             83.518
dim 6   0.086                  6.145                             89.663
dim 7   0.078                  5.542                             95.205
dim 8   0.067                  4.795                            100.000
```

A continuación, aplicaremos el método de ligamiento promedio (argumento `method = "average"`) sobre los *scores* resultantes del CA con la función `HCPC()`. Debido a que en Apodaca *et al.* (2019a) se sugiere la presencia de tres grupos, especificaremos el número de agrupamientos con el argumento `nb.clust = 3`.

```
> hcpc <- HCPC(ca, nb.clust = 3, method = "average", graph = FALSE)
```

Para visualizar el dendrograma (Fig. 6.33A) utilizaremos la función `fviz_dend()` del paquete factoextra (Kassambara y Mundt 2017). Previamente, debemos modificar el objeto dentro de `hcpc` que define el número de grupos a graficar.

```
> library(factoextra)
> hcpc$call$t$nb.clust <- 3
> fviz_dend(hcpc, k_colors = rep("black", 3), rect = TRUE, rect_fill = TRUE,
+           color_labels_by_k = FALSE, horiz = TRUE)
```

Con el argumento `k_colors` se indican los colores de cada grupo (se utilizó la función `rep()` para crear un vector que repita el color negro tres veces), los argumentos `rect` y `rect_fill` se utilizan para graficar rectángulos llenos asociados a cada grupo, el argumento `color_labels_by_k` especifica si se deben colo-

rear las etiquetas de las UE de acuerdo a los colores especificados por `k_colors` y, finalmente, `horiz` indica si el gráfico debe tener orientación horizontal. El dendrograma sostiene la formación de tres grandes grupos (Fig. 33A): Río Uruguay Superior-Río Uruguay Medio (grupo 1), Río Paraguay Inferior-Río Paraná Superior-Río Paraná Medio-Esteros del Iberá (grupo 2) y Delta Superior-Delta Inferior-Río de la Plata (grupo 3).

Con la función `fviz_cluster()` es posible graficar e identificar (mediante colores y figuras geométricas) las UE sobre los ejes principales del CA, según el grupo al que pertenecen (Fig. 33B). Esta función toma como argumentos, el resultado del HCPC, los datos (`data`) y los ejes a graficar (`axes`). También podemos indicar si queremos añadir un gráfico en estrella para cada grupo (`star.plot`), etiquetas no superpuestas (`repel`), elipses delimitando cada grupo (`ellipse` y `ellipse.type`) y colores asociados a cada grupo (`palette`).

```
> fviz_cluster(hcpc, data = RLP[, -1], axes = c(1, 2), repel = TRUE,
+              star.plot = TRUE, ellipse = TRUE, ellipse.type = "confidence",
+              palette = rep("black", 3))
```

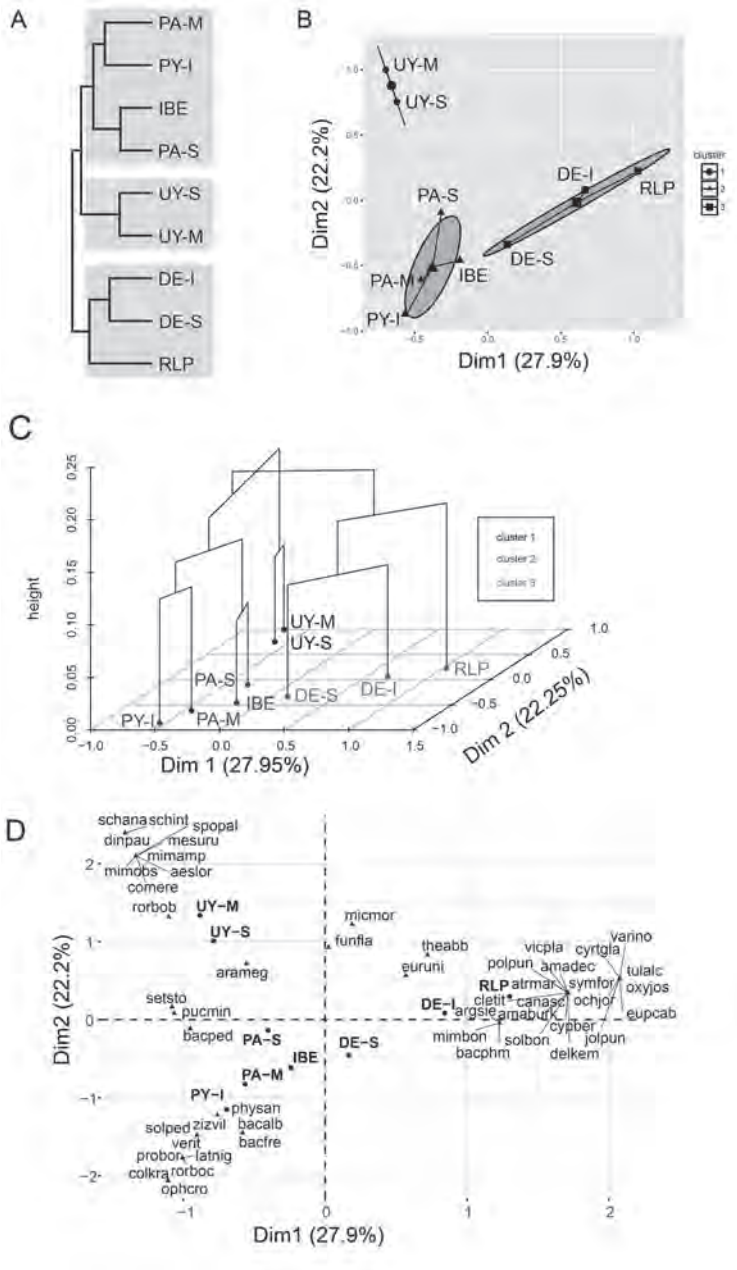


Fig. 6.33. HCPC. (A) Agrupamiento jerárquico (UPGMA) sobre los ejes principales del CA; (B) CA (ejes 1 y 2) donde se muestran los grupos formados por el agrupamiento; (C) dendrograma superpuesto sobre el CA; (D) *biplot* simétrico de áreas y especies. PY-I: Río Paraguay Inferior, PA-S: Río Paraná Superior, PA-M: Río Paraná Medio, IBE: Esteros del Iberá, UY-S: Río Uruguay Superior, UY-M: Río Uruguay Medio, DE-I: Delta Inferior, DE-S: Delta Superior, RLP: Río de la Plata. Aeslor: *Aeschynomene lorentziana*, amaburk: *Amauropelta burkartii*, amadec: *A. decurtata* var. *platensis*, arameg: *Araujia megapotamica*, argsie: *Argenteohyla siemersi siemersi*, atrmar: *Atrichonotus marginatus*, bacalb: *Baccharis albida*, bacfre: *B. frenguellii*, bacped: *B. pedersenii*, bacphm: *B. phyteumoides*, canasc: *Canna ascendens*, cletit: *Cleome titubans*, colkra: *Colobosaura kraepelini*, comere: *Commelina erecta* fo. *dielsii*, cypber: *Cyperus berroi*, cyrtgla: *Cyrtomon glaucus*, delkem: *Deltamys kempi kempi*, dinpau: *Dinogeophilus pauropus*, eupcab: *Eupatorium cabrerae*, euruni: *Eurymetopus unicolor*, funfla: *Funastrum flavum*, jolpun: *Jollas puntalara*, latnig: *Lathyrus nigrivalvis*, mesuru: *Mesabolivar uruguayensis*, micmor: *Micrognamma mortoniana*, mimamp: *Mimosa amphigena*, mimbon: *M. bonplandii*, mimobs: *M. obstrigosa*, ochjor: *Ochlerotatus* jorgi, ophcro: *Ophioglossum crotalophoroides* var. *nanum*, oxyjos: *Oxymycterus josei*, physan: *Physalaemus santafecinus*, polpun: *Polybetes punctulatus*, probor: *Progonyleptes borellii*, pucmin: *Pucrolia minuta*, rorbob: *Rorippa bonariensis* var. *burkartii*, rorboc: *R. bonariensis* var. *chacoensis*, schana: *Schendylops anamariae*, schint: *S. interfluvius*, setsto: *Setaria stolonifera*, solbon: *Solanum bonariense*, solped: *S. pedersenii*, spopal: *Sporophila palustris*, symfor: *Sympathica formosa*, theabb: *Thelypteris abbiattii*, tulalc: *Tullbergia alcirae*, varino: *Varinodulia*, verit: *Verita*, vicpla: *Vicia platensis*, zizvil: *Zizaniopsis villanensis*.

A pesar de que los dos primeros ejes explican poca variación de la MBD (~50%), los tres grupos se diferencian bastante bien (el grupo 1 en el cuadrante superior izquierdo, el grupo 2 en el cuadrante inferior izquierdo y el grupo 3 en los cuadrantes superior e inferior derecho; Fig. 6.33B). También es posible combinar el dendrograma con la ordenación (Fig. 6.33C) mediante la función `plot.HCPC()`. Previamente, cambiaremos los colores del gráfico utilizando una paleta en tono de grises con la función `palette()`. En la función `plot.HCPC()`, el argumento `choice = "3D.map"` indica el tipo de gráfico (tres dimensiones en este caso), mientras que el argumento `angle` especifica su ángulo de rotación.

```
> palette(gray(c(0, 0.3, 0.5, 0.7)))
> plot.HCPC(hcpc, axes = c(1, 2), choice = "3D.map", angle = 40)
```

También podemos graficar un *biplot* simétrico para identificar asociaciones entre áreas y especies (Fig. 33D). Debido a la gran cantidad de especies, sólo se mostrarán las primeras 50 que más contribuyen a la ordenación (argumento `select.col = list(contrib = 50)`).

```
> fviz_ca_biplot(ca, map = "symbiplot", col.row = "black", col.col = "black",
+                select.col = list(contrib = 50), repel = TRUE)
```

En las Figuras 6.33B y D se observa que el grupo 1 se asocia con la presencia de *Sporophila palustris* (spopal), *Mimosa obstrigosa* (mimobs) y *M. amphigena* (mimamp); el grupo 2 se asocia con la presencia de *Physalaemus santafecinus* (physan), *Baccharis albida* (bacalb) y *B. frenguellii* (bacfre); y el grupo 3 (particularmente el Río de la Plata y Delta Inferior) se asocia con la presencia de *B. phyteumoides* (bacphy), *Mimosa bonplandii* (mimbon) y *Argenteohyla siemersi siemersi* (argsie).

Una vez realizados los gráficos podemos analizar cuáles UE y ejes principales contribuyen más a los agrupamientos, información contenida en los objetos `hcpc$desc.ind` y `hcpc$desc.axes`, respectivamente.

```
> hcpc$desc.ind$para
Cluster: 1
     UY-S      UY-M
0.6254744 0.6254744
------------------------------------------------------------------------
Cluster: 2
     PA-S      PA-M       IBE      PY-I
0.9583097 0.9834212 1.0926326 1.1510409
------------------------------------------------------------------------
Cluster: 3
     DE-I      DE-S       RLP
0.6654775 0.9147964 0.9316826
```

En la salida anterior se muestran las UE “modelo” o más representativas de cada grupo, cuyos valores corresponden a la distancia entre una UE y su centroide correspondiente. Por ejemplo, las áreas más representativas de los grupos 2 y 3 son el Río Paraná Superior y el Delta Inferior, respectivamente.

También podemos evaluar qué ejes del CA contribuyen más a cada grupo (las cifras se redondearán a tres dígitos).

```
> lapply(hcpc$desc.axes$quanti, round, 3)
$`1`
      v.test Mean in category Overall mean sd in category Overall sd p.value
Dim.2 13.438            0.889       -0.067          0.122   0.543      0.000
```

```
Dim.3  2.532            0.116       -0.048          0.015   0.493      0.011
Dim.1 -8.437           -0.664        0.016          0.038   0.615      0.000

$`2`
       v.test Mean in category Overall mean sd in category Overall sd p.value
Dim.4   6.591            0.223        0.034          0.261      0.423   0.000
Dim.7   3.250            0.077        0.017          0.227      0.272   0.001
Dim.6   3.084            0.063        0.002          0.106      0.289   0.002
Dim.3  -5.153           -0.220       -0.048          0.724      0.493   0.000
Dim.1  -9.512           -0.381        0.016          0.140      0.615   0.000
Dim.2 -12.517           -0.528       -0.067          0.281      0.543   0.000

$`3`
      v.test Mean in category Overall mean sd in category Overall sd p.value
Dim.1 15.785            0.670        0.016          0.343      0.615   0.000
Dim.3  3.261            0.061       -0.048          0.073      0.493   0.001
Dim.2  2.496            0.024       -0.067          0.217      0.543   0.013
Dim.6 -2.580           -0.048        0.002          0.385      0.289   0.010
Dim.7 -3.111           -0.040        0.017          0.205      0.272   0.002
Dim.4 -6.101           -0.139        0.034          0.537      0.423   0.000
```

Para cada grupo (números asociados a los signos `$`) se muestran la media y el desvío estándar (`Mean in category`, `sd in category`) de los valores de un eje principal (`Dim`), la media y el desvío estándar global de un eje principal para todas las UE (`Overall mean`, `Overall sd`) y una prueba estadística que, en términos generales, compara la media de un grupo con su respectiva media global (`v.test`) y evalúa su probabilidad (`p.value`) (Husson *et al.* 2017). Valores menores a 0,05 se pueden considerar como una asociación significativa entre el eje y el grupo en cuestión. Por mencionar sólo algunos ejemplos, el grupo 1 se asocia positivamente con el eje 2 y negativamente con el eje 1; el grupo 2 se asocia positivamente con el eje 4 y negativamente con el eje 2; y el grupo 3 se asocia positivamente con el eje 1 y negativamente con el eje 4.

CAPÍTULO 7

ESTIMACIÓN DE LA HISTORIA EVOLUTIVA: FUNDAMENTOS DEL ANÁLISIS FILOGENÉTICO Y EL MÉTODO DE PARSIMONIA

¿Tiene la vida una historia? Esta pregunta fue respondida afirmativamente por Darwin (1859) quien, en su libro "El Origen de las Especies" publicó una única ilustración, un árbol genealógico hipotético de la representación de esa historia. Haeckel (1866) denominó a esa representación jerárquica de la historia de la vida "filogenia". Esta historia es producto de los procesos de la evolución biológica (Apodaca *et al.* 2019b). En el análisis filogenético las variables están representadas por caracteres y las unidades de estudio (UE) por taxones (desde dominios a subespecies), por poblaciones o por individuos.

Las preguntas que desde Darwin hasta nuestros días están vigentes son dos:

¿Dónde está escrita la historia de la vida (tipo de carácter)? y ¿cómo se lee la historia de la vida (método)? En este capítulo y en el próximo trataremos de responder ambas preguntas.

Los pasos para estimar esa historia son: (1) registrar la variación en los caracteres utilizados, y (2) aplicar un método a esa variación que genere una hipótesis genealógica.

Los caracteres utilizados para reconstruir la filogenia pueden ser morfológicos (externos, internos -anatomía-, embriológicos, palinológicos, citológicos, histológicos y ultra-estructurales), fisiológicos, químicos, etológicos, ecológicos, genéticos y moleculares (Normark y Lanteri 1998). A pesar de que todos ellos tienen su valor en la estimación filogenética, la mayoría de los árboles filogenéticos de grupos actuales son generados en primer lugar, utilizando caracteres moleculares y, en segundo lugar, morfológicos (Crisci *et al.* 2019).

Las rutinas de R utilizadas para reconstruir filogenias serán discutidas al final del Capítulo 8.

HOMOLOGÍA

Para estimar la historia de la vida, es fundamental discutir el concepto de homología y el de su complemento, la homoplasia. Homología es la posesión por dos o más taxones de un carácter que, con o sin modificaciones, ha sido heredado de un ancestro común. Este concepto es incluso anterior a la teoría evolutiva de Darwin: el mismo órgano bajo una gran variedad de formas y funciones (Owen 1843). Actualmente se aplica el moderno concepto de homología a los distintos niveles de análisis, el organísmico y el molecular. En el Box 7.1 se exponen distintas definiciones de homología.

Box 7.1. Definiciones del término homología

Existen numerosas definiciones del término homología. Aquí se presentan cinco, de las cuales las primeras cuatro se encuentran dentro de un marco evolutivo, mientras que la última está basada en similitud.

Hennig (1968): estados homólogos son aquellos que pueden ser considerados como sucesivos estados de transformación de un mismo estado inicial. Se entiende por transformación al proceso histórico real de la evolución.
Mayr (1969): homólogos son los caracteres de dos o más organismos, cuyo origen puede determinarse en el mismo carácter del ancestro común de esos organismos.
Wiley (1975): dos o más caracteres son homólogos si ellos son estados de transformación de un mismo carácter original, que estaba presente en el ancestro común de los organismos que presentan los caracteres.
Bertalanffy (1987): dos caracteres son homólogos si se originan filogenéticamente uno del otro, o de una base hereditaria común.
Sneath y Sokal (1973): dos caracteres son homólogos cuando se corresponden en su composición y en su estructura. Por correspondencia en la composición se entiende la similitud cualitativa desde el punto de vista biológico y/o químico de sus constituyentes. Por correspondencia estructural se entiende la similitud, en cuanto al orden de sus partes u orden espacio-temporal, o en la estructura de sus fenómenos bioquímicos, o en el orden secuencial de las sustancias o estructuras organizadas.

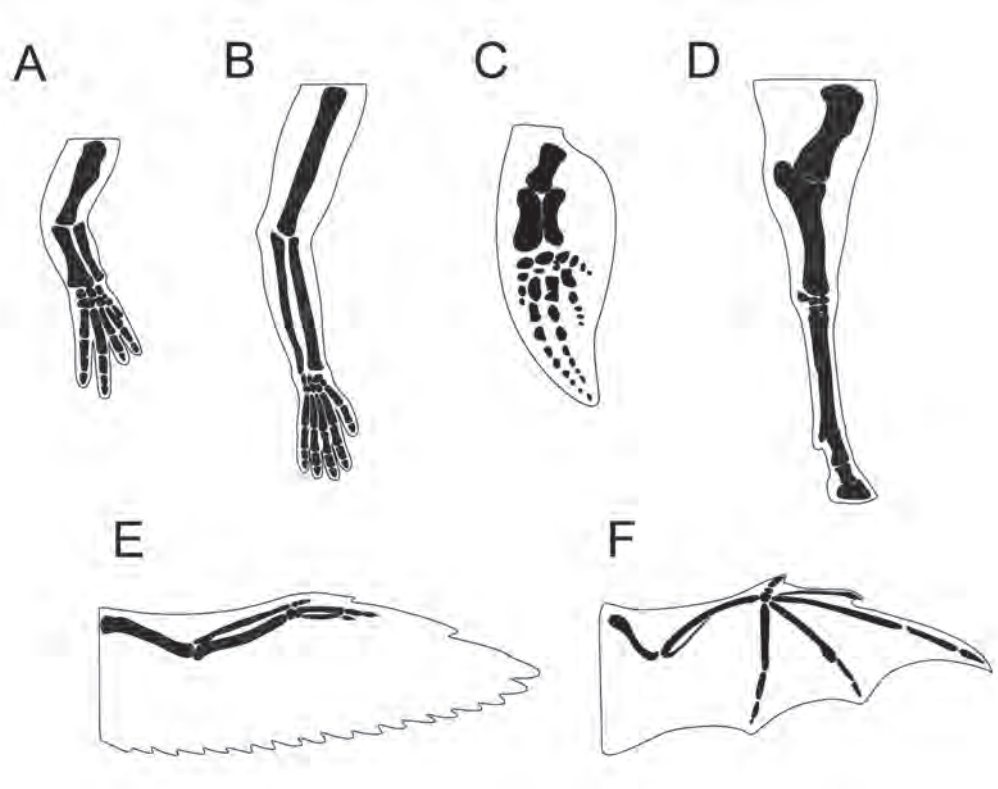


Fig. 7.1. Miembros anteriores de distintos grupos de tetrápodos. (A) reptil; (B) humano; (C) cetáceo; (D) caballo; (E) ave; (F) murciélago.

Un ejemplo de homología son los miembros de los grupos de tetrápodos que derivan todos de un miembro quiridio ancestral (Fig. 7.1).

La homología morfológica es la más discutida en la literatura, y a través de los años se han establecido al menos tres criterios para identificarla:

1. Similitud morfológica, topológica o estructural. Es el criterio más antiguo de todos para establecer homologías. Ya Owen (1843) lo utilizaba como criterio para establecerla.
2. Congruencia con las homologías de otros caracteres. Según algunos autores (Patterson 1988), este es el criterio más fuerte para establecer homologías.
3. No coexistencia. Dos caracteres homólogos no pueden coexistir en el mismo individuo. Cuando coexisten copias de más de un carácter homólogo en un mismo individuo se denomina homonimia (por ejemplo las patas de los animales segmentados).

HOMOPLASIA

Podemos decir que hay una homoplasia cuando dos o más taxones comparten caracteres similares o idénticos, y éstos no derivan del ancestro común (similares características con distinto origen evolutivo). El término fue acuñado por Lankester en 1870 e incluye a los paralelismos, las convergencias y las reversiones. A pesar de los 150 años transcurridos y los innumerables trabajos en donde se discute este concepto, la homoplasia sigue siendo un tema controversial. En la raíz de las controversias se encuentran la cuestión del origen evolutivo de la similitud entre los organismos y el término homología. Algunos autores consideran que las diferencias entre paralelismos, convergencias y reversiones son innecesarias e irrelevantes (por ejemplo, Nelson y Platnick 1981, Wiley 1981) y que son necesarios sólo dos conceptos: homologías y no homologías (homoplasias). Sin embargo, en las últimas décadas la homoplasia ha sido rescatada por diversos autores como un concepto importante dentro de la teoría evolutiva (por ejemplo, Sanderson y Hufford 1996). A pesar de las controversias, es posible distinguir entre los procesos de convergencia, paralelismo y reversión (Hall 2007).

La convergencia es la evolución independiente de características similares en diferentes líneas evolutivas, derivadas de diferentes caracteres ancestrales y con diferentes caminos del desarrollo (ontogenias). Un ejemplo son los ojos de los vertebrados y los ojos de los cefalópodos (Futuyma y Kirkpatrick 2017).

El paralelismo es la evolución independiente en líneas cercanamente relacionadas filogenéticamente, de características similares o idénticas, usualmente basadas en los mismos caminos de desarrollo. Por ejemplo la aparición de manera independiente de maxilipedios (apéndices torácicos con función alimenticia) en diferentes líneas de crustáceos.

La reversión constituye el retorno a una característica ancestral a partir de una derivada. Por ejemplo, los insectos actuales presentan alas y evolucionaron de ancestros sin alas, pero algunos grupos perdieron las alas secundariamente.

Se puede distinguir un paralelismo de una convergencia aplicando los siguientes criterios:

1. Si hay correspondencia estructural es un paralelismo; en caso contrario, es una convergencia.
2. Si hay características similares no ancestrales en taxones cercanos, es un paralelismo; en taxones lejanos es una convergencia.
3. Si hay características independientes causadas por una base génica y desarrollo compartidos o una predisposición ancestral, es un paralelismo; en caso contrario, es una convergencia.

Para algunos autores (Arendt y Reznick 2007) la distinción entre convergencia y paralelismo es una falsa dicotomía, pues representan los extremos de un continuo.

La Tabla 7.1 muestra la aplicación de los criterios a los distintos tipos de relaciones entre homología, paralelismo y convergencia.

Tabla 7.1. Criterios utilizados para distinguir entre homología, paralelismo, convergencia y homonimia. Modificada de Patterson (1988).

	Criterios			
Relación	**Congruencia**	**Similitud estructural**	**No coexistencia**	**Desarrollo (ontogenia)**
Homología	Positivo	Positivo	Positivo	Similar
Paralelismo	Negativo	Positivo	Positivo	Similar
Convergencia	Negativo	Negativo	Positivo	Diferente
Homonimia	Positivo	Positivo	Negativo	Similar

El concepto de analogía utilizado por Owen (1843) como contraposición a la homología, hoy día podría sinonimizarse con una convergencia funcional. Por ejemplo, las alas de los murciélagos, de las aves y de los pterosaurios son convergentes (funcionalmente) ya que no se heredaron de un ancestro común con alas. Sin embargo, son homólogas como miembros anteriores, ya que derivan de un ancestro común con miembro quiridio (Fig. 7.2).

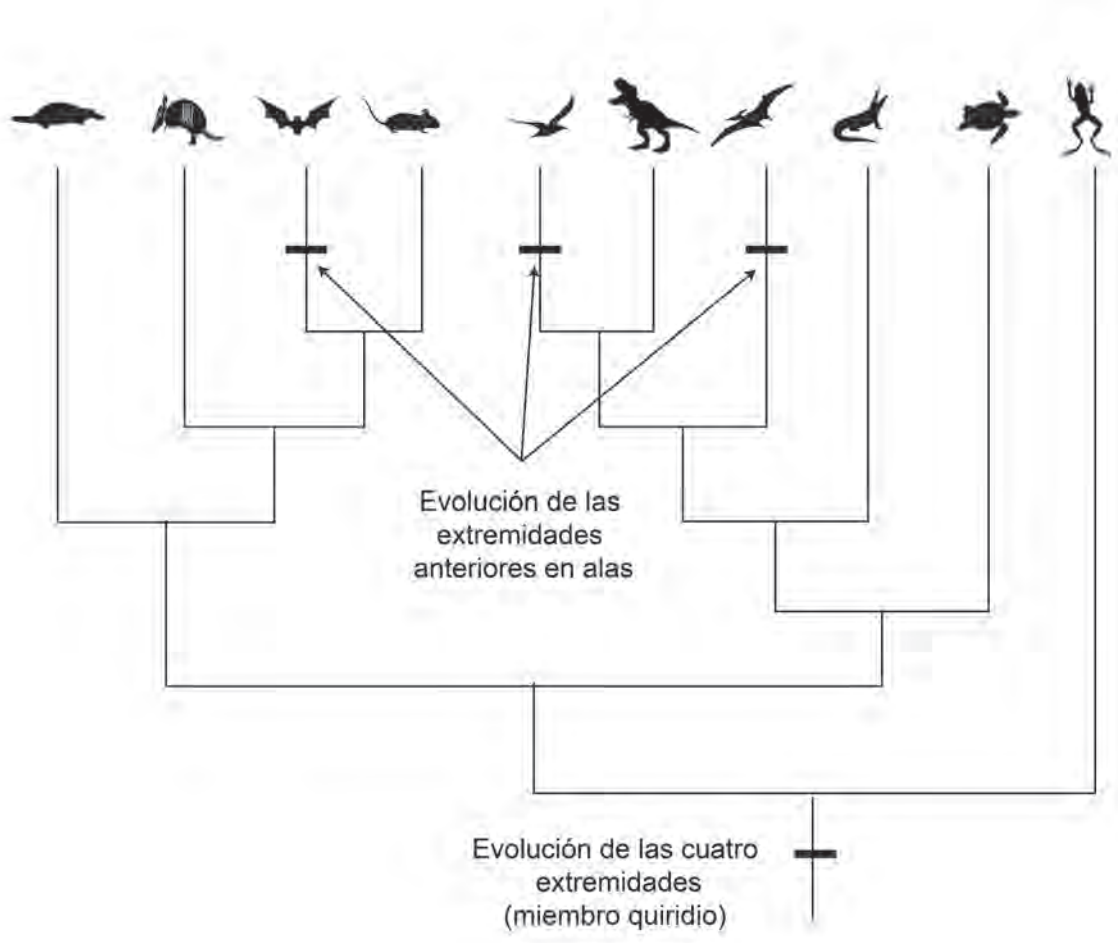


Fig. 7.2. Filogenia de los tetrápodos donde se muestra la evolución de las extremidades anteriores en alas como un caso de convergencia funcional.

LA HOMOLOGÍA EN LA BIOLOGÍA MOLECULAR

La homología y la homoplasia no sólo se definen para caracteres fenotípicos, sino también para otros caracteres como por ejemplo las secuencias de ADN. Los estudios filogenéticos se han visto revolucionados por los datos moleculares, que revelan la variación de miles e incluso millones de posiciones de pares de bases en secuencias homólogas de ADN.

Para analizar segmentos comparables de diferentes organismos, es necesario alinear las secuencias de modo tal que los sitios (posiciones) de nucleótidos sean homólogos (Forey *et al.* 1992, Schuh 2000). En este caso, el criterio de homología es el de la similitud entre los sitios, que impliquen un alineamiento con la menor cantidad de cambios posibles. Una vez alineadas las secuencias, cada posición (sitio) en una de las dos cadenas representa un carácter, y su identidad es uno de los cuatro nucleótidos (A, T, C, G), los que representan un estado de ese carácter.

Cuando comparamos secuencias de ADN de diferentes especímenes que han tenido un ancestro común, éstas pueden diferir entre sí, lo que significa que hubo cambios originados por mutaciones. Estos cambios pueden deberse a:

- Transiciones: sustituciones de nucleótidos en las que una purina se reemplaza por otra purina (adenina o guanina), o una pirimidina es reemplazada por otra pirimidina (timina o citosina).
- Transversiones: sustituciones de nucleótidos en las que una purina se reemplaza por una pirimidina, o viceversa.
- Inserciones/deleciones: cambios mutacionales donde se adiciona o pierde una base o fragmento de ADN. En ese caso, es preciso insertar *gaps* para alinear las secuencias. Así, al alinear las secuencias de modo manual, se tratan de minimizar las sustituciones y las deleciones simultáneamente, utilizando distintos criterios (Fig. 7.3).

Fig. 7.3. Tres modos distintos de realizar un alineamiento de secuencias de manera manual. En gris se muestran las sustituciones y las líneas representan los *gaps.*

Existen diversos métodos con distintos modelos de costos para alinear secuencias de ADN y software para aplicar esos métodos. Un tratamiento más profundo sobre el tema de alineamiento puede hallarse en Higgins y Lemey (2009).

La filogenia no sólo se puede aplicar a los organismos, sino también a sus genes. Hay casos en los que la filogenia de los genes puede no coincidir con la filogenia de los organismos a los que pertenecen dichos genes. Un ejemplo es cuando se produce la duplicación de un gen sin una especiación que la acompañe. A los genes producto de esta situación se los denomina genes parálogos. En los casos que la duplicación de un gen sea acompañada por un evento de especiación, a los genes se los denomina genes ortólogos.

Otra incongruencia entre la filogenia de organismos y genes se observa cuando hay transferencia horizontal de genes. Por ejemplo, un mismo gen se ha encontrado en algunas especies de felinos y monos

del viejo mundo (Catarrhini). Si construyéramos la filogenia de este gen obtendríamos como resultado que estos félidos están más relacionados con los monos que con otros félidos, lo que es claramente incongruente con el resto de los genes y la morfología. La explicación para este fenómeno es que el gen ha sido transferido de los monos a los félidos por un virus (Li y Graur 1991). A este tipo de gen que se transfiere entre especies se lo denomina gen xenólogo (Koonin *et al.* 2001). Esto sugiere que para que una filogenia de organismos sea confiable, se deben usar caracteres homólogos y genes ortólogos.

POLARIDAD: DIRECCIÓN DEL CAMBIO EVOLUTIVO

Una vez elegidos y codificados los caracteres, es necesario establecer la polaridad o dirección del cambio evolutivo. Para reconocer la polaridad es preciso determinar el estado ancestral (plesiomórfico) de cada carácter utilizado y en consecuencia, reconocer el estado o los estados derivados (avanzados, apomórficos o evolucionados), con el objetivo de enraizar el árbol filogenético.

El estado ancestral de un carácter es aquel que se halla o se infiere que se hallaba en el ancestro común más reciente del grupo, cuya historia evolutiva se está determinando (Crisci y Stuessy 1980, Fernández *et al.* 2005). Este concepto de ancestralidad es relativo: el estado de un carácter puede ser ancestral en un taxón particular, pero no necesariamente en otro. Por lo tanto, el uso del término ancestral, y por ende el de derivado, carece de sentido si no se refiere a un grupo o taxón particular.

El criterio para determinar el estado plesiomórfico de un carácter es el de la comparación con el o los grupos externos (GE), también denominados *outgroups*. Por ejemplo, supóngase que se está intentando reconstruir la filogenia de un género y el carácter color de la flor varía dentro del mismo con dos estados, blanco y azul. Para determinar cuál de estos dos estados es el ancestral debemos ir al género más cercano al estudiado y observar cuál es el color de la flor, supongamos blanco. De esto deduciríamos que en el género en estudio, el blanco es el estado ancestral. El mejor grupo externo para un determinado taxón es aquel con el que comparte un ancestro que sólo los origina a ellos dos.

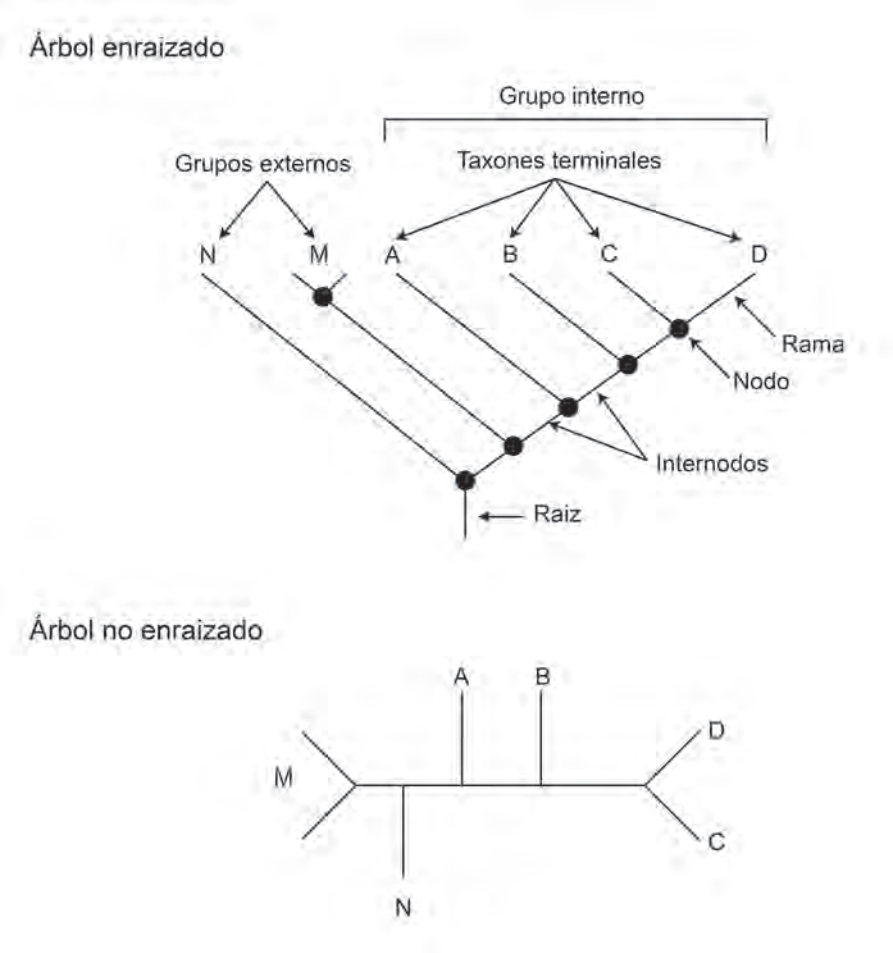
El estado de un carácter apomórfico que se encuentra en dos o más taxones se denomina sinapomorfía, el cual se considera que surgió del ancestro más reciente de dichos taxones. Una autapomorfía es un estado apomórfico en un solo taxón. Una simplesiomorfia es el estado de un carácter plesiomórfico que se encuentra en dos o más taxones. Estos términos no son absolutos, sino que dependen del nivel sistemático en el que se realiza el análisis filogenético.

ÁRBOLES FILOGENÉTICOS: TERMINOLOGÍA Y CONCEPTOS BÁSICOS

Los árboles filogenéticos, también llamados cladogramas, son representaciones de estructuras jerárquicas que simbolizan las relaciones evolutivas entre los organismos o los genes. Los árboles pueden estar enraizados (cuando se utiliza el grupo externo para polarizar los caracteres) o no enraizados (cuando no se aplica el concepto de polaridad en la interpretación del árbol). En el Box 7.2 se describen los términos vinculados a los árboles filogenéticos.

Los árboles pueden ser completamente dicotómicos (resueltos), en los que cada nodo está conectado con dos nodos, o con taxones terminales, o con un nodo y un taxón terminal; o politómicos (parcialmente resueltos), en el que uno o más nodos están conectados con tres o más nodos o taxones terminales. Una politomía puede representar dos situaciones: divergencia simultánea de varios taxones descendientes (por ejemplo radiación adaptativa) o incertidumbre en las relaciones filogenéticas entre los taxones incluidos en la politomía. La Figura 7.4 muestra un árbol totalmente dicotómico y un árbol parcialmente politómico.

Box 7.2. Terminología básica de los árboles filogenéticos



Topología: estructura de las relaciones entre los taxones terminales.
Raíz: base o punto de partida del árbol filogenético.
Nodo: punto de ramificación del árbol, representa un ancestro hipotético de otros nodos o de taxones terminales a los que origina.
Grupo hermano: grupo que comparte un nodo que sólo los origina a ellos (por ejemplo, B es el grupo hermano del grupo C-D).
Grupo interno (*ingroup*): conjunto de taxones terminales bajo estudio.
Grupo externo (*outgroup*): taxón o taxones terminales utilizados como grupo hermano del grupo interno, con el objetivo de enraizar el árbol.
Internodo (rama interna): segmento que une a dos nodos contiguos.
Rama: segmento que une nodos con taxones terminales.
Taxón terminal: unidad del análisis que se sitúa en el extremo de una rama.

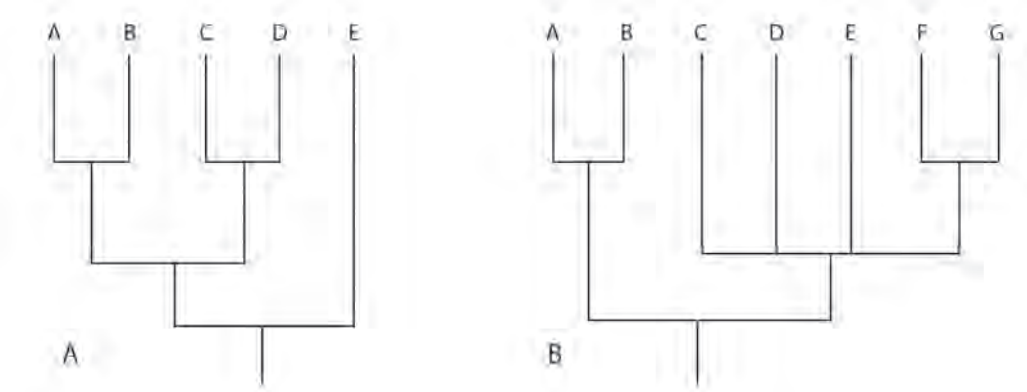


Fig. 7.4. (A) Árbol completamente dicotómico; (B) árbol con politomías.

Los árboles pueden ser representados de distintas formas sin modificar sus relaciones, siguiendo las reglas mostradas en la Figura 7.5.

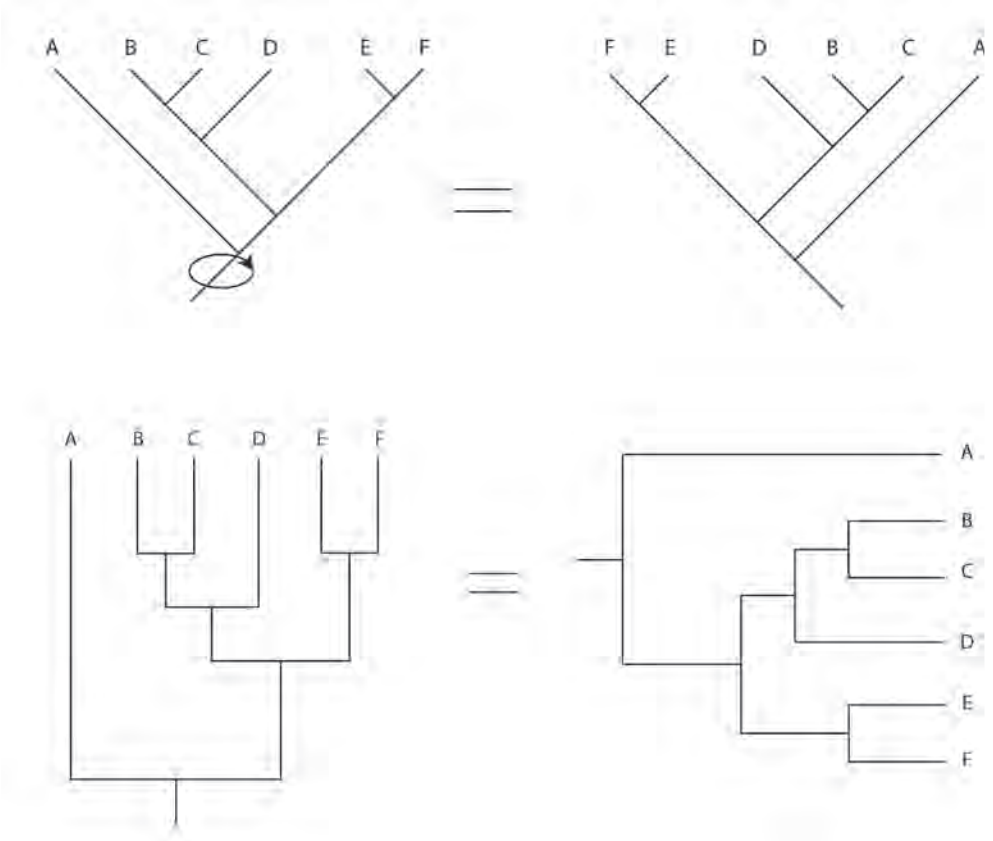


Fig. 7.5. Diferentes formas de representar un mismo árbol.

A su vez, las longitudes de las ramas de los árboles pueden ser proporcionales a la cantidad de cambios evolutivos que se producen a lo largo de ellas (Fig. 7.6).

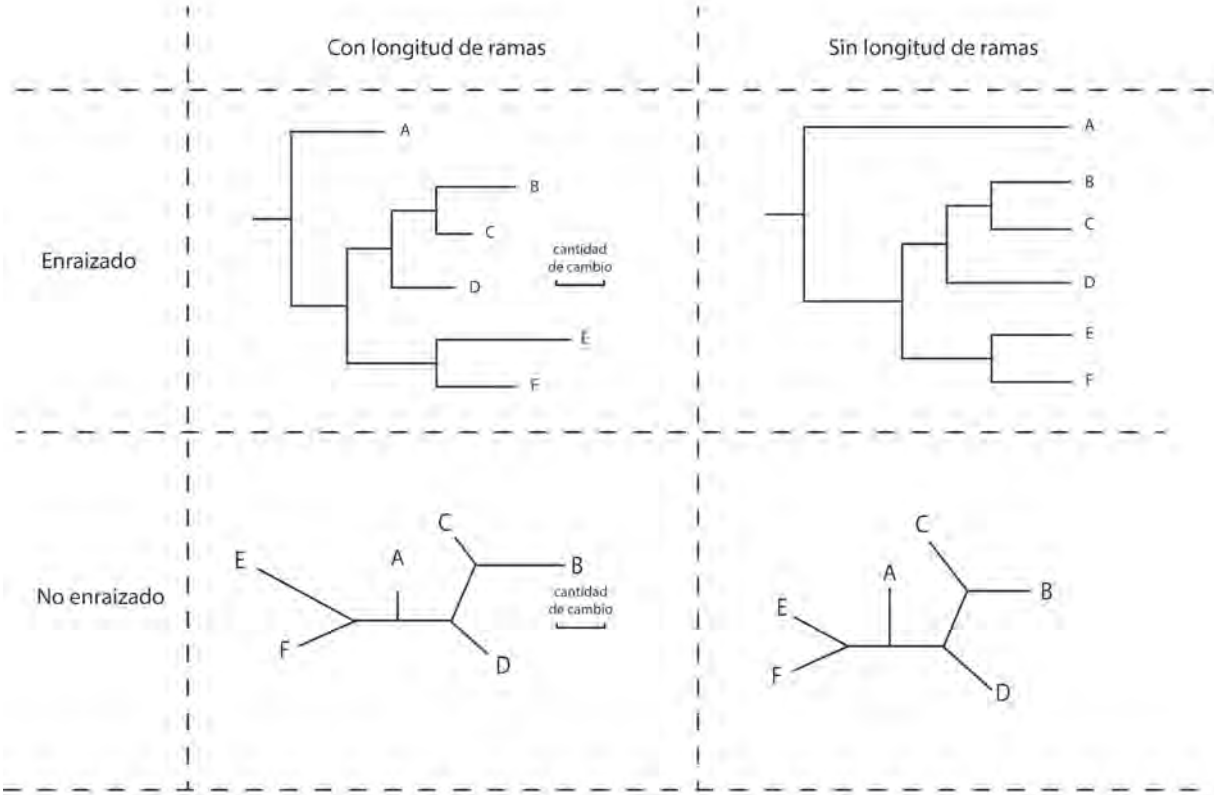


Fig. 7.6. Árboles filogenéticos enraizados, no enraizados, con y sin longitudes de las ramas.

Un árbol no enraizado se puede enraizar de forma imaginaria suponiendo que tiramos del punto donde queremos ubicar la raíz (Fig. 7.7). La Figura 7.8 muestra cinco relaciones filogenéticas distintas entre los taxones (todas derivadas del mismo árbol no enraizado). Se puede observar la importancia de la posición de la raíz, que determina la topología final del árbol sin cambiar su longitud (cantidad de cambios).

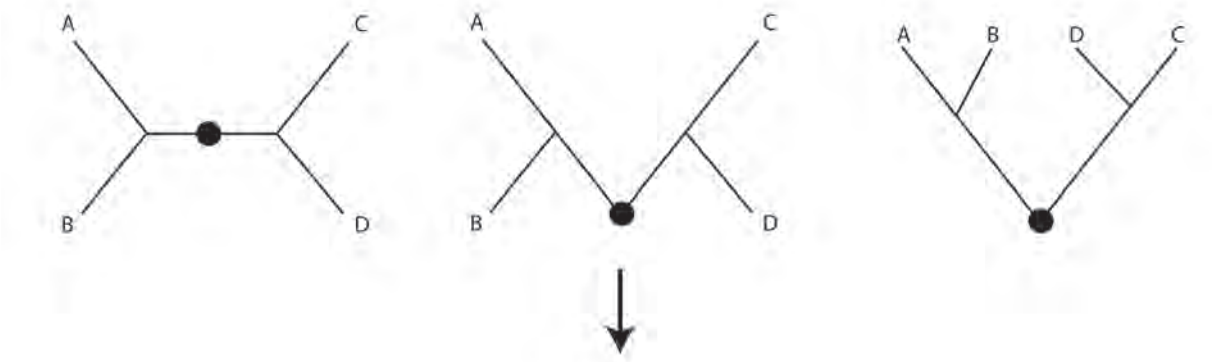


Fig. 7.7. Ejemplo hipotético del modo en que se enraíza un árbol filogenético.

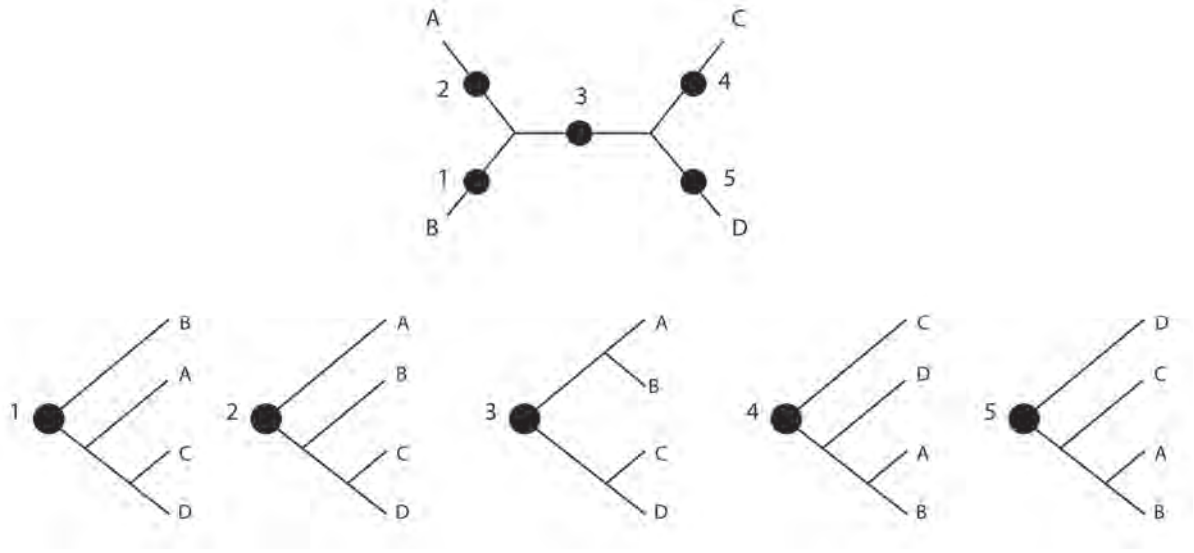


Fig. 7.8. Formas posibles de enraizar un árbol de cuatro taxones. Los números representan distintas posiciones posibles de la raíz.

Los grupos formados a partir del árbol filogenético pueden ser monofiléticos, parafiléticos o polifiléticos. Un grupo monofilético contiene al ancestro y a todos sus descendientes, por lo tanto está definido por una o más sinapomorfías (Fig. 7.9A). Un grupo parafilético contiene al ancestro pero no a todos sus descendientes, por lo tanto está definido por una o más simplesiomorfías (Fig. 7.9B). Un grupo polifilético es un grupo formado a partir de dos ancestros, por lo tanto está definido por una o más homoplasias (Fig. 7.9C).

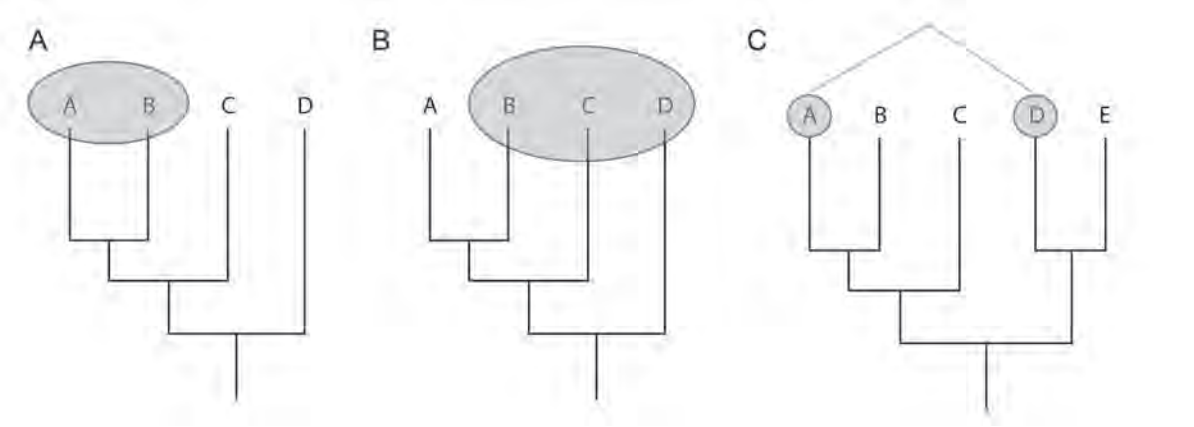


Fig. 7.9. Tipos de grupos basados en la filogenia: (A) grupo monofilético; (B) grupo parafilético; (C) grupo polifilético.

MÉTODOS DE ESTIMACIÓN FILOGENÉTICA

Existen al menos cuatro métodos de estimación filogenética:

- Parsimonia (simplicidad)
- Métodos de distancia

- Máxima verosimilitud
- Análisis bayesiano

Todos los métodos filogenéticos comparten los siguientes postulados básicos:

1. La naturaleza tiene una estructura jerárquica.
2. Esa estructura jerárquica puede representarse mediante árboles filogenéticos.
3. La estructura jerárquica de la naturaleza puede rescatarse mediante un muestreo de caracteres.

PARSIMONIA

La palabra *parsimony* en el contexto de árboles filogenéticos fue utilizada por primera vez por Camin y Sokal (1965), y en español se traduce como parsimonia. Los postulados específicos de este método, que se suman a los tres postulados básicos mencionados anteriormente, son los siguientes:

4. Los grupos se forman en función de la posesión de novedades evolutivas en común (sinapomorfías).
5. Una vez construidas las hipótesis (cladogramas = árboles filogenéticos) se elige aquella que presente menor cantidad de pasos evolutivos (parsimonia).

Este método fue originalmente propuesto por Hennig (1950, 1968) y profundizado por autores posteriores (por ejemplo, Kluge y Farris 1969, Farris *et al*. 1970, Wiley 1981, Farris 1983, Goloboff 1998). Una posible historia ilustrativa de los postulados de la parsimonia es la que se presenta en la Tabla 7.2.

Tabla 7.2. Historia ilustrativa, no exhaustiva, de los postulados fundamentales de la sistemática filogenética. El postulado 5 se le adjudica a un principio auxiliar de Hennig (Farris y Kluge 1985), que algunos autores no aceptan (Duncan 1984) y proponen que la parsimonia fue planteada por primera vez por Edwards y Cavalli-Sforza (1964).

Autor	Año	Postulados
Darwin	1859	1 y 2
Müller	1864	1, 2 y 3
Haeckel	1866	1 y 2
Ameghino	1884	1, 2 y 3
Rosa	1918	1, 2, 3 y 4
Hennig	1950	1, 2, 3, 4 y 5

El método de parsimonia es de simple aplicación manual cuando los taxones terminales son pocos, pero de una imposible aplicación manual cuando los taxones superan un cierto número. Con más de 30 taxones (aproximadamente) es imposible garantizar una solución óptima, incluso computacionalmente. Esto se debe a que el número posible de árboles aumenta de manera exponencial con el aumento del número de taxones terminales, como muestra la Figura 7.10 y la Tabla 7.3 para árboles bifurcados.

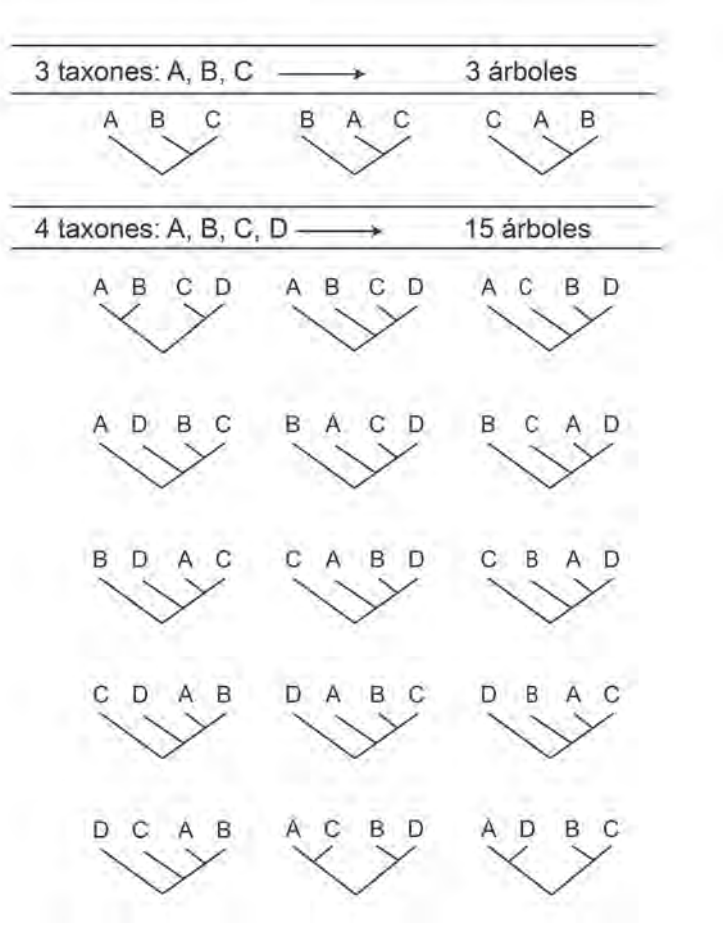


Fig. 7.10 Árboles bifurcados posibles para tres y cuatro taxones.

Tabla 7.3. Número de árboles bifurcados en función del número de taxones.

N° de taxones	N° de árboles bifurcados
3	3
4	15
5	105
6	945
7	10395
8	135135
9	2027025
10	34459425
30	$4{,}9518 \times 10^{38}$
40	$1{,}00985 \times 10^{57}$

En los casos de más de 30 taxones la búsqueda del árbol más parsimonioso se denomina búsqueda heurística y el resultado es aproximado, no exacto. A este problema se lo denomina problema del viajante o *Travelling Salesman Problem* (Prim 1957) que intenta responder la siguiente pregunta: dada una lista de ciudades y las distancias entre ellas ¿cuál es la ruta más corta posible que visita cada ciudad una vez y al finalizar regresa a la ciudad de origen? Este es el ejemplo de un problema que corresponde a los denominados problemas NP-Completos donde el tiempo de resolución del problema no depende de una función polinómica sino de funciones que crecen más rápidamente (como una función exponencial), y difícilmente se pueda resolver computacionalmente (Graham y Foulds 1982).

Un ejemplo de aplicación manual del método de parsimonia se muestra en la Figura 7.11 expresado para un carácter, y en la Figura 7.12 expresado para cuatro caracteres superpuestos sobre los árboles. La línea simple representa una sinapomorfía, mientras que la línea doble representa un paralelismo. En todos los casos, el árbol más parsimonioso es el que tiene la menor cantidad de pasos y es por lo tanto el elegido.

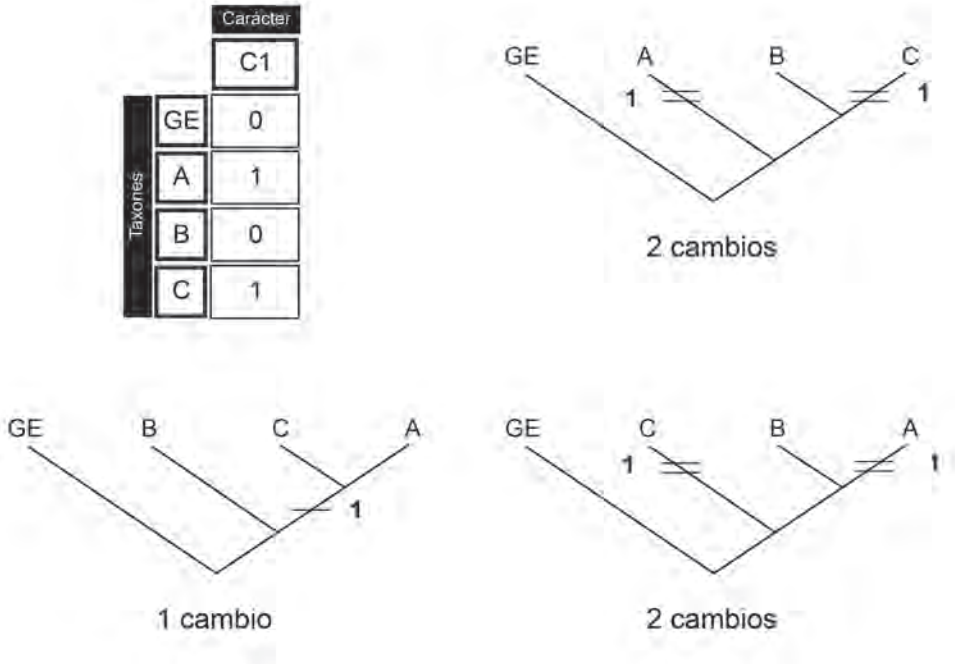


Fig. 7.11. Ejemplo simple de aplicación del método de parsimonia para tres taxones (A a C), el *outgroup* (GE) y un carácter (1).

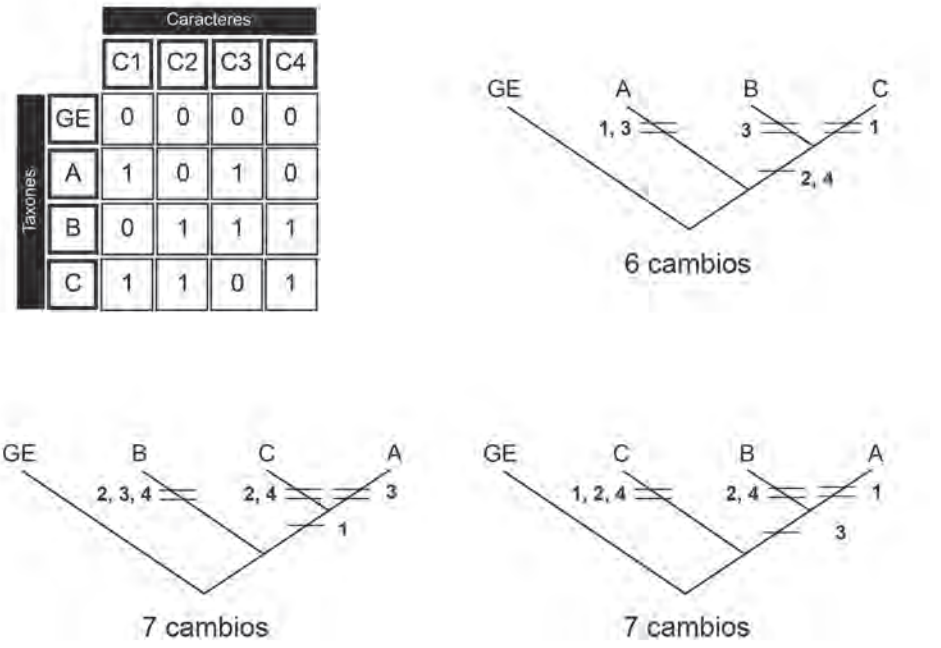


Fig. 7.12. Ejemplo simple de aplicación del método de parsimonia para tres taxones (A a C), el *outgroup* (GE) y cuatro caracteres (1 a 4).

En la Figura 7.13 se puede apreciar un mapa de conceptos de los pasos en la aplicación del método de parsimonia (1 a 4).

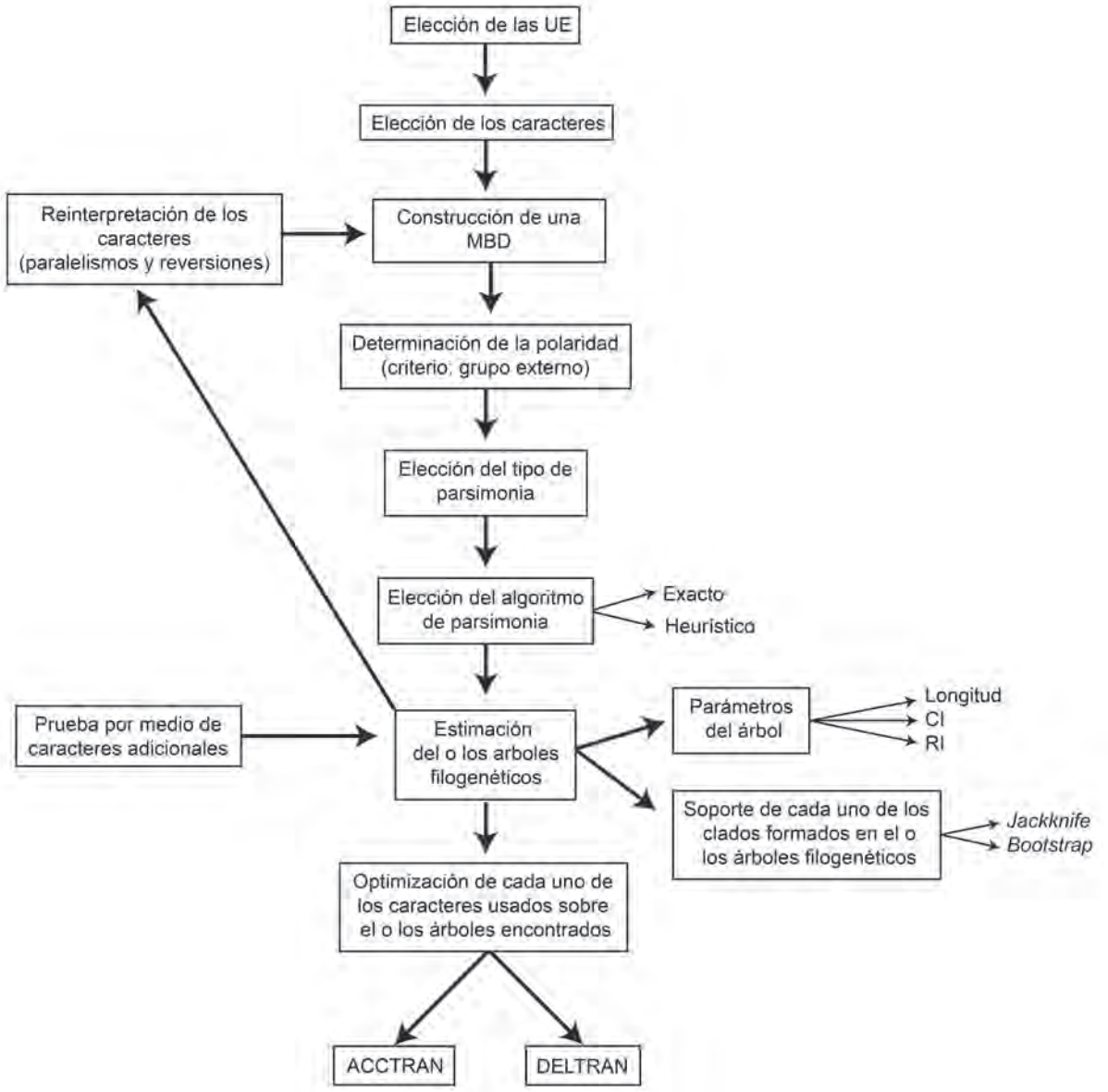


Fig. 7.13. Mapa de conceptos de los pasos necesarios para realizar un análisis de parsimonia. CI: índice de consistencia, RI: índice de retención.

Tipos de Parsimonia

Existen al menos cuatro variantes de la parsimonia que valoran de distinta forma los cambios entre estados de caracteres, en términos de "pasos" o eventos evolutivos: Wagner, Fitch, Dollo y Camin-Sokal (Forey *et al.* 1992). Una quinta variante es la propuesta por Sankoff (1975), utilizada sólo para datos moleculares.

Parsimonia de Wagner

Las posibilidades de cambios en ambos sentidos son iguales. Permite que los estados de un carácter reviertan a su condición ancestral y la aparición en paralelo de un estado de carácter. Por cada cambio se suma un solo paso. Los caracteres multiestado son considerados aditivos u ordenados, por lo que el cambio de 0 a 1 vale un paso, el cambio de 1 a 2 también vale un paso (Fig. 7.14) y el cambio de 0 a 2 vale 2 pasos. Este tipo de parsimonia fue formalizada por Farris (1970), quien se basó en el trabajo de Wagner (1961).

Parsimonia de Fitch

Este tipo de parsimonia fue formalizada por Fitch (1971) y es similar a la parsimonia de Wagner, ya que permite que los estados de los caracteres reviertan o aparezcan en paralelo. Sin embargo, difiere en que los caracteres multiestado son desordenados o no aditivos. Cada cambio de un estado puede variar a cualquier otro estado, por lo que todas las transformaciones tienen un mismo costo, que es igual a 1 (Fig. 7.14). En el caso de caracteres doble estado (0/1), la parsimonia de Wagner genera el mismo resultado que la de Fitch.

Parsimonia de Dollo

Este tipo de parsimonia fue propuesta por Farris (1977) utilizando como modelo la regla de Dollo (1893). Bajo este tipo de parsimonia cada estado derivado de un carácter puede ganarse sólo una vez (los paralelismos de estados de caracteres se penalizan con altos valores de cambio). Esto significa que los caracteres complejos evolucionan una sola vez y que toda homoplasia se considera como una pérdida secundaria. De este modo, se permite el cambio de 0 a 1 una sola vez, pero se admite cualquier número de reversiones de 1 a 0.

Parsimonia de Camin-Sokal

Este tipo de parsimonia fue propuesta por Camin y Sokal (1965), los estados derivados de un carácter pueden ser ganados cuantas veces sea necesario, pero la pérdida (reversiones) de ese estado está prohibida. Esto significa que por ejemplo para un carácter dado, éste puede pasar del estado 0 al 1 cuantas veces sea necesario, pero nunca se puede pasar del estado 1 al 0. Todas las homoplasias son debidas a paralelismos (Fig. 7.14).

	Wagner				Fitch				Dollo				Camin -Sokal			
	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0	–	1	2	3	–	1	1	1	–	M	2M	3M	–	1	2	3
1	1	–	1	2	1	–	1	1	1	–	M	2M	∞	–	1	2
2	2	1	–	1	1	1	–	1	2	1	–	M	∞	∞	–	1
3	3	2	1	–	1	1	1	–	3	2	1	–	∞	∞	∞	–

Fig. 7.14. Comparación de matrices de costo para las simplicidades de Wagner, Fitch, Dollo y Camin-Sokal. El cambio de un estado a otro se lee por filas. Para Wagner, el costo entre estados es una serie acumulativa donde los estados tienen un orden. Para Fitch, el costo entre dos estados siempre es el mismo (los estados no tienen un orden). En el caso de Dollo, los cambios de un estado ancestral a uno derivado sólo pueden ocurrir una vez en el árbol, penaliza los paralelismos con un costo alto (M) para que en lo posible ocurran sólo una vez, ya que no se pueden prohibir absolutamente. En el caso de Camin-Sokal se permite la cantidad de paralelismos que sean necesarios, mientras que se prohíben las reversiones a un estado ancestral, asignandole un valor infinito.

Parsimona de Sankoff

Una posible aplicación de matrices de costo para datos de secuencias de ADN fue propuesta por Sankoff (1975). El costo de la transversión es igual a 2,5 y el costo de la transición es igual a 1 (Fig. 7.15). Estos costos pueden ser modificados por el usuario, generando su propio tipo de parsimonia.

	A	C	G	T
A	0	2,5	1	2,5
C	2,5	0	2,5	1
G	1	2,5	0	2,5
T	2,5	1	2,5	0

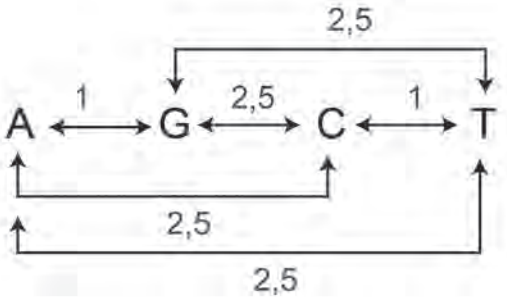


Fig. 7.15. Matriz de costo para cambios en secuencias de ADN aplicando la parsimonia de Sankoff. A: adenina, C: citosina, G: guanina, T: timina.

Métodos computacionales para hallar el o los árboles más parsimoniosos

Tal como se discutió anteriormente, el problema computacional de la estimación filogenética hace que en los casos de más de 30 taxones, sea imposible garantizar que el o los árboles encontrados sean los más parsimoniosos. Para menos de 30 taxones se ha propuesto un método computacional denominado *branch and bound* (Felsenstein 2004), que asegura encontrar el o los árboles más parsimoniosos. Otro método que permite encontrar el o los árboles más parsimoniosos es el denominado exhaustivo (Cigliano *et al.* 2005) que sólo puede ser utilizado cuando se analizan menos de 10 taxones.

Para un mayor número de taxones existen los métodos heurísticos, que no garantizan el hallazgo del o de los árboles más cortos. Estos métodos buscan los árboles más cortos a base de prueba y error, usando como punto de partida uno o más árboles iniciales. El método aplica un procedimiento de permutación de ramas (*branch swapping*) que consiste en mover las ramas del árbol inicial a diferentes partes del mismo, medir la longitud del árbol y guardar aquellos que posean igual o menor longitud.

Los métodos de permutación de ramas más comunes son (Swofford 2002): TBR (*tree bisection and reconnection*), NNI (*nearest neighbor interchange*), SPR (*subtree pruning and regrafting*) y, para matrices de gran tamaño, el método de ratchet (Nixon 1999). En la Figura 7.16 se resumen los métodos de búsqueda del árbol más parsimonioso y en la Figura 7.17 se resumen los métodos de permutación de ramas. El método de NNI intercambia las conexiones entre subárboles dentro de un mismo árbol; dado que hay tres formas posibles de combinar cuatro subárboles, y una corresponde al árbol original, cada intercambio crea dos nuevos árboles. En SPR y TBR se divide el árbol en dos subárboles y se conectan todos los pares de ramas entre los dos subárboles. En SPR los dos subárboles siempre se unen por el mismo nodo de uno de los subárboles, mientras que en TBR la unión entre los dos subárboles se da en todos los nodos posibles (Fig 7.17).

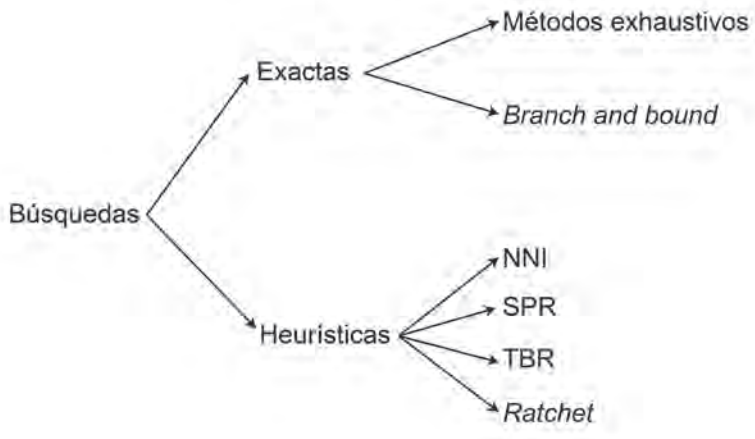


Fig. 7.16. Métodos de búsqueda del árbol más parsimonioso.

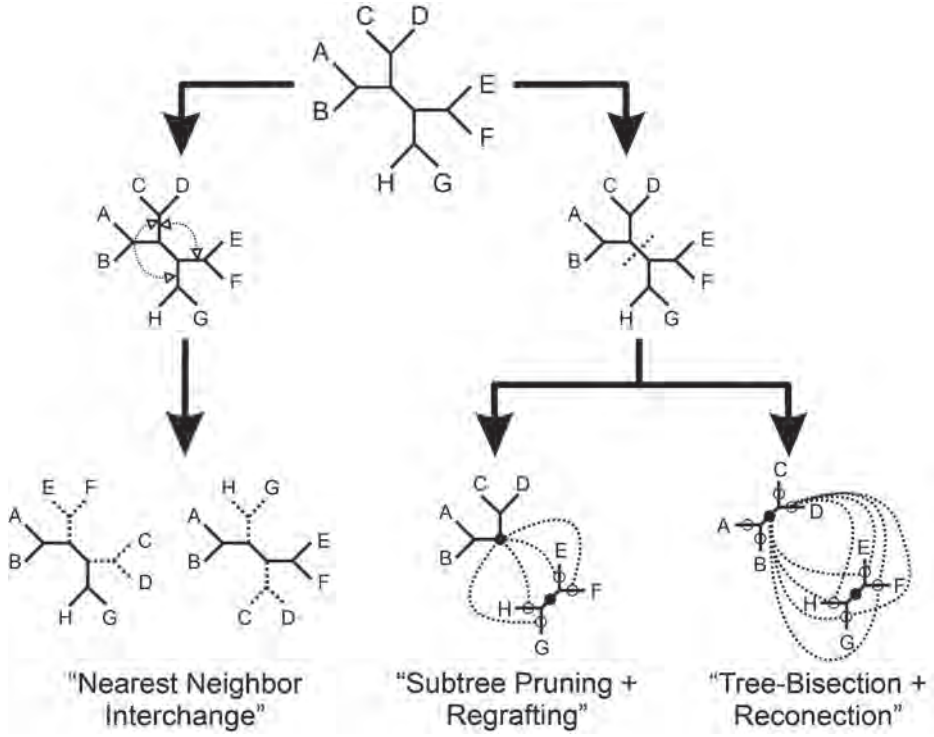


Fig. 7.17. Métodos de permutación de ramas. Las líneas punteadas conectan todos los pares de ramas entre los dos subárboles, excepto los círculos llenos (ya que daría como resultado el árbol completo). Modificada de Schmidt y von Haeseler (2009).

Optimización de los caracteres sobre el o los árboles encontrados

Una vez hallado el árbol más parsimonioso, es posible analizar la evolución de los caracteres utilizados superponiéndolos sobre el mismo. Una opción consiste en favorecer a los paralelismos por sobre las reversiones (en el caso que sean equivalentes en la cantidad de cambios que implican). De esta forma, se retrasan las transformaciones que tengan homoplasias en la topología del árbol, técnica denominada DELTRAN (Swofford y Maddison 1987). La opción opuesta es preferir las reversiones por sobre los paralelismos, que es el método de optimización original de Farris (1970). En este caso, los cambios se dan siempre cercanos a la raíz del árbol, procedimiento denominado ACCTRAN (Forey *et al.* 1992). En cualquiera de los dos casos la longitud del árbol no se ve afectada. ACCTRAN es el más utilizado en bases biológicas debido a que los caracteres complejos se supone que aparecen una única vez (de Pinna 1991). Sin embargo, Agnarsson y Miller (2008) sugieren que cada carácter debe ser analizado de manera independiente para decidir entre ambas opciones (Fig. 7.18).

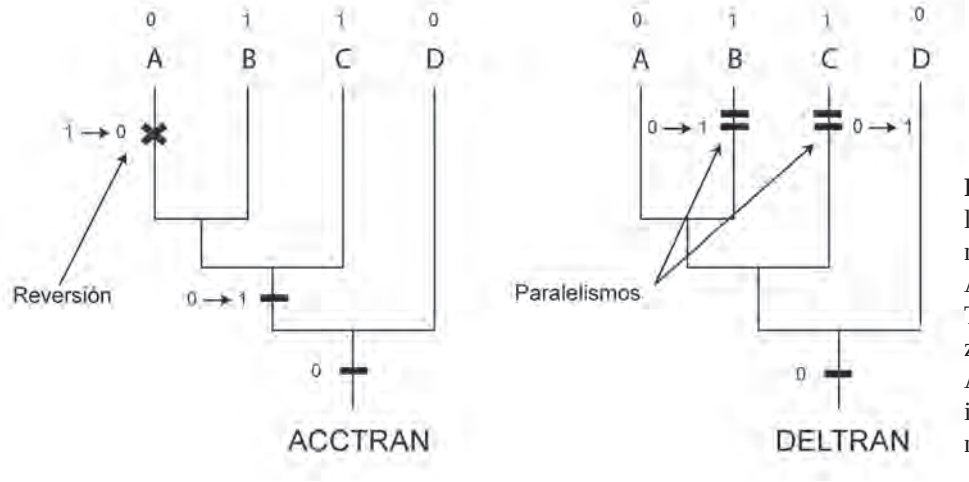


Fig. 7.18. Árboles filogenéticos hipotéticos mostrando los métodos ACCTRAN y DELTRAN para la optimización de caracteres. Ambas opciones son igualmente parsimoniosas.

Parámetros del árbol

El árbol hallado se define en función de la cantidad de cambios o pasos entre los estados de los caracteres que sustentan sus relaciones. Cuanto mejor sea el ajuste de los datos al árbol, menor será la cantidad de homoplasias necesarias. La longitud del árbol corresponde a la cantidad de cambios requeridos. El árbol que minimiza la cantidad de homoplasias es el más parsimonioso.

Un criterio para evaluar el ajuste global del árbol es el índice de consistencia (CI; Farris 1969b). Éste se calcula como CI = M/S, donde M es la mínima cantidad de pasos posibles y S es el número de pasos observados en el árbol. La mínima cantidad de pasos se calcula suponiendo que todos los caracteres son sinapomorfías o autapomorfías, sin homoplasias. Este índice varía entre 0 y 1 (ajuste perfecto). El CI aumenta artificialmente si se cuentan las autapomorfías de los taxones terminales y/o los caracteres no informativos (Forey *et al.* 1992).

Otra medida de la cantidad relativa de homoplasia para el árbol es el índice de retención (RI), que mide la cantidad de homoplasia observada en función de la máxima homoplasia posible en los datos (Farris 1989). Éste se calcula como RI = $(G - S)/(G - M)$, donde G es el número máximo de cambios que podría tener el carácter en el árbol; S es el número de pasos observados en el árbol y M es la mínima cantidad de pasos posibles. Este índice varía entre 0 (peor ajuste) y 1 (ajuste perfecto). A continuación, se presenta el cálculo de los índices de consistencia y retención para el árbol más parsimonioso de la Figura 7.12. M corresponde al cambio mínimo posible, que en el caso de los caracteres doble estado (0-1) es igual al número de caracteres (en el ejemplo el cambio mínimo es 4). S corresponde al número de pasos observados en el árbol. G corresponde al cambio máximo posible, que en el caso de los caracteres doble estado (0-1) es igual al número de caracteres por el número de taxones (en el ejemplo el cambio máximo es 12). Por lo tanto, CI = 4/6 = 0,67 y RI = (12 – 6)/(12 – 4) = 0,75.

Soporte de los grupos formados en el árbol

Se han propuesto pruebas estadísticas para medir el grado de confianza de los grupos hallados en los árboles filogenéticos. Entre ellas están el *bootstrap* (Efron 1979), aplicado por primera vez por Felsenstein (1985) y el *jackknife,* propuesto por Quenouille (1956) y aplicado por primera vez a filogenia por Mueller y Ayala (1982).

En el *bootstrap*, que es el método más ampliamente utilizado, se extraen de forma aleatoria con reemplazo tantas columnas como caracteres tenga la MBD. Por lo tanto, un carácter puede estar presente más de una vez, o ninguna vez en la nueva matriz (denominada pseudo-réplica). A partir de cada pseudo-réplica se reconstruye un árbol (Fig. 7.19A). Este procedimiento se repite como mínimo 100 veces. El porcentaje de veces que aparece un grupo en los árboles resultantes es una medida del soporte del grupo. Por ejemplo, si un grupo aparece en todos los árboles reconstruidos, entonces el valor de soporte de *boostrap* es del 100%. Se considera que un valor menor a 70% debería tratarse con cautela en términos de soporte de grupos.

Similar al *bootstrap* es el *jackknife*, donde se extrae un porcentaje de los caracteres aleatoriamente y sin reemplazo (Fig. 7.19B). La cantidad de caracteres que se extrae generalmente varía entre 1 y la mitad del número de caracteres (*delete-half jackknife*; Felsenstein 2004).

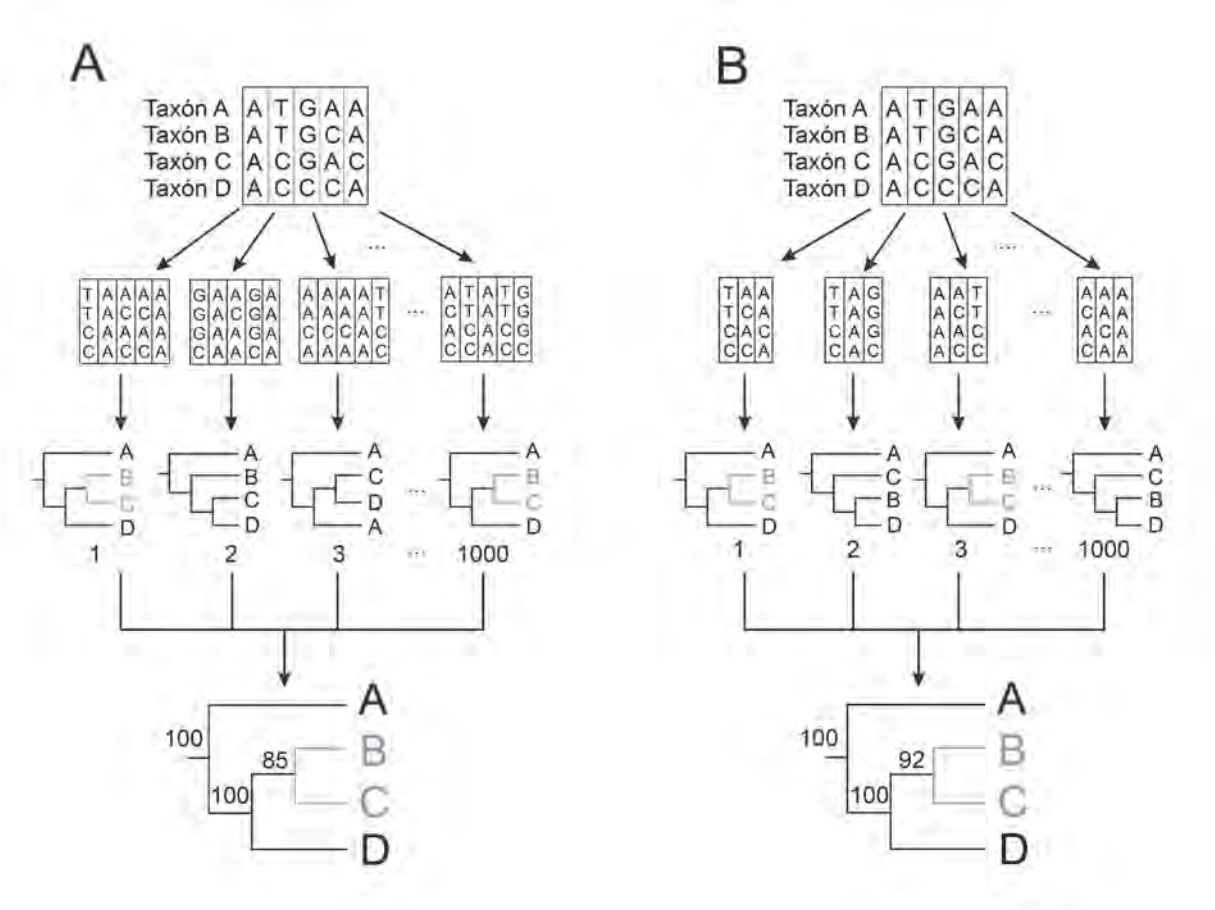


Fig. 7.19. (A) *Bootstrap*; (B) *jackknife*. Los números en los internodos representan los valores (en porcentaje) de soporte de los grupos.

Árboles de consenso

En múltiples ocasiones es posible obtener como resultado más de un árbol óptimo. Por ejemplo, en el caso de la parsimonia es posible obtener más de un árbol de la misma longitud. También puede darse el caso de obtener árboles óptimos diferentes como resultado de utilizar distintos tipos de caracteres (por ejemplo, datos morfológicos *vs.* moleculares). En estos casos se han propuesto los árboles de consenso como una manera de resumir la información compartida por estos árboles.

Existen varios métodos de árboles de consenso (Miyamoto 1985, Page 1996, Felsenstein 2004), pero los más usados son el consenso estricto y el consenso de mayoría. En el método de consenso estricto, se construye un árbol que contiene sólo los grupos presentes en todos los árboles comparados. El consenso de mayoría contiene a aquellos grupos que se repiten en más del 50% de los árboles. En la Figura 7.20 se muestra un ejemplo de tres árboles generados con la misma matriz (o alternativamente, árboles generados con distintas matrices de datos, pero con los mismos taxones) a los que se les aplican los métodos de consenso estricto y de mayoría. Cuando los árboles comparados son sólo dos, el consenso de mayoría es igual al consenso estricto.

Existen otros métodos de consenso como, por ejemplo, el consenso reducido (Adams 1972, Wilkinson 1995), que recobra estructuras comunes entre los árboles originales cuando uno o pocos taxones tienen diferentes posiciones en ellos. Estos árboles de consenso reducido pueden contener grupos que no aparecen en ninguno de los árboles originales y que no están sustentados por caracteres. Un método de consenso reducido es el de Adams (1972), que consiste en colocar los taxones conflictivos en la base del árbol y tratar de rescatar, sin la participación de estos taxones, las estructuras comunes del conjunto de árboles que se están comparando, como se muestra en la Figura 7.21.

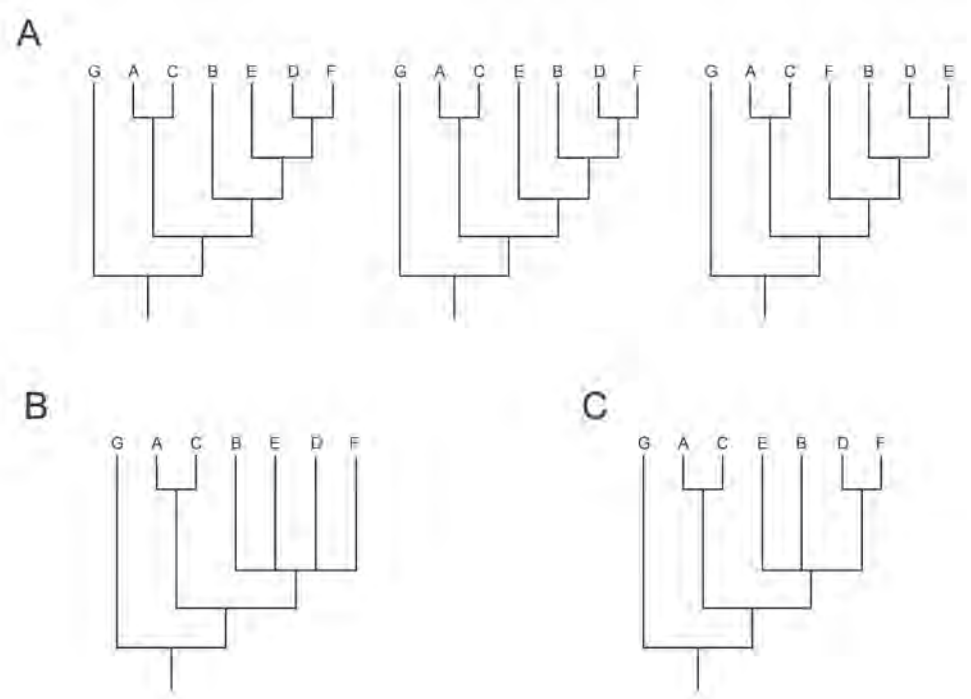


Fig. 7.20. (A) Conjunto de tres árboles igualmente parsimoniosos obtenidos a partir de la misma MBD; (B) consenso estricto; (C) consenso de mayoría.

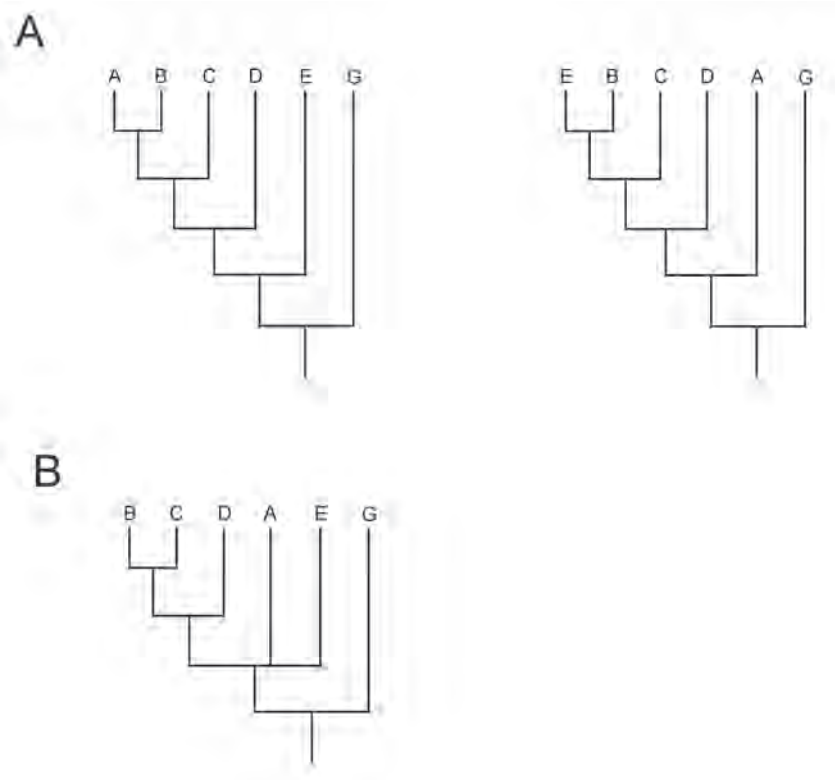


Fig. 7.21. (A) Conjunto de dos árboles igualmente parsimoniosos obtenidos a partir de la misma MBD, donde sólo los taxones A y E difieren en su posición, lo que causaría un consenso estricto irresuelto, hecho que se evita con un consenso reducido; (B) consenso reducido.

MÉTODOS DE DISTANCIA

Por último, mencionaremos brevemente los métodos de distancia aplicados a datos moleculares. Éstos se basan en el supuesto de que las diferencias entre las secuencias, medidas como el número de cambios acumulados, pueden representarse mediante relaciones filogenéticas (Cavalli-Sforza y Edwards 1967, Fitch y Margoliash 1967, Forey *et al.* 1992, Lemey *et al*. 2009, Roch 2010). Uno de los métodos más utilizados es el método de *neighbor-joining* (NJ) en donde las secuencias de ADN se convierten en una matriz de distancias a partir de la cual se construye el árbol filogenético mediante análisis de agrupamientos (Saitou y Nei 1987).

El método es bastante rápido en términos computacionales y es muy útil cuando se deben analizar cientos de especies. Sin embargo, esta metodología ha sido muy criticada por el modo en que realiza la medición de las distancias entre los taxones y su significado filogenético (Farris 1985, Forey *et al.* 1992, Schuh y Brower 2009).

CAPÍTULO 8
ESTIMACIÓN DE LA HISTORIA EVOLUTIVA: MÉTODOS PROBABILÍSTICOS

A diferencia de la parsimonia en la que no existe un modelo de evolución para los caracteres (tema controvertido, pues la parsimonia puede ser considerada un modelo en sí mismo; Kelchner y Thomas 2007), la estimación de la filogenia mediante métodos probabilísticos utiliza un modelo de evolución de los caracteres (Whelan *et al.* 2001, Yang y Rannala 2012). En la mayoría de los casos los caracteres corresponden a datos moleculares, incluyendo genomas completos (Rannala y Yang 2008), aunque también se han utilizado datos morfológicos (Lewis 2001, Wright y Hillis 2014, Smith 2019). Para este último caso, Lewis (2001) desarrolló el modelo Mkv, una extensión del modelo Jukes-Cantor (Jukes y Cantor 1969) aplicado a estados morfológicos discretos. Sin embargo, ha sido criticado debido a que el supuesto de un mecanismo común de evolución a todos los caracteres no aplica a la morfología (Goloboff *et al.* 2018).

En los últimos años los métodos probabilísticos han generado un enorme cuerpo teórico y empírico (Edwards 2009). Uno de los métodos más conocidos es el de máxima verosimilitud (*maximum likelihood*), introducido en estudios filogenéticos por Edwards y Cavalli-Sforza (1964) y Neyman (1971) y desarrollado por Felsenstein (1981). En la década de 1990 se introdujo la inferencia bayesiana aplicada al análisis filogenético (Rannala y Yang 1996, Mau y Newton 1997), revolucionando el estudio de las filogenias aplicadas a caracteres moleculares (Huelsenbeck *et al.* 2001). En este capítulo brindaremos una breve introducción a estos métodos. Para un desarrollo más profundo de máxima verosimilitud e inferencia bayesiana aplicados a filogenias, se recomienda la lectura de Felsenstein (2004) y Chen *et al.* (2014), respectivamente.

MÁXIMA VEROSIMILITUD

Concepto de máxima verosimilitud

El objetivo de la máxima verosimilitud (MV) es intentar establecer el modelo (hipótesis estadística) que mejor describa el proceso que generó un cierto conjunto de datos. El método fue desarrollado y popularizado por Fisher entre 1912 y 1922, pero ya había sido utilizado por Gauss, Laplace, Thiele y Edgeworth (Aldrich 1997). La verosimilitud (L) se define como la probabilidad de un conjunto de datos, dado un parámetro (o conjunto de parámetros):

$$L = P(\text{datos} \mid \text{parámetros})$$

La idea consiste en asumir distintos valores del parámetro (por ejemplo, la media) y calcular la probabilidad de obtener la muestra bajo cada uno de estos valores. Luego, se elige aquel valor del parámetro que maximiza la probabilidad de obtener la muestra. Por lo tanto, el principio de MV plantea elegir aquel parámetro que hace que los datos sean más probables.

Supongamos que se tiene el siguiente conjunto de datos:

$$x = (2, 4, 5, 3, 2, 3, 3, 6, 5, 4)$$

La probabilidad de observar esta muestra es el producto de las probabilidades de cada evento:

$$P(\text{datos}) = P(2) \times P(4) \times P(5) \times P(3) \times P(2) \times P(3) \times P(3) \times P(6) \times P(5) \times P(4)$$
$$P(\text{datos}) = \prod P(x_i)$$

Π es el símbolo de productoria. El producto de estas probabilidades sólo es válido si se asume que cada evento es independiente de cualquier otro. En la práctica se suele trabajar con el logaritmo de este producto dado que es más fácil de analizar, conocido como como log-verosimilitud:

$$\log L = \log[P(\text{datos} \mid \text{parámetros})]$$

Nos podemos hacer la siguiente pregunta: ¿Cuál es el valor de media poblacional que maximiza la probabilidad de obtener los datos observados usando como modelo la distribución normal? Para responderla, debemos tener en cuenta que la probabilidad de cada evento variará según el valor que adoptemos de media poblacional y desvío estándar (que desconocemos). Por ejemplo, si asumimos que la población que dio origen a los datos tiene una distribución normal con una media poblacional $\mu = 2$ y desvío estándar poblacional $\sigma = 1$, entonces la probabilidad de estos datos será igual a $1{,}2\times10^{-12}$; si consideramos $\mu = 2{,}3$, la probabilidad de los datos asciende a $4{,}6\times10^{-12}$; y así sucesivamente (estos valores fueron calculados mediante el uso de software, manteniendo constante el desvío estándar). Por lo tanto, es necesario calcular la probabilidad de los datos tomando múltiples valores de media poblacional (Fig. 8.1A).

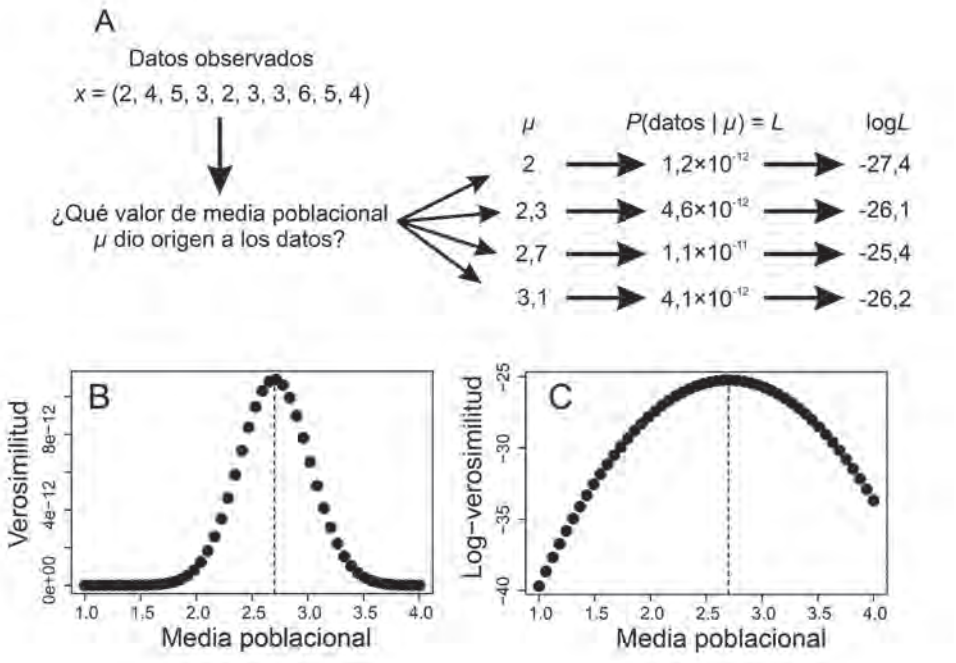


Fig. 8.1. Visualización del concepto de MV. La línea punteada indica qué valor de media poblacional maximiza la probabilidad de obtener el conjunto de datos observados, asumiendo un desvío estándar constante igual a 1. (A) Datos observados, múltiples valores de media poblacional y sus verosimilitudes; (B) gráfico de valores de media poblacional *vs.* la verosimilitud; (C) gráfico de valores de media poblacional *vs.* el logaritmo de la verosimilitud.

En la Figura 8.1B se observa que el valor μ = 2,7 es el que maximiza la probabilidad de obtener el conjunto de datos observado (de forma no casual coincide con el promedio de los datos) y se lo conoce como valor de MV. En la práctica la mayoría de las veces se obtiene este valor analíticamente, derivando la función log*L,* igualando a cero la ecuación (lo cual generalmente garantiza que hay un máximo) y despejando el parámetro de interés. Además de ser más fácil de resolver analíticamente, log*L* modifica la escala de forma tal que facilita la interpretación de los valores (Fig. 8.1C).

Máxima verosimilitud y filogenia

En un marco filogenético, los datos observados son secuencias génicas de individuos que representan taxones. Los parámetros estadísticos corresponden a la topología del árbol, las longitudes de las ramas y los parámetros del modelo de evolución de secuencias (Huelsenbeck y Crandall 1997). Asignando valores a estos parámetros es posible calcular la probabilidad de los datos observados bajo estos parámetros (Fig. 8.2):

$$L = P(\text{secuencias alineadas} \mid \text{árbol, modelo evolutivo})$$

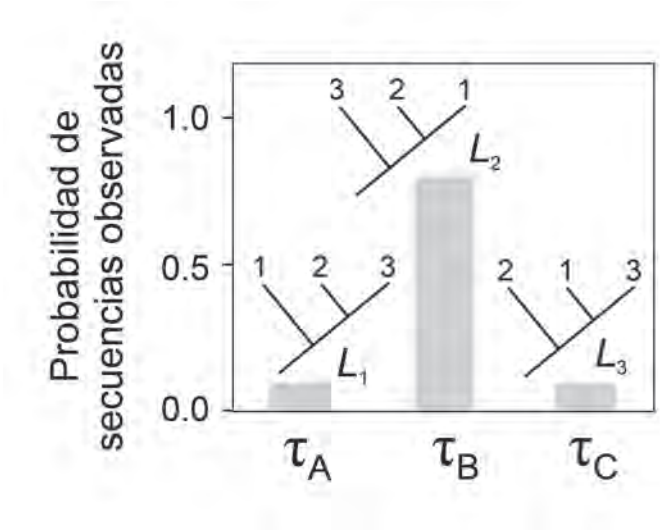


Fig. 8.2. Concepto de MV en un contexto filogenético. Se muestran tres topologías posibles (τ_A a τ_C) para tres secuencias (1 a 3) y un mismo modelo hipotético de evolución de secuencias. Para cada topología se calcula la probabilidad de obtener las secuencias observadas (verosimilitudes, L_1 a L_3). El árbol elegido es el que maximiza la probabilidad de obtener las secuencias observadas (MV), en este caso, τ_B.

Idealmente, desearíamos encontrar la probabilidad de un árbol, dado un conjunto de secuencias *P*(árbol | secuencias) (ver en este capítulo *Análisis filogenético bayesiano*). Esto requeriría encontrar todos los árboles posibles, lo cual suele ser imposible en la práctica. En cambio, el concepto de MV implica calcular la probabilidad de los datos, dado un árbol filogenético, y elegir aquel árbol que hace que los datos sean más probables.

Hay que recordar que, como resultado de utilizar el método de MV, la probabilidad de mutación en un sitio de la secuencia se asume independiente de la probabilidad de mutación en cualquier otro sitio de la misma secuencia. Este supuesto permite calcular la verosimilitud de cada sitio independientemente del resto. El producto de todos estos valores es igual a la verosimilitud de la secuencia completa. Por lo tanto, el mayor problema es encontrar el árbol óptimo para un único sitio de la secuencia de ADN.

Ejemplo más simple: máxima verosimilitud para un par de nucleótidos

Supongamos el siguiente árbol de tres secuencias, S_0 a S_2 (Fig. 8.3), donde *d* es el número de sustituciones por sitio o distancia genética (longitudes de las ramas). Considere el árbol τ con sus longitudes de las ramas (es decir, el número de sustituciones del sitio) y el modelo de evolución del sitio *M* con sus parámetros (por ejemplo, la relación transición/transversión). El objetivo es calcular la probabilidad de obtener las bases observadas. Asumiremos que el ancestro S_0 evolucionó a lo largo de las ramas del árbol τ con longitudes *d*.

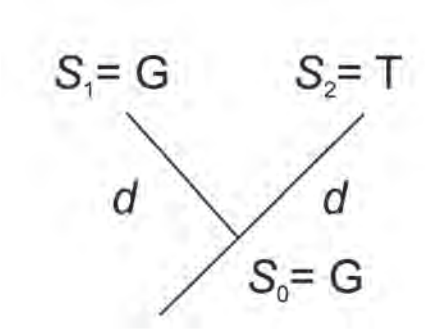


Fig. 8.3. Ejemplo hipotético de un árbol de dos secuencias (S_1 y S_2) analizando un sitio nucleotídico; asumiendo que el ancestro (S_0) en ese sitio presentaba una G. G: guanina, T: timina, *d*: longitud de la rama.

La verosimilitud de este árbol será igual al producto entre la frecuencia de la base del ancestro (f_{S_0}) y las probabilidades de sustitución en cada rama. En este caso, la verosimilitud *L* será igual a la frecuencia de G (asumida como presente en el ancestro; f_G) por las probabilidades de observar una G y una T, asumiendo que el ancestro poseía una G, P_{GG} y P_{GT}. Estas dos probabilidades dependen de la cantidad de cambios acumulados en cada rama (o longitud de la rama) *d*, por lo que debe recordarse que son funciones de *d*.

$$L = f_{S_0} \times P_{S_0S_1} \times P_{S_0S_2}$$
$$L = f_G \times P_{GG} \times P_{GT}$$

¿Cómo podemos calcular las probabilidades de sustitución? Estas dependerán del modelo evolutivo tomado como supuesto (proceso) que dio origen a los nucleótidos observados. En este ejemplo asumimos que el ancestro es conocido. En la práctica, sin embargo, las secuencias ancestrales son siempre hipotéticas. Para "resolver" esta estimación ancestral se calculan las verosimilitudes para todos los estados ancestrales posibles y se suman, método denominado MV promedio. Este es un punto importante de diferencia con respecto a la parsimonia. En esta última, si dos secuencias comparten el mismo nucleótido se asume que estaba presente en el ancestro más reciente. En el enfoque de MV este nucleótido es compartido con el ancestro con una cierta probabilidad, que se reduce si las secuencias están remotamente emparentadas.

Por ejemplo, la verosimilitud de la base A (L_A) será igual a la frecuencia de A (asumida como presente en el ancestro; f_A) por las probabilidades de observar una G y una T, asumiendo que A estaba en el ancestro, P_{AG} y P_{AT}. Este procedimiento se repite asumiendo que el resto de las bases nitrogenadas (C, G y T) estaban presentes en el ancestro, y luego se suma el total. Dicho de otra forma, la probabilidad de cada sitio se calcula sumando las verosimilitudes de todos los estados ancestrales posibles.

$$L_A = f_A \times P_{AG} \times P_{AT}$$
$$L_C = f_C \times P_{CG} \times P_{CT}$$
$$L_G = f_G \times P_{GG} \times P_{GT}$$
$$L_T = f_T \times P_{TG} \times P_{TT}$$
$$L = L_A + L_C + L_G + L_T$$

A continuación, se presenta un ejemplo sencillo para dos secuencias de nucleótidos. Éste se basa en un árbol con tres secuencias (S_0 a S_2) y por lo tanto hay un único par de ramas que las conecta (Fig. 8.4). Asumiremos que las secuencias evolucionan de acuerdo al modelo de Jukes y Cantor (JC69), donde cada sitio evoluciona independientemente de los demás y con la misma tasa y probabilidad de sustitución (Jukes y Cantor 1969).

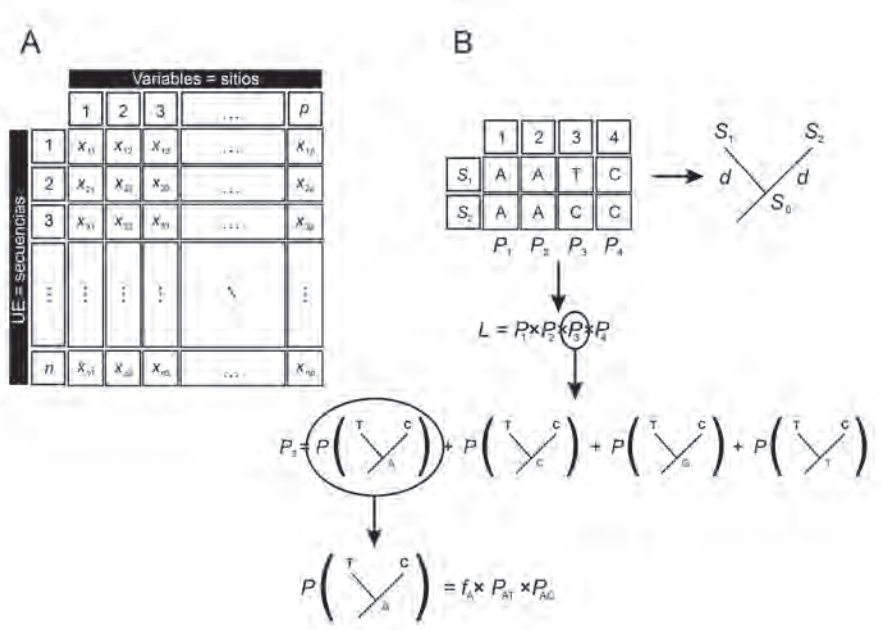


Fig. 8.4. (A) MBD aplicada a secuencias de ADN; (B) ejemplo hipotético de árbol para dos secuencias de nucleótidos (S_1 y S_2) y cuatro sitios, junto con el cálculo de las probabilidades *P* y verosimilitud *L*. Las letras grises muestran las bases nitrogenadas asumidas como presentes en el nodo S_0.

Las secuencias S_1 y S_2 corresponden a los taxones terminales del árbol filogenético, mientras que S_0 representa el nodo ancestral. El alineamiento tiene longitud $p = 4$ para las dos secuencias $S_1 = (X_{11}, X_{12}, \ldots, X_{1p})$ y $S_2 = (X_{21}, X_{22}, \ldots, X_{2p})$, donde X_{ij} es el nucleótido de la secuencia *i* en el sitio *j*. Por lo tanto, el alineamiento corresponde a la MBD y *p* al número de variables o sitios (Fig. 8.4A).

La verosimilitud de este árbol es el producto de las probabilidades de observar dos A en el sitio 1 (P_1), dos A en el sitio 2 (P_2), una T y una C en el sitio 3 (P_3), y dos C en el sitio 4 (P_4):

$$L = P_1 \times P_2 \times P_3 \times P_4$$

A modo de ejemplo, la probabilidad de observar una T y una C en el sitio 3 (P_3) se calcula como la probabilidad del árbol considerando a todos los estados ancestrales posibles (Fig. 8.4B):

$$P_3 = f_A \times P_{AT} \times P_{AC} + f_C \times P_{CT} \times P_{CC} + f_G \times P_{GT} \times P_{GC} + f_T \times P_{TT} \times P_{TC}$$

Esta probabilidad corresponde al caso del ejemplo con un único sitio. Este procedimiento se repite para cada sitio (Fig. 8.4B). La probabilidad de observar el nucleótido *y*, dado que el nucleótido *x* estaba presente anteriormente, podemos expresarla como P_{xy} y depende de *d*, es decir del número de sustituciones por sitio o distancia genética (longitud de la rama). Esta probabilidad, según el modelo JC69, puede calcularse como:

$$P_{xy} = \frac{1}{4}\left(1 + 3e^{kd}\right)$$

$$P_{xx} = \frac{1}{4}\left(1 - e^{kd}\right)$$

En otras palabras, P_{xy} es la probabilidad de mutación luego de *d* sustituciones (independientemente del nucléotido en cuestión) y P_{xx} es la probabilidad de que el nucleótido no haya mutado ($x = y$). En este modelo, *k* es igual a –4/3. Para calcular *d*, el estadístico relevante es el número de pares de nucleótidos idénticos (p_0) y el número de pares diferentes (p_1), donde $p_0 + p_1 = p$. Se puede demostrar que el valor de *d* que maximiza la log-verosimilitud es:

$$d = -\frac{3}{4}\ln\left(1-\frac{4}{3}\times\frac{p_1}{p_0+p_1}\right)$$

Para el ejemplo de la Figura 8.4, $p_0 = 3$ y $p_1 = 1$. Por lo tanto,

$$d = -\frac{3}{4}\log\left(1-\frac{4}{3}\times\frac{1}{4}\right)$$
$$d = -\frac{3}{4}\log\left(1-\frac{1}{3}\right)$$
$$d = 0,30$$

Con esta información podemos calcular las probabilidades P_{xx} y P_{xy} como:

$$P_{xx} = \frac{1}{4}\left(1+3e^{-\frac{4}{3}0,3}\right) = \frac{1}{4}\left(1+3\times 0,67\right) = 0,75$$
$$P_{xy} = \frac{1}{4}\left(1-e^{-\frac{4}{3}0,3}\right) = \frac{1}{4}\left(1-0,67\right) = 0,08$$

Teniendo en cuenta que las frecuencias de cada base son $f_A = 4/8$, $f_C = 3/8$, $f_G = 0$ y $f_T = 1/8$, podemos calcular la probabilidad de observar una T y una C en el sitio 3 (P_3):

$$P_3 = f_A \times P_{AT} \times P_{AC} + f_C \times P_{CT} \times P_{CC} + f_G \times P_{GT} \times P_{GC} + f_T \times P_{TT} \times P_{TC}$$
$$P_3 = 0,25\times 0,08\times 0,08 + 0,375\times 0,08\times 0,75 + 0\times 0,08\times 0,08 + 0,125\times 0,75\times 0,08$$
$$P_3 = 0,0032 + 0,0225 + 0,0075$$
$$P_3 = 0,033$$

Este valor es irrelevante en sí mismo y sólo cobra sentido si se compara con el valor de otro árbol (si un árbol tiene mayor valor de L se considera relativamente mejor). Este ejemplo tiene solución analítica porque es el modelo más simple de evolución y hay únicamente dos secuencias (que pueden ser relacionadas por un único árbol).

Máxima verosimilitud para más de dos secuencias de ADN

Cuando los datos consisten en más de dos secuencias, en lugar de calcular la probabilidad P_{xy} de observar dos nucléotidos x e y en un sitio para dos secuencias, se calcula la probabilidad de encontrar una cierta columna o patrón de nucleótidos en la MBD. Supongamos que D_j es el conjunto de nucleótidos observados en el sitio j del alineamiento, la probabilidad de observar ese conjunto de nucleótidos para ese sitio dependerá del modelo de evolución de secuencias M y del árbol τ que relaciona las n secuencias con el número de sustituciones a lo largo de cada rama del árbol (longitudes de las ramas). En teoría, puede asignarse a cada sitio su propio modelo de evolución de secuencias y sus longitudes de las ramas. Sin embargo, esta situación se vuelve inabordable en términos computacionales, por lo que se necesitan algunas simplificaciones. Se asume que cada sitio de la secuencia evoluciona con el mismo modelo M (por ejemplo, JC69). Este supuesto también implica que todos los sitios evolucionan con la misma tasa de sustitución. Para superar esta simplificación, la tasa en un sitio dado se modifica por un factor sitio-específico $\rho_j > 0$. Así, la probabilidad de un determinado patrón nucleotídico para el sitio j es:

$$P_j = P\left(D_j \mid M, \tau, \rho_j\right)$$

Por lo tanto, la probabilidad de observar un conjunto de alineamientos (MBD) D = (D_1, …, D_p) es el producto de las probabilidades de cada sitio:

$$L = \prod P_j = \prod P\left(D_j \mid \tau, M, \rho_j\right)$$

Cálculo de probabilidades para un determinado árbol

Considere el árbol τ con sus longitudes de las ramas (número de sustituciones por sitio), el modelo de evolución de secuencias M con sus parámetros (por ejemplo, relación transición/transversión, composición de bases) y el factor sitio-específico $\rho_j = 1$ para cada sitio j. La Figura 8.5 muestra un ejemplo hipotético de un árbol no enraizado de cuatro secuencias terminales y dos nodos (Schmidt y von Haeseler 2009). El objetivo es calcular la probabilidad de observar uno de los 4^n posibles patrones de alineamiento de las n secuencias. La Figura 8.6 muestra la MBD asumiendo que el ancestro S_5 evolucionó a lo largo de las ramas del árbol τ con longitudes d_1, d_2, d_3, d_4 y d_5 (Fig. 8.6).

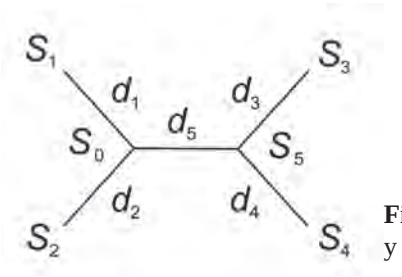


Fig. 8.5. Árbol no enraizado de cuatro secuencias (S_1 a S_4), dos nodos (S_0 y S_5) y cinco longitudes de las ramas (d_1 a d_5).

Para calcular la probabilidad en un sitio específico j, primero es necesario conocer los estados ancestrales de S_0 y S_5. La probabilidad de los datos, dados los estados ancestrales, es la probabilidad de cada transición ancestro-descendiente como función de la longitud de la rama (d) que las conecta:

$$P_j = P_{S_0S_1}\left(d_1\right) \times P_{S_0S_2}\left(d_2\right) \times P_{S_5S_3}\left(d_3\right) \times P_{S_5S_4}\left(d_4\right) \times P_{S_5S_0}\left(d_5\right)$$

Como se dijo anteriormente, las secuencias ancestrales no se conocen, por lo que se aplica el método de MV promedio.

Método de "poda" de Felsenstein

En la práctica, para calcular la verosimilitud de un árbol para un sitio dado en una determinada secuencia se utiliza el método de "poda" de Felsenstein (1981), cuyos pasos son los siguientes:

1. Comenzar a trabajar desde las terminales hacia la raíz.
2. Calcular en cada nodo la probabilidad de los datos, dado el árbol que se está calculando.
3. Una vez en la raíz, se obtiene P(datos | árbol).

Este método sólo da la P(datos | árbol) para un árbol con longitudes de las ramas especificadas. La búsqueda de todas las topologías posibles es aún un problema sin resolver.

Para las terminales, los valores de MV para cada base nitrogenada corresponden a los valores observados (1 si está presente y 0 si está ausente; es decir que no representan probabilidades). Para los nodos, el cálculo de las probabilidades se hace utilizando el método de MV promedio, como se muestra a continuación.

Considere un árbol de cuatro secuencias (S_1 a S_4) y un patrón de alineamiento para el sitio j = 4, D_j = (C, G, C, C), con todas las longitudes de las ramas d = 0,1 (Fig. 8.6A). Este sitio se convierte a una matriz de ceros y unos (Fig. 8.6B).

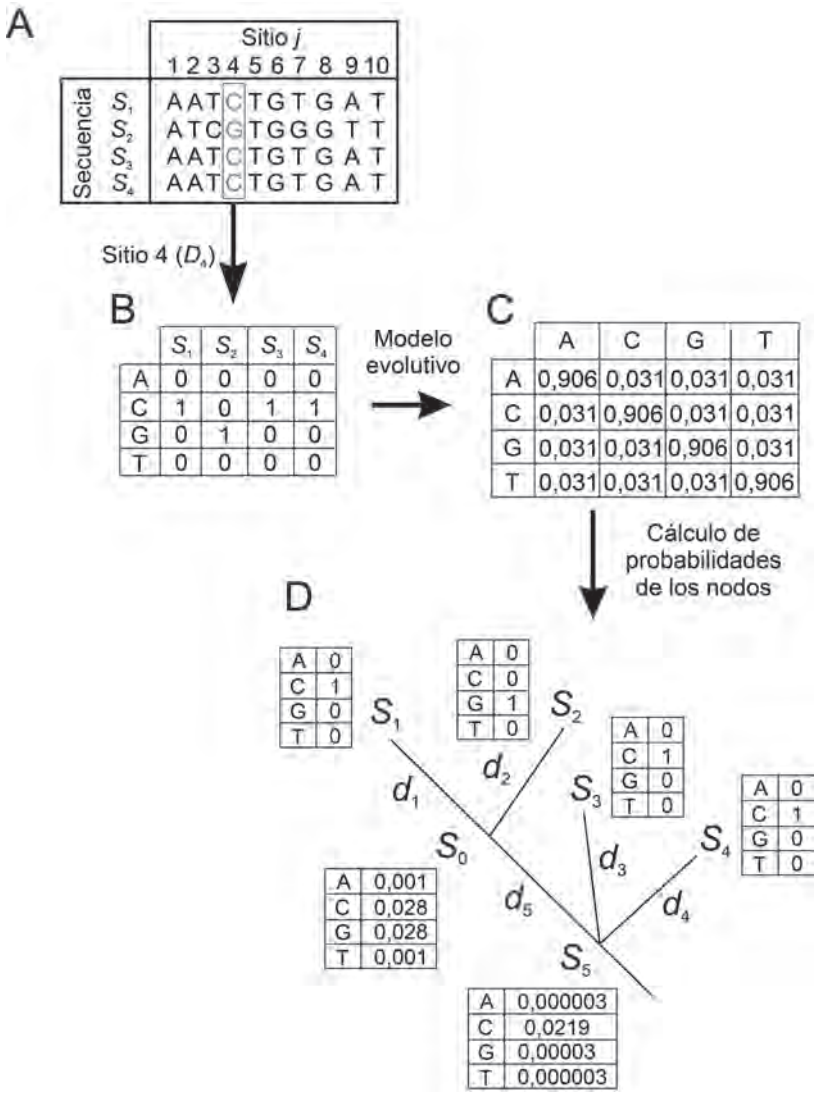


Fig. 8.6. Método de "poda" de Felsenstein. (A) Alineamiento de cuatro secuencias; (B) matriz codificada con ceros y unos del sitio 4 para las cuatro secuencias; (C) matriz de transición que muestra las probabilidades de sustitución entre bases, de acuerdo a un modelo de evolución de secuencias; (D) filogenia de las cuatro secuencias con las probabilidades (estimadas por MV) para cada uno de los cuatro tipos de nucleótidos. Modificada de Schmidt y von Haeseler (2009).

A continuación, se construye una matriz de transición que muestra las probabilidades de sustitución entre bases según un modelo de evolución de secuencias (Fig. 8.6C). En este caso, la probabilidad de no observar ninguna mutación al cabo de $d = 0{,}1$ asumiendo el modelo JC69 es:

$$P_{xx} = \frac{1}{4}\left(1 + 3e^{-\frac{4}{3}0{,}1}\right)$$

$$P_{xx} = \frac{1}{4}\left(1 + 2{,}62\right)$$

$$P_{xx} = 0{,}906$$

Mientras que la probabilidad de observar una mutación es:

$$P_{xy} = \frac{1}{4}\left(1 - e^{-\frac{4}{3}0,1}\right)$$
$$P_{xy} = \frac{1}{4}(1 - 0,87)$$
$$P_{xy} = 0,031$$

Por último, el análisis hace foco sobre los estados ancestrales de los nodos. A diferencia de la parsimonia donde se asume un estado definitivo, en MV se calcula la probabilidad de obtener cada uno de los estados posibles (A, C, G, y T). Por ejemplo, para obtener la probabilidad de que C sea la base en el sitio 4 del nodo S_0, se calcula el producto de las probabilidades de sustitución de C a C para la secuencia 1 (teniendo en cuenta la longitud de la rama d_1) y de sustitución de C a G para la secuencia 2 (teniendo en cuenta la longitud de la rama d_2), ya que el resto de los productos es igual a 0 (por ejemplo, la probabilidad de sustitución de C a A para la secuencia 1 es igual a 0 porque la base observada en la secuencia 1 es C).

$$L_0(C) = P_{CC}(d_1)P_{CG}(d_2)$$
$$L_0(C) = 0,906 \times 0,031$$
$$L_0(C) = 0,028$$

Para obtener la probabilidad de T en el nodo S_5, calculamos la probabilidad de que T haya mutado a C en las secuencias S_3 y S_4 (con longitud de ramas d_3 y d_4, respectivamente), por la probabilidad de que T haya mutado a una A, C, G o T en el nodo S_0 (P_{TA}, P_{TC}, P_{TG} y P_{TT} con distancia d_5). Esto se debe, nuevamente, a que el estado ancestral es desconocido. A su vez, cada una de estas probabilidades se multiplican por la probabilidad de que cada nucleótido haya mutado a C para la secuencia S_1 (con longitud de rama d_1) y a G para la secuencia S_2 (con longitud de rama d_2), como se muestra en el ejemplo de la ecuación anterior.

$$L_5(T) = P_{TC}(d_3)P_{TC}(d_4)\left[P_{TA}(d_5) \times 0,001 + P_{TC}(d_5) \times 0,028 + P_{TG}(d_5) \times 0,028 + P_{TT}(d_5) \times 0,001\right]$$
$$L_5(T) = 0,031 \times 0,031 \times \left[0,031 \times 0,001 + 0,031 \times 0,028 + 0,031 \times 0,028 + 0,906 \times 0,001\right]$$
$$L_5(T) = 0,031 \times 0,031 \times 0,003$$
$$L_5(T) = 0,000003$$

Encontrar el árbol de máxima verosimilitud

Los cálculos anteriores muestran cómo calcular la probabilidad de un determinado alineamiento, si todos los parámetros del árbol se conocieran. En la práctica, las longitudes de las ramas se desconocen y se calculan numéricamente mediante maximización de la función de MV. Sin embargo, encontrar las longitudes de las ramas óptimas no es el mayor problema, sino encontrar el árbol entre todas las posibles topologías que maximice la función de MV. Como se vio en el Capítulo 7, no hay un método que garantice encontrar el mejor árbol entre todas las topologías posibles.

Cuando se estima el árbol de MV se necesitan calcular todos los parámetros del modelo y longitud de ramas para cada árbol, y luego se selecciona aquel que arroje el mayor valor de verosimilitud (o sea, el que maximiza la probabilidad de obtener los datos observados). Debido a la cantidad innumerable de topologías que puede haber, tampoco es posible estimar todos los parámetros para cada árbol. Por lo tanto, se han sugerido varios métodos heurísticos para encontrar árboles relativamente razonables, incluyendo *stepwise addition* (Felsenstein 1993), *star decomposition* (Adachi y Hasegawa 1996) y *neighbor-joining* (Saitou y Nei 1987).

Métodos de permutación de ramas

Debido a la imposibilidad de encontrar todas las topologías de árboles posibles para un número de taxones mayor a 30, se utiliza la permutación de ramas (búsqueda heurística) con el fin de encontrar árboles con mayor verosimilitud. Al igual que en la parsimonia, se genera un número de árboles a partir de un árbol inicial (conjunto de subárboles) utilizando un método de permutación. Para cada árbol resultante se calcula la verosimilitud. Se selecciona el árbol con MV y se repite el procedimiento (en parsimonia se selecciona el árbol con menor número de pasos). Las permutaciones finalizan cuando no se encuentra un árbol con mayor verosimilitud. Este árbol se dice que es el árbol localmente óptimo. La probabilidad de encontrar el árbol globalmente óptimo entre todos los árboles posibles depende de los datos y la cantidad de subárboles iniciales.

Los métodos de permutación son los mismos utilizados en parsimonia: *nearest neighbor interchange* (NNI), *subtree pruning and regrafting* (SPR) y *tree bisection and reconnection* (TBR). Dependiendo de la operación, la cantidad de subárboles crece de forma lineal (NNI), cuadrática (SPR) o cúbica (TBR) con el número de taxones en el árbol completo (ver Fig. 7.17).

Una de las primeras búsquedas heurísticas de árboles por MV fue el método de adición por pasos (*stepwise addition*; Fig. 8.7). El método comienza por un árbol no enraizado para tres taxones seleccionados aleatoriamente del total de *n* taxones (Fig. 8.7A). Se construye el árbol de MV (Fig. 8.7B). A continuación, se selecciona otro taxón de forma aleatoria de los restantes *n* – 3 taxones. Se inserta el taxón en cada una de las ramas del árbol y se selecciona el árbol con mayor verosimilitud. La rama donde se insertó el taxón que genera el mayor valor de verosimilitud se denomina "rama de inserción" (Fig. 8.7C). Se repite este procedimiento hasta que se hayan incluido todos los taxones. Luego de *n* – 3 pasos se obtiene el árbol con MV que es, al menos, localmente óptimo. Dado que se ha utilizado una pequeña proporción de posibles topologías, es posible que otra inserción de taxones genere un mayor valor de verosimilitud.

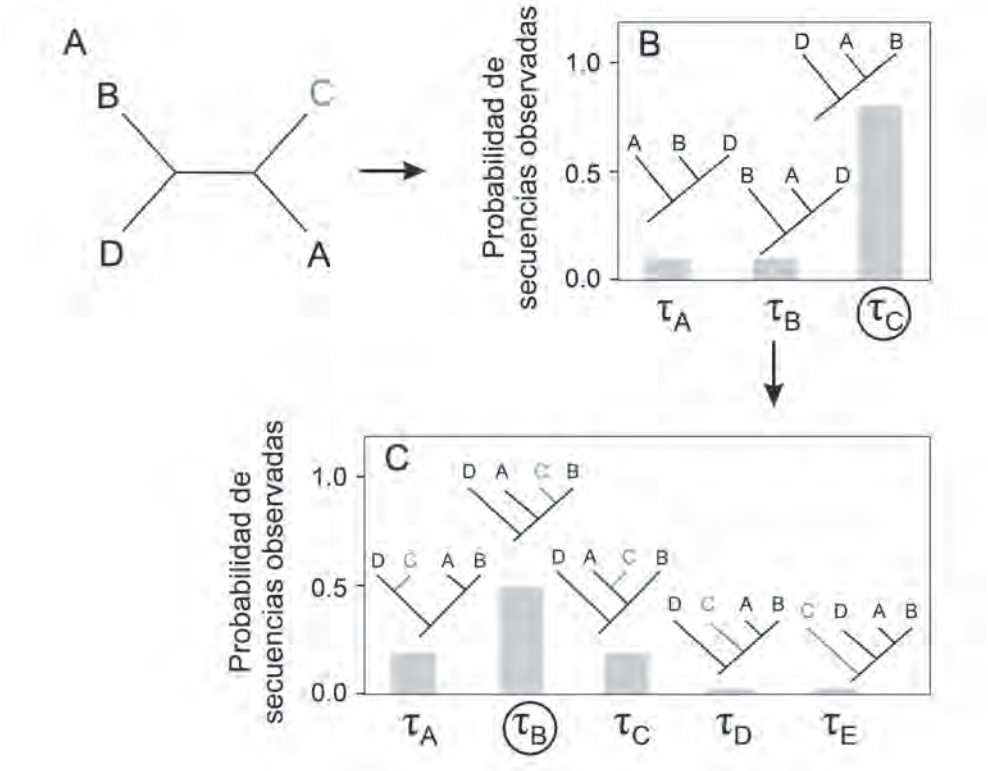


Fig. 8.7. Método de adición por pasos para obtener el árbol de MV. (A) Se seleccionan *n* – 3 taxones aleatoriamente (letras negras); (B) se identifica aquel árbol de MV (círculo); (C) se añade un nuevo taxón (letra gris) y se vuelve a identificar el árbol de MV. El procedimiento se repite hasta que todos los taxones hayan sido incluidos en el árbol.

ANÁLISIS FILOGENÉTICO BAYESIANO

Lógica bayesiana

Supongamos que tenemos una urna con una gran cantidad de pelotas, algunas negras y otras blancas. Conociendo la proporción de pelotas negras y blancas ¿cuál es la probabilidad de obtener, por ejemplo, cinco pelotas negras y cinco pelotas blancas si se extraen 10 pelotas? Este es un problema de probabilidad directo. Bayes resolvió una situación inversa a dicho problema (Box y Tiao 2011). El planteo inverso es: dada una muestra de pelotas negras y blancas ¿cuál es la proporción de pelotas negras y blancas en la urna? Este es el tipo de pregunta que se quiere responder en la inferencia bayesiana.

Nuestro pensamiento y toma de decisiones en la vida cotidiana tiene un fuerte componente bayesiano. Por ejemplo, supongamos que tenemos tos un día de verano. ¿Cuál es la probabilidad de que sea producto de una gripe? Dado lo que se sabe de la prevalencia de gripe en verano, podemos suponer que es sólo tos. Considere la misma situación pero en invierno. Con esta información adicional, probablemente supondríamos que la probabilidad de tener gripe sería mucho mayor. Este uso del sentido común y el conocimiento previo es algo que hacemos inconscientemente. La inferencia bayesiana es una forma de utilizar esta información para realizar predicciones más precisas.

La probabilidad de tener gripe dado que presentamos tos es una probabilidad condicional expresada como *P*(gripe | tos) pero que se ve alterada con la información previa de la estación del año (denominada en la jerga bayesiana probabilidad *a priori*). Si consideramos la baja prevalencia de la enfermedad en el verano, *P*(gripe | tos), será baja, mientras que si estamos en invierno, *P*(gripe | tos) será alta. Estas probabilidades condicionales se denominan probabilidades *a posteriori*, porque son el resultado final del cálculo luego de incorporar la información previa.

Si no supiéramos nada sobre los porcentajes de personas que tienen gripe en nuestra área, es decir asumiendo que el 50% de las personas tienen gripe durante todo el año, difícilmente podamos concluir si tenemos gripe o no (en cuyo caso se denomina probabilidad *a priori* vaga, plana o no informativa). De esta forma, nuestro conocimiento se actualiza continuamente en la vida cotidiana como resultado de incorporar información nueva sobre distintos fenómenos. La idea fundamental del teorema de Bayes es la modificación de nuestras propias creencias una vez que observamos los datos (evidencia). De aquí surge la principal crítica al enfoque bayesiano: basar el análisis en creencias subjetivas del investigador. Sin embargo, cuando realmente hay creencias fuertes y consensuadas sobre determinados parámetros ¿por qué no hacerlas explícitas a través del análisis bayesiano?

La pregunta en términos probabilísticos es, dado que la persona presenta tos ¿cuál es la probabilidad de tener gripe, *P*(gripe)? Resulta que es imposible conocer este valor sin especificar nuestras creencias previas acerca del valor de *P*(gripe). Esto se hace asumiendo alguna distribución de probabilidad con los posibles valores de *P*(gripe). Como se dijo anteriormente, si no hay información previa podemos asignarles el mismo valor a *P*(gripe) en invierno y en verano.

Bayes encontró que la probabilidad de que un fenómeno observado (gripe en nuestro ejemplo) se deba a una causa en particular (tos en nuestro caso), puede obtenerse con la siguiente fórmula:

$$P(\text{gripe} \mid \text{tos}) = \frac{P(\text{gripe})\,P(\text{tos} \mid \text{gripe})}{P(\text{tos})}$$

Esto se conoce como teorema o regla de Bayes (1763). La probabilidad *P*(gripe | tos) se conoce como probabilidad *a posteriori* porque especifica la probabilidad de tener gripe, luego de que la probabilidad *a priori* se ha actualizado con los datos disponibles (en este caso, el síntoma de la tos en invierno o en verano). En la Figura 8.8. se muestra una representación de esta situación utilizando el teorema de Bayes.

¿Cómo traducimos esta situación al análisis bayesiano? Para la situación 1 (estación verano; Fig. 8.8A) supongamos que la prevalencia de gripe es del 5% (5 personas de 100), es decir que *P*(gripe) = 0,05. De las personas que no tienen gripe (95), 40 tienen tos y 55 no la tienen. De las personas con gripe (5), una no tiene tos y cuatro sí la tienen, es decir que *P*(tos | gripe) = 4/5 = 0,80. Finalmente, la probabilidad de presentar tos, *P*(tos) = (4 + 40)/100 = 0,44.

Hasta aquí no es necesario realizar ningún cálculo, sino simplemente observar que hay cuatro personas con gripe y tos *vs.* 40 personas sin gripe y con tos, por lo que es una apuesta "segura" que la persona no tiene gripe, porque en verano es común la tos sin gripe. Este tipo de razonamiento es el que hacemos inconscientemente.

Reemplazando valores:

$$P(\text{gripe} \mid \text{tos}) = \frac{0,05 \times 0,80}{0,44}$$
$$P(\text{gripe} \mid \text{tos}) = 0,09$$

El valor de 0,09 es la probabilidad de tener gripe, dado que se tiene tos en verano. De la misma forma podemos calcular la probabilidad de no tener gripe teniendo tos aplicando la propiedad del complemento:

$$P(\text{no gripe} \mid \text{tos}) = 1 - 0,09$$
$$P(\text{no gripe} \mid \text{tos}) = 0,90$$

En el invierno la situación cambia (Fig. 8.8B). Supongamos que la prevalencia de gripe es del 60% (en nuestro ejemplo 60 personas), es decir $P(\text{gripe}) = 0,60$. De las personas que no tienen gripe (40), cinco tienen tos y 35 no la tienen. De las personas con gripe (60), cinco no tienen tos y 55 sí la tienen, es decir, $P(\text{tos} \mid \text{gripe}) = 55/60 = 0,92$. Como hay cinco personas sin gripe pero con tos *vs.* 55 personas con gripe y tos, la apuesta segura es que la persona tiene gripe. Ahora, $P(\text{tos}) = (5 + 55)/100 = 0,60$. Si reemplazamos estos valores en la fórmula de Bayes:

$$P(\text{gripe} \mid \text{tos}) = \frac{0,60 \times 0,92}{0,60}$$
$$P(\text{gripe} \mid \text{tos}) = 0,92$$

De esta forma, el teorema de Bayes ha capturado nuestra intuición sobre la situación. Lo que es más importante, ha incorporado nuestro conocimiento preexistente de que hay muchos más casos de gripe en invierno que en verano. Usando este conocimiento previo, el teorema actualizó nuestras creencias sobre el fenómeno.

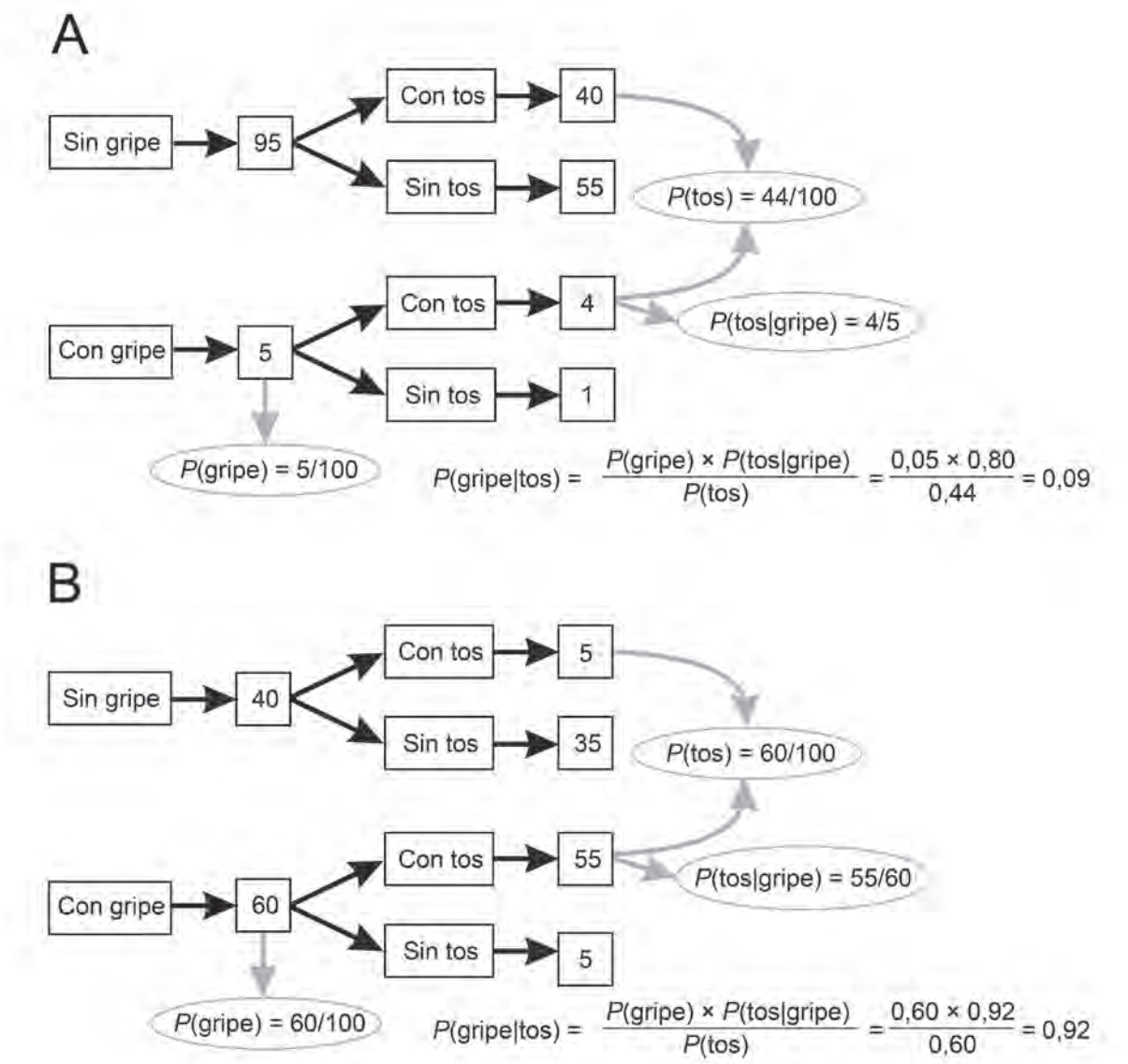


Fig. 8.8. Teorema de Bayes aplicado a un ejemplo sencillo de casos de gripe con la tos como evidencia. (A) En verano; (B) en invierno. Las líneas y elipses grises representan aquellas probabilidades necesarias para calcular el teorema.

Para generalizar, en nuestro ejemplo, gripe es el parámetro de interés, mientras que tos en verano o invierno representa el dato observado o evidencia. Reemplazando estos nuevos conceptos:

$$P\left(\text{parámetro} \mid \text{datos}\right) = \frac{P\left(\text{parámetro}\right) P\left(\text{datos} \mid \text{parámetro}\right)}{P\left(\text{datos}\right)}$$

El denominador es una constante de normalización que asegura que los valores de probabilidad varíen entre 0 y 1. Observe que en el numerador se encuentra el concepto de verosimilitud $L = P$(datos | parámetro). Finalmente, podemos conceptualizar este teorema como:

$$\text{Probabilidad } a \textit{ posteriori} = \frac{\text{Probabilidad } a \textit{ priori} \times L}{cte}$$

En resumen, la probabilidad *a posteriori* corresponde al producto entre la información previa (probabilidad *a priori*) y la información contenida en los datos observados.

Inferencia filogenética bayesiana

¿Cómo se aplica el teorema de Bayes a la filogenia? Supongamos que estamos interesados en las relaciones entre tres taxones, para los cuales hay tres topologías posibles (Fig. 8.9). Antes de comenzar el análisis, debemos especificar nuestras creencias *a priori* acerca de las relaciones entre los taxones. En ausencia de información previa, una solución simple es asignar igual probabilidad a cada árbol. Dado que hay tres posibilidades, a cada árbol le corresponde una probabilidad de 1/3. Esta probabilidad *a priori* no informativa es apropiada cuando no tenemos conocimiento previo o no queremos realizar el análisis tomando como base resultados previos.

Para actualizar la probabilidad *a priori* se necesitan datos observados típicamente en forma de secuencias de ADN y un modelo explícito de evolución de secuencias (Lemmon y Moriarty 2004, Kelchner y Thomas 2007). En principio, la regla de Bayes se utiliza para obtener la probabilidad *a posteriori*, que es el resultado del análisis. La probabilidad *a posteriori* especifica la probabilidad de un árbol, dados un modelo, la probabilidad *a priori* y los datos (Huelsenbeck *et al.* 2001, Ronquist 2004). Cuando los datos son informativos, la mayor parte de la distribución de la probabilidad *a posteriori* se concentra en un solo árbol (o en un subconjunto de árboles del total de árboles posibles). El procedimiento se resume en la Figura 8.9.

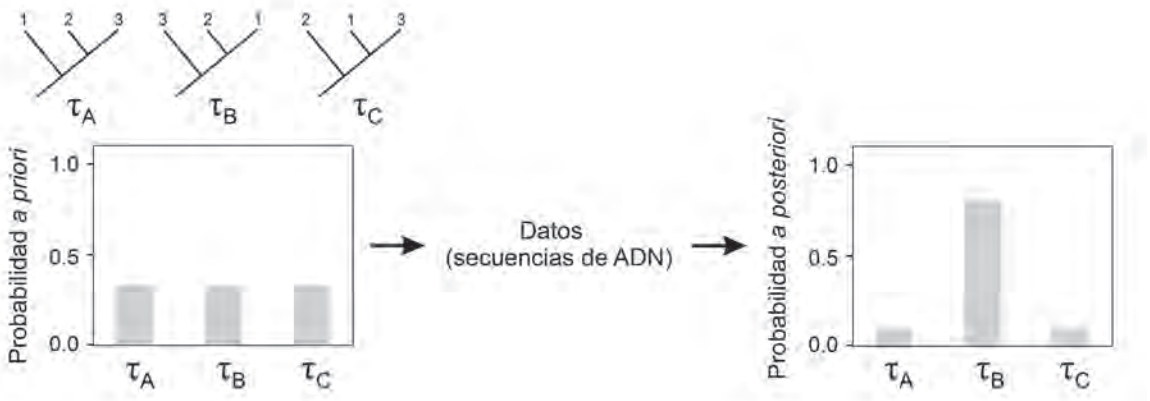


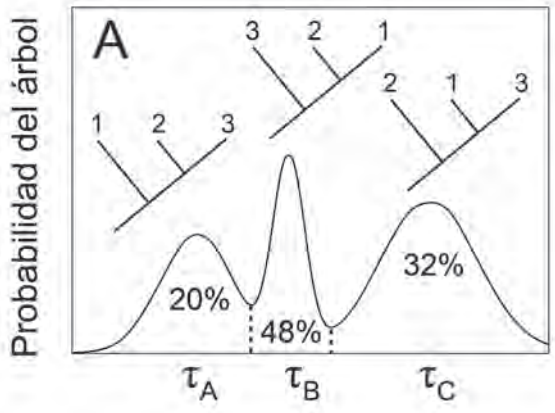
Fig. 8.9. Enfoque bayesiano aplicado al análisis filogenético. Se comienza suponiendo que todos los árboles tienen la misma probabilidad *a priori* de haber dado origen a las secuencias observadas. Una vez incorporados los datos (observaciones) en formas de secuencias de ADN, se puede identificar el árbol que tiene la mayor probabilidad de haber dado origen a las secuencias observadas.

Para realizar el análisis necesitamos una MBD de secuencias alineadas y un modelo de evolución de secuencias. Este modelo en el caso ideal, contendría un único parámetro de topología del árbol, τ, con tres valores posibles. Sin embargo, esto no es suficiente porque se necesitan también las longitudes de las ramas d que están asociadas a un modelo de sustitución. Al igual que en MV, los parámetros del árbol son la topología y el modelo de evolución de las secuencias:

$$P(\text{árbol} \mid \text{secuencias}) = \frac{P(\text{árbol})\,P(\text{secuencias} \mid \text{árbol})}{P(\text{secuencias})}$$

Dicho de otra forma, el teorema de Bayes aplicado a la filogenia calcula la probabilidad de un árbol dada una serie de secuencias de ADN y un modelo de sustitución (Huelsenbeck *et al.* 2001). El árbol con la mayor probabilidad *a posteriori* será seleccionado como la mejor estimación de la filogenia.

Aunque el ejemplo mencionado anteriormente es el más sencillo, es imposible representar el espacio de parámetros en una sola dimensión (ya que hay dos parámetros, topología y longitudes de las ramas). Sin embargo, imagine por un instante que es posible (Fig. 8.10A), entonces el eje de los parámetros presentaría tres regiones distintas correspondientes a tres topologías diferentes. Dentro de cada región, los diferentes puntos en el eje representan diferentes longitudes de las ramas. Presumiblemente mostraría tres picos, cada uno correspondiente a una combinación óptima de topología y longitud de las ramas.



B

	τ_A	τ_B	τ_C	
d_1 = 0,1, 0,2, 0,3, ...	0,10	0,07	0,12	0,29
d_2 = 0,1, 0,2, 0,3, ...	0,05	0,22	0,06	0,33
d_3 = 0,1, 0,2, 0,3, ...	0,05	0,19	0,14	0,38
	0,20	0,48	0,32	

Fig. 8.10. (A) Probabilidad *a posteriori* de cada topología de árbol; (B) tabla de topologías (τ) *vs.* longitudes de las ramas de cada taxón (d). Las celdas en gris representan las probabilidades marginales (acumuladas). Por ejemplo, para una topología dada, éstas representan la probabilidad acumulada de todas las combinaciones de longitudes de las ramas posibles. Es importante resaltar que al ser este un ejemplo hipotético, el modelo de evolución de secuencias no es explícito, pero en la práctica se necesita para estimar las longitudes de las ramas del árbol. Modificada de Ronquist *et al.* (2009).

Para obtener la distribución de probabilidad *a posteriori* de cada topología, calculamos el área bajo cada una de las curvas. Al hacer esto, estamos considerando todas las longitudes de las ramas posibles para cada topología, por lo que las longitudes de las ramas en este punto, no son de interés. En la jerga, el conjunto de estas distribuciones se denomina distribución de probabilidad marginal de las topologías. Imaginemos que representamos el espacio de parámetros en una tabla de doble entrada (Fig. 8.10B). Las columnas podrían representar las tres topologías y las filas distintas longitudes de las ramas. Dado que las longitudes de las ramas toman valores continuos, habría en realidad infinitas filas. A los fines prácticos, agrupamos los valores de longitud de ramas en categorías discretas. De este modo, podemos obtener las probabilidades *a posteriori* en cada una de las celdas de la tabla. Estas son probabilidades conjuntas, porque representan la probabilidad de una topología y longitud de las ramas en particular. Si sumamos todas las probabilidades conjuntas a lo largo de una fila o columna de la tabla, obtendríamos las probabilidades marginales del parámetro correspondiente. Para obtener la probabilidad marginal de las topologías se suman las celdas a lo largo de cada columna (Fig. 8.10B). Esta probabilidad marginal representa la probabilidad de una determinada topología, considerando todas las combinaciones de longitudes de las ramas posibles.

También es posible obtener las probabilidades marginales de las longitudes de las ramas sumando los valores de cada fila. Típicamente, estos valores son de poco interés, pero muestran una propiedad importante de la inferencia bayesiana: no hay una distinción clara entre diferentes tipos de parámetros (no hay parámetros más importantes que otros). Una vez que se obtiene la distribución de probabilidad *a posteriori*, se puede calcular cualquier probabilidad marginal de interés. No hay necesidad de decidir cuál es el parámetro de interés *a priori* del análisis.

Distribuciones de probabilidad *a priori*

La probabilidad *a priori* debería resumir el mejor conocimiento del investigador acerca del modelo o parámetros antes de analizar los datos (Nascimento *et al.* 2017). Estas probabilidades *a priori* son importantes debido a que representan una parte esencial del análisis bayesiano. En este sentido, debe tenerse en cuenta que las probabilidades *a priori* por defecto en muchos programas pueden no ser apropiadas y deben utilizarse con cuidado. La especificación de las probabilidades *a priori* es una responsabilidad del usuario, a pesar de no ser una tarea fácil debido al gran número de parámetros para estimar que suele haber. Por lo tanto, el análisis de robustez también debería formar parte de todo análisis bayesiano (Nascimento *et al.* 2017). Mediante la evaluación de las probabilidades *a posteriori* generadas bajo diferentes probabilidades *a priori*, el investigador puede evaluar si las probabilidades *a posteriori* son robustas (Nascimento *et al.* 2017).

Por el contrario, el uso de las probabilidades *a priori* no informativas "deja hablar a los datos" y no sesgan las conclusiones con la subjetividad inherente de las probabilidades *a priori* subjetivas. En este caso la inferencia bayesiana es similar a MV, dado que la probabilidad *a priori* es igual para cualquier valor del parámetro (Huelsenbeck *et al.* 2001).

$$\text{Probabilidad } a\ posteriori = \frac{c \times L}{k}$$

Sin embargo, el uso de probabilidades *a priori* no informativas ha sido criticado, ya que la presencia de información *a priori* genuina representa el centro del análisis bayesiano (Efron 2013). Además, muchos conjuntos de datos muestran probabilidades *a posteriori* extremadamente altas, incluso para clados aparentemente incorrectos (Yang y Rannala 2005, Rannala *et al.* 2011). Las probabilidades *a posteriori*, tanto de árboles como de clados, son sensibles a las probabilidades *a priori* de las longitudes de las ramas internas (internodo), y aquellas probabilidades *a priori* que asumen longitudes de las ramas largas conllevan altas probabilidades *a posteriori* de los árboles (Yang y Rannala 2005).

En el otro extremo, basar la inferencia sobre fuertes probabilidades *a priori* puede ser contraproducente. Si bien mejoramos nuestra estimación al hacer algunas suposiciones sobre el parámetro, el propósito de medir algo es aprender sobre él. Si suponemos que ya conocemos la respuesta podemos estar censurando los datos. Dicho de otra forma, si comenzamos el análisis con una fuerte suposición previa sobre un pequeño conjunto de árboles, nunca detectaríamos si otro árbol es más probable. Nuestra probabilidad *a priori* asignaría una probabilidad cero a los otros árboles.

Mark Twain (1835-1910) expuso el peligro de utilizar fuertes probabilidades *a priori*: "*Lo que te mete en problemas no es lo que no sabes. Es lo que sabes con certeza que simplemente no es así*". Sin embargo, hay una forma de evitar eliminar ciegamente ciertas posibilidades, asignando al menos una pequeña probabilidad a cada árbol. Esta es una de las razones por la cual generalmente se usa la distribución normal como distribución de probabilidad *a priori*. Esta distribución concentra la mayor parte de nuestra creencia en un pequeño rango de resultados, pero tiene colas muy largas que nunca se vuelven completamente cero, sin importar cuánto se alarguen.

Cadenas de Markov Monte Carlo

Una vez seleccionados los datos, el modelo y las probabilidades *a priori*, el siguiente paso es obtener la distribución *a posteriori*. En la mayoría de los casos es imposible obtener esta distribución analíticamente. Aún peor, ésta no se puede estimar ni siquiera tomando muestras aleatorias de esta misma distribución. Esto se debe a que la mayor parte de la distribución *a posteriori* se encuentra concentrada en una pequeña parte del vasto espacio de parámetros. También es evidente en un contexto filogenético, donde hay un gran número de topologías posibles, pero es una sola la que dio origen a las secuencias de ADN observadas. La solución es estimar la distribución *a posteriori* utilizando el método de cadenas de Markov Monte Carlo (MCMC) o método de Metropolis-Hastings (Yang y Rannala 1997).

El método de MCMC surge en la estadística bayesiana como un método para aproximar la distribución de probabilidad *a posteriori* de un parámetro mediante un muestreo aleatorio en el espacio probabilístico (Yang y Rannala 1997, Larget y Simon 1999, Li *et al.* 2000). La idea central es realizar pequeños cambios aleatorios en los valores de un parámetro, y aceptar ese valor si el cambio propuesto aumenta la probabilidad con respecto al valor anterior (Fig. 8.11). Si la probabilidad del parámetro en la nueva posición es menor que en la posición anterior, se acepta la nueva posición dependiendo del valor del cociente entre ambas probabilidades (Fig. 8.11). En el caso de las filogenias, la idea es proponer un determinado árbol y realizar pequeños cambios aleatorios sobre sus parámetros para obtener un nuevo árbol, y luego aceptarlo o rechazarlo según las probabilidades de cada uno. Si éste último se acepta, se vuelve a realizar un pequeño cambio generando otro árbol para compararlo con el anterior, y así sucesivamente hasta lograr la convergencia (Huelsenbeck *et al.* 2001). El conjunto de todos estos cambios forma una cadena.

El método comienza a partir de un valor arbitrario θ, luego se propone mover este valor hacia otra posición. El método para adoptar este cambio de posición puede ser simple o sofisticado, y se denomina método de propuesta (*proposal method*). El método de Metropolis-Hastings (Metropolis *et al.* 1953, Hastings 1970) por ejemplo, toma un valor aleatorio de una distribución normal con media en el valor θ y un cierto desvío estándar σ (que determinará qué tan lejos pueden tomar los valores de θ). Luego, se evalúa si esa nueva posición θ^* es un buen lugar hacia el cual desplazarse. Si la probabilidad resultante del parámetro en la nueva posición aumenta (explica mejor los datos que la posición anterior), entonces se acepta ese nuevo valor. El método calcula el cociente de las probabilidades *a posteriori* (r) entre ambos valores, pudiendo obtener dos resultados: (1) si r es mayor a 1, el nuevo valor θ^* tiene mayor probabilidad que el valor anterior, en cuyo caso siempre se acepta como nuevo punto de partida en la siguiente iteración de la cadena, y (2) si r es menor a 1, el nuevo valor θ^* tiene menor probabilidad que el valor anterior, en cuyo caso se acepta con una probabilidad proporcional a la magnitud del cociente (por ejemplo, si el valor actual tiene cuatro veces más chances que el valor propuesto, $r = 1/4$, habrá un 25% de probabilidad de moverse al valor propuesto). Esto último tiene la finalidad de evitar quedar atrapado en un máximo local, al explorar un amplio rango del espacio de probabilidades de las topologías posibles y favorecer la posibilidad de encontrar el máximo global (Fig. 8.11B). Si siempre se aceptara el valor con mayor probabilidad, se podría identificar un pico que no es el máximo global, sino uno local, de la distribución *a posteriori* (Fig. 8.11).

Observe que el valor del parámetro en la siguiente iteración depende del valor actual, pero no de los valores visitados anteriormente. Este método se dice que “no tiene memoria”, propiedad conocida como Markoviana, de aquí viene el nombre de cadena de Markov (Nascimento *et al.* 2017).

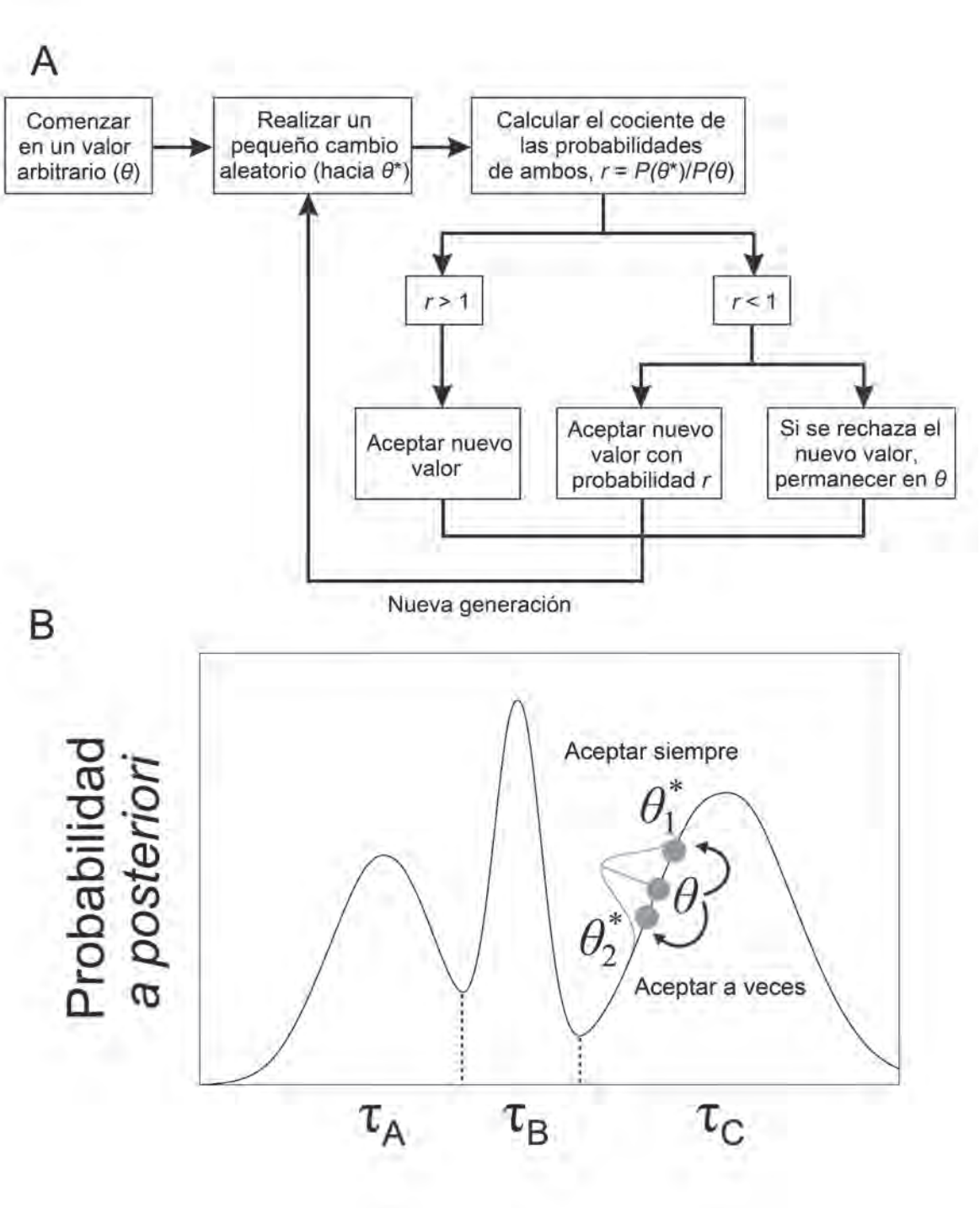


Fig. 8.11. Método de MCMC. (A) Mapa conceptual del método de MCMC con los pasos a seguir para estimar la probabilidad *a posteriori* de un parámetro; (B) gráfico que resume el método de MCMC. Los máximos en τ_A y τ_C son máximos locales, mientras que el máximo en τ_B es el máximo global. El método de Metropolis-Hastings utiliza una distribución normal (curva gris) con media θ y desvío estandar σ. Si $\theta^* > \theta$, se acepta siempre el nuevo valor; si $\theta^* < \theta$, se acepta el nuevo valor con una cierta probabilidad r.

Burn-in, mezcla y convergencia

Si la cadena comenzó a partir de un árbol aleatorio y longitudes de las ramas arbitrarias, la probabilidad de que esos parámetros sean los correctos es muy baja. A medida que la cadena se mueve hacia

regiones con mayor probabilidad, la verosimilitud típicamente aumenta muy rápido. Esta fase inicial de la corrida se denominada *burn-in*, y suele descartarse porque está fuertemente influenciada por el valor inicial (Fig. 8.12).

A medida que las generaciones (iteraciones) aumentan, la verosimilitud tiende a estabilizarse en la denominada fase estacionaria (Fig. 8.12). Esta fase es el primer signo de convergencia de la cadena. Así, el gráfico del número de generaciones *vs.* la verosimilitud, conocido como gráfico traza, es importante para monitorear el desempeño de la MCMC (Fig. 8.12). Sin embargo, es necesario confirmar la convergencia con otras herramientas diagnósticas, porque además de esta condición, es necesario que las probabilidades recorran adecuadamente la distribución *a posteriori*. La velocidad con la cual las cadenas cubren zonas de alta probabilidad de la distribución *a posteriori* se conoce como comportamiento de mezcla (*mixing*). Mientras mejor sea la mezcla, más rápido la cadena generará una muestra adecuada de probabilidad *a posteriori* (Fig. 8.13).

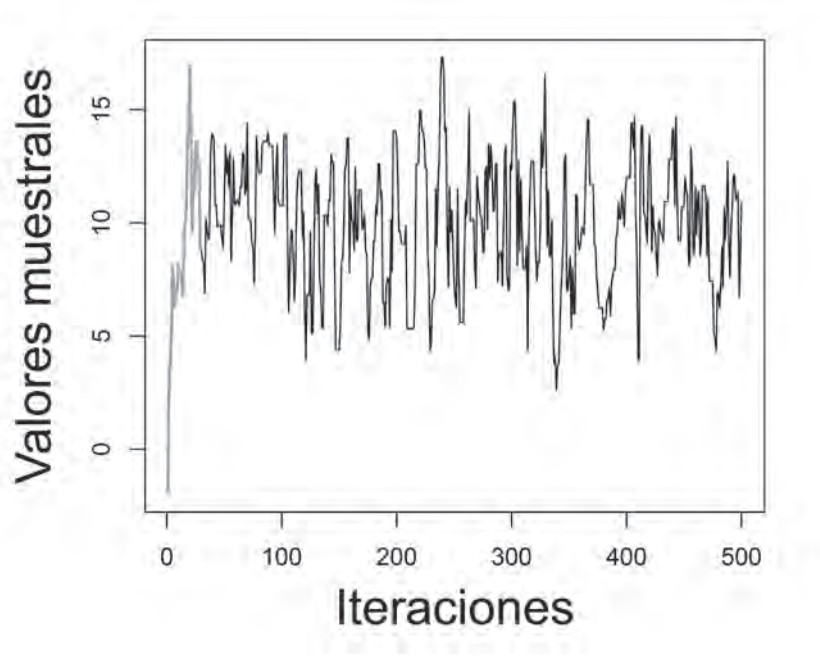


Fig. 8.12. Traza de una corrida de MCMC, mostrando las iteraciones *vs.* los valores muestrales del parámetro de interés. Se muestran el *burn-in* (sección en gris) y la fase estacionaria (sección en negro).

El comportamiento de mezcla del muestreo de Metropolis-Hastings puede regularse utilizando un parámetro de ajuste (*tuning parameter*). En este método, el parámetro de ajuste es el desvío estándar de la distribución normal. Si el desvío es muy pequeño, el valor propuesto será muy similar al anterior. La probabilidad *a posteriori* será muy similar a la actual, por lo que se tenderá a aceptar el valor propuesto. Cada valor propuesto moverá la cadena una pequeña distancia en el espacio de parámetros, por lo que tomará un largo tiempo cubrir completamente la región de interés; la mezcla es pobre y la cadena es ineficiente (Fig. 8.13A). Un desvío muy grande también tendrá una mezcla pobre y la cadena será ineficiente. Bajo estas condiciones, el valor propuesto casi siempre será muy diferente del valor actual. Si se alcanza una región de alta probabilidad *a posteriori*, es muy posible que el valor propuesto tenga una probabilidad muy baja con respecto al valor actual. Por lo tanto, la cadena permanecerá en el mismo valor un largo tiempo, resultando en una mezcla pobre (Fig. 8.13B). Así, el muestreo más eficiente consiste en utilizar parámetros de ajuste intermedios (Fig. 8.13C), con tasas de aceptación del 30-40% (Nascimento *et al.* 2017).

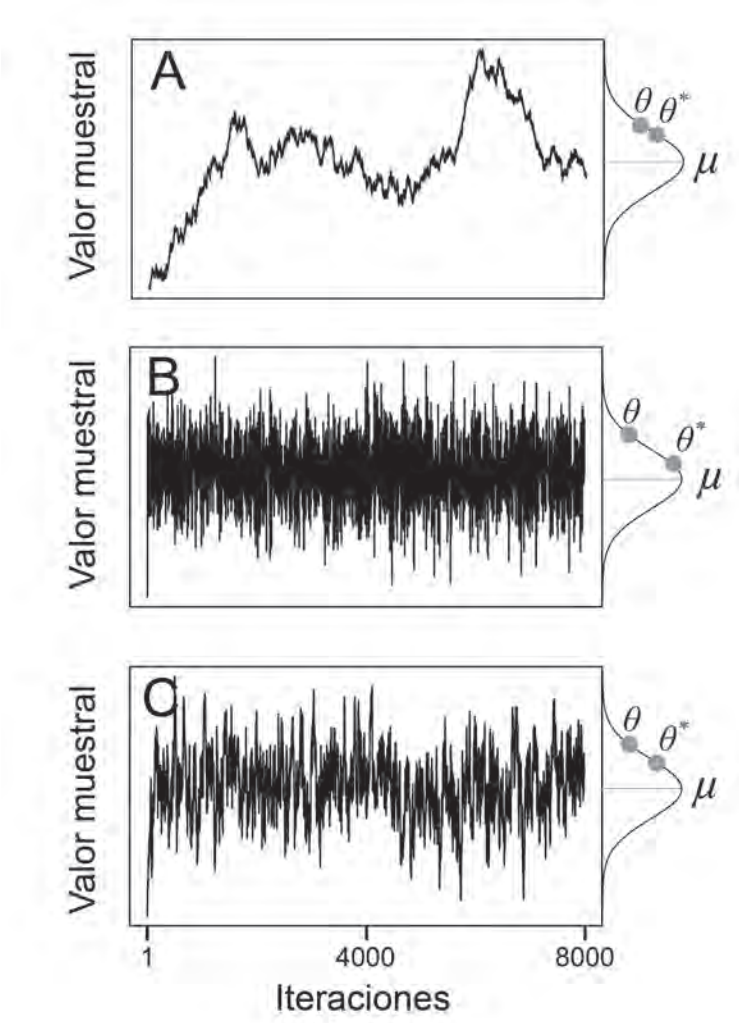


Fig. 8.13. Comportamiento de mezcla para diferentes valores de desvío estándar. Las distribuciones normales representan la distribución poblacional desconocida que se quiere describir. (A) Diferencias entre θ y θ* pequeñas (σ = 0,1) resultan en tasas de aceptación del valor propuesto altas; el espacio de probabilidades toma mucho tiempo en recorrerse, dando como resultado una mezcla pobre; (B) diferencias entre θ y θ* altas (σ = 5) resultan en tasas de aceptación del valor propuesto muy bajas; el espacio de probabilidades no es completamente recorrido, dando también como resultado una mezcla pobre; (C) diferencias entre θ y θ* moderadas (σ = 1) resultan en tasas de aceptación del valor propuesto intermedias; el espacio de probabilidades es recorrido relativamente bien, dando como resultado una buena mezcla.

En una larga corrida de MCMC la cadena debería pasar la mayor parte de las iteraciones recorriendo regiones de alta probabilidad *a posteriori*. La tasa de convergencia es la tasa con la cual una cadena que comienza desde cualquier posición inicial se mueve a altas regiones de probabilidad *a posteriori* (Nascimento *et al.* 2017). Los diagnósticos de convergencia ayudan a determinar la calidad de la muestra de la distribución *a posteriori*. Esencialmente, existen tres tipos de diagnósticos ampliamente utilizados (Ronquist *et al.* 2009): (1) examinar las series de autocorrelación, tamaños de muestra efectivos y otras medidas del comportamiento de una única cadena, (2) comparar muestras de segmentos de tiempos sucesivos de una única cadena, y (3) comparar muestras de diferentes cadenas. El último enfoque es sin dudas el más poderoso para detectar problemas de convergencia. Sin embargo, implica cierto desperdicio de tiempo computacional, ya que genera múltiples muestras *burn-in* que deben ser descartadas. Por lo tanto, el primer enfoque es ampliamente utilizado. En este caso, la fase estacionaria de la cadena se muestrea cada una cierta cantidad de iteraciones (proceso denominado *thinning*), por ejemplo cada 50 ó 100 generaciones. Esto se debe a que las cadenas van cambiando de manera muy lenta, por lo que los valores de muestras sucesivas son muy parecidos entre sí (ya que el valor actual es igual al anterior si se rechaza el valor propuesto, o es una modificación de éste si se acepta), efecto denominado autocorrelación. Por otro lado, el *thinning* permite ahorrar espacio en el disco, dado que una MCMC puede generar fácilmente millones de muestras.

En el estudio de las filogenias, la topología del árbol es típicamente el parámetro más difícil de obtener. Así, tiene sentido enfocarse en este parámetro cuando se evalúa la convergencia de las cadenas. Si se realizan varias corridas de MCMC para un mismo árbol, inicialmente cada uno muestreará distintas

zonas del espacio de parámetros. A medida que se acercan a la convergencia, todas las corridas serán similares. Por lo tanto, una forma de evaluar la convergencia es calcular la varianza entre y dentro de las diferentes cadenas. Este diagnóstico se denomina *Potential Scale Reduction Factor* (PSRF) propuesto por Gelman y Rubin (1992). Si las cadenas comienzan a partir de puntos de inicio muy dispersos, la varianza entre las cadenas será inicialmente mayor a la varianza dentro de las cadenas. A medida que las cadenas convergen, las varianzas serán cada vez más similares y el PSRF se aproximará a 1.

Autocorrelaciones fuertes indican que la cadena es poco eficiente en atravesar la distribución de probabilidad *a posteriori*. Una medida de la eficiencia es el cociente entre la varianza de una muestra independiente del mismo tamaño que la muestra *a posteriori* y la varianza de la cadena generada. Supongamos que generamos una muestra MCMC de $n = 100$ iteraciones. Si la eficiencia es de 0,25 significa que nuestra muestra es tan eficiente como una muestra con observaciones independientes de tamaño $n \times 0{,}25 = 100 \times 0{,}25 = 25$ (Nascimento *et al.* 2017). Dicho de otra forma, la eficiencia indica que nuestro tamaño de muestra independiente no es realmente 100, sino 25. Por lo tanto, para obtener una muestra independiente igual a 100 deberíamos realizar $100 \times 25 = 2500$ iteraciones. Este valor se denomina tamaño de muestra efectivo (ESS), y a modo de criterio se deberían tener cadenas con ESS de entre 1000 y 10000 iteraciones (Nascimento *et al.* 2017). Los métodos bayesianos sin embargo, son computacionalmente intensivos, por lo que se recomiendan valores de al menos un ESS = 200.

Resumen de resultados

Una vez que se obtiene una muestra adecuada, es decir que las cadenas convergen y no presentan problemas de mezcla o eficiencia, se vuelve trivial resumir la información, por ejemplo a través de un histograma. La mayoría de los parámetros filogenéticos son variables continuas y se resumen con estadísticos como la media, la mediana y la varianza. En la estadística bayesiana también se reportan los intervalos de credibilidad bayesianos (ICB) al 95%, que se obtienen de eliminar el 2,5% de los valores inferiores y superiores de la muestra. Este intervalo contiene al verdadero parámetro poblacional con una probabilidad de 0,95.

La distribución *a posteriori* de las topologías y longitudes de las ramas es más difícil de resumir. Si hay pocas topologías con altas probabilidades *a posteriori*, se puede hacer una lista con todas las topologías y sus probabilidades, o simplemente reportar la topología con la máxima probabilidad *a posteriori*. Sin embargo, la mayoría de las probabilidades *a posteriori* contienen muchas topologías con una probabilidad razonablemente alta, por lo que es necesario usar otros métodos. El enfoque más común para resumir las topologías es reportar la frecuencia de los clados más comunes, dado que hay muchos menos clados que topologías. Además, todos los clados presentes en al menos un 50% pueden ser visualizados en un árbol de consenso por mayoría (ver Cap. 7).

Para resumir las longitudes de las ramas quizás la mejor forma sea mostrar la distribución de las longitudes de las ramas para cada topología. Sin embargo, si hay muchas topologías podría no haber suficientes muestras de longitudes de las ramas para cada una. Un enfoque razonable consiste en agrupar las muestras de longitud de las ramas que corresponden al mismo clado. Estas longitudes de las ramas agrupadas pueden mostrarse también en un árbol de consenso. Sin embargo, las distribuciones agrupadas pueden ser multimodales (múltiples picos) dado que los valores muestreados en la mayoría de los casos vienen de diferentes topologías, y una medida simple como la media puede ser errónea.

Selección de modelos de evolución de secuencias

Una vez identificada la topología del árbol para un determinado modelo de evolución de secuencias, podemos comparar distintos modelos y evaluar el grado de ajuste a los datos, P(secuencias | modelo de evolución). Supongamos que se comparan dos modelos de evolución M_0 y M_1, con probabilidades *a priori* $P(M_0)$ y $P(M_1)$. Podemos calcular el cociente de las probabilidades (denominado *odd*) *a posteriori* como:

$$\frac{\text{probabilidad } a\ posteriori(M_1)}{\text{probabilidad } a\ posteriori(M_0)} = \frac{\text{probabilidad } a\ priori(M_1)}{\text{probabilidad } a\ priori(M_0)} \times \frac{L(M_1)}{L(M_0)} \times \frac{P(\text{secuencias})}{P(\text{secuencias})}$$

$$odd\ a\ posteriori = odd\ a\ priori \times \frac{L(M_1)}{L(M_0)}$$

Note que *P*(secuencias) es igual para ambos términos, independientemente del modelo, por lo que se simplifican. Así, el *odd a posteriori* es igual al *odd a priori* por el cociente de las verosimilitudes, este último denominado factor de Bayes (Jeffreys 1935). Este factor resume la evidencia de una hipótesis o modelo dados los datos, con respecto a otro modelo. En la práctica, si el factor de Bayes (*B*) es grande, implica que el modelo del numerador tiene muchas más chances de ser el modelo correcto con respecto al modelo del denominador. Si bien "qué tan grande" resulta subjetivo y depende del investigador, se han sugerido algunos criterios que se resumen en la Tabla 8.1 (Kass y Raftery 1995).

Tabla 8.1. Criterios para establecer evidencia a favor de un modelo M_0 utilizando el factor de Bayes (*B*).

$B = L(M_1) / L(M_0)$	Evidencia a favor de M_1
1–3	Muy débil
3–20	Moderada
20–150	Fuerte
>150	Muy fuerte

Cadenas de Markov Monte Carlo *vs. bootstrap*

Como se mencionó anteriormente las MCMC también generan una muestra de árboles. Por lo tanto, el número de veces que un grupo aparece en la muestra de árboles puede utilizarse como medida de soporte del grupo, al igual que en el *bootstrap* (Simmons *et al.* 2004; ver Cap. 7) y se denomina soporte de credibilidad bayesiano.

A diferencia del *bootstrap,* las MCMC producen una muestra de árboles mucho mayor en el mismo tiempo computacional, porque generan un árbol para cada propuesta *vs.* un árbol por búsqueda (que evalúa numerosas alternativas; Holder y Lewis 2003). Sin embargo, los árboles generados por las MCMC están altamente correlacionados, es decir que son muy parecidos entre sí debido a que los cambios que se van proponiendo son pequeños. Por lo tanto, se requieren varios millones de réplicas, mientras que se requieren muchas menos en el *bootstrap* (en el orden de miles) para la mayoría de los problemas (Holder y Lewis 2003).

Si bien los valores de soporte *bootstrap* pueden compararse con las probabilidades *a posteriori* (Holmes 2003), se han observado grandes discrepancias entre los valores de soporte de credibilidad bayesianos y *boostrap* (Alfaro *et al.* 2003), y que estos últimos son más propensos a soportar hipótesis filogenéticas falsas (Douady *et al.* 2003, Erixon *et al.* 2003). Sin embargo, también se ha demostrado que las probabilidades *a posteriori* pueden ser excesivamente liberales (es decir, rechazan hipótesis con facilidad; Suzuki *et al.* 2002).

Finalmente en la Tabla 8.2 se resumen las principales diferencias entre la MV y la inferencia bayesiana. Si bien ambos enfoques operan conceptualmente de forma inversa (probabilidad de un conjunto de secuencias dado un árbol *vs.* probabilidad de un árbol dado un conjunto de secuencias), pareciera a simple vista que deberían dar el mismo resultado. Sin embargo, basta considerar un ejemplo muy sencillo para ver la diferencia. La probabilidad de que haya estado nublado dado que llovió (enfoque bayesiano) será muy alta (cercana al 100%), mientras que la probabilidad de que llueva dado que está nublado (verosimilitud) no necesariamente será alta (puede ser del 50% o incluso menos).

Tabla 8.2. Principales diferencias entre la MV y la inferencia bayesiana.

	Máxima verosimilitud	**Inferencia bayesiana**
Probabilidad	Probabilidad de un conjunto de datos, dado un parámetro	Probabilidad de un parámetro, dado un conjunto de datos
Enfoque	Frecuentista. Definida en el contexto de frecuencias relativas a largo plazo	Describe cualquier fenómeno incierto
Parámetros	Fijos y desconocidos	Aleatorios
Naturaleza del método	Objetivo	Subjetivo

Box 8.1. Terminología asociada a los métodos probabilísticos

Probabilidad: valor entre 0 y 1 que mide la certidumbre asociada a un suceso o evento que ocurre aleatoriamente.
Modelo evolutivo o de sustitución de secuencias: modelo probabilístico que describe los eventos de mutación de un nucleótido a otro en un sitio.
Verosimilitud: probabilidad de obtener un conjunto de datos, dado un conjunto de parámetros estadísticos. En un contexto filogenético, corresponde a la probabilidad de obtener un conjunto de secuencias de ADN, dado un determinado árbol y modelo de sustitución de secuencias. Los parámetros del árbol incluyen la topología y un modelo de sustitución de secuencias (que determinan entre otros parámetros, las longitudes de las ramas). La log-verosimilitud corresponde al logaritmo de la verosimilitud.
Máxima verosimilitud: método cuyo objetivo es encontrar el conjunto de parámetros que maximiza la probabilidad de obtener los datos observados. En un contexto filogenético, este método intenta encontrar aquel árbol (dado un modelo de sustitución de secuencias) que maximiza la probabilidad de obtener las secuencias de ADN observadas.
Distribución de probabilidad: función matemática que asigna a cada evento, definido sobre una variable aleatoria, la probabilidad de que dicho evento ocurra.
Teorema de Bayes: proposición que expresa la probabilidad de un parámetro estadístico, dado un conjunto de datos. En un contexto filogenético, permite calcular la probabilidad de un determinado árbol, dado un conjunto de secuencias de ADN y un modelo de sustitución de secuencias.
Probabilidad *a priori*: probabilidad de que ocurra un evento antes de tener en cuenta los datos u observaciones. En un contexto filogenético, representa la probabilidad de un árbol bajo conocimiento previo (sin observar ningún dato).
Probabilidad *a posteriori*: probabilidad de que ocurra un evento después de tener en cuenta los datos u observaciones. En un marco filogenético, representa la probabilidad de un determinado árbol, dado un conjunto de secuencias de ADN y la *probabilidad a priori*.
Cadenas de Markov Monte Carlo (MCMC): técnica de simulación cuyo objetivo es obtener muestras de una distribución de probabilidad. En el análisis bayesiano, se utiliza para generar una muestra de la distribución *a posteriori*, a partir de la cual pueden utilizarse estimadores estadísticos clásicos, como la media o el desvío estándar para resumir la información. Un árbol de consenso por MCMC es simplemente un resumen de la distribución *a posteriori* de la topología de un árbol.
Fase estacionaria: fase de una cadena de MCMC donde se estabiliza (converge) el valor de verosimilitud.
***Burn-in*:** fase inicial de una cadena de MCMC que finaliza antes de alcanzar la fase estacionaria, suele descartarse del análisis.
Mezcla: velocidad o tasa a la cual una cadena atraviesa eficientemente la distribución *a posteriori* una vez alcanzada la fase estacionaria.

Thinning: método que consiste en extraer valores de la cadena cada una cierta cantidad de iteraciones.
Eficiencia: medida que relaciona la varianza de la muestra *a posteriori* con la varianza de una muestra independiente del mismo tamaño que la cadena.
Tamaño de muestra efectivo: medida real del tamaño de muestra *a posteriori* (con observaciones independientes). Se calcula como el producto entre el número de iteraciones de la cadena y la eficiencia.
Potential Scale Reduction Factor **(PSRF):** medida de la convergencia de las cadenas. Se calcula como el cociente entre la varianza entre cadenas y la varianza dentro de las diferentes cadenas. A medida que las cadenas convergen, las varianzas son cada vez más similares y el PSRF se aproxima a 1.
Gráfico traza: gráfico que muestra la verosimilitud o valores del parámetro *vs.* las iteraciones de la MCMC.
Odd: cociente entre dos probabilidades. Es una medida de cuántas veces es más probable que ocurra un evento en relación a otro.
Odd a priori: cociente entre las probabilidades *a priori* de dos modelos estadísticos.
Factor de Bayes: cociente entre las verosimilitudes de dos modelos estadísticos.
Odd a posteriori: producto entre el *odd a priori* y el factor de Bayes de dos modelos estadísticos. Es una medida del soporte o evidencia de una hipótesis, representada por un modelo estadístico dados los datos, en relación a otro modelo. Cuando ambos modelos son igualmente probables *a priori*, el *odd a posteriori* es igual al factor de Bayes.
Intervalo de credibilidad bayesiano (ICB): intervalo que incluye al verdadero parámetro poblacional, con probabilidad 1 – *P*, dados los datos. Así, un ICB del 95% (1 – 0,05) incluye al parámetro poblacional con una probabilidad de 0,95.
Soporte de credibilidad bayesiano: número de veces que un grupo aparece en la muestra de árboles de la distribución *a posteriori*.
Criterio de información: medida de la calidad relativa de un modelo estadístico, para un conjunto de datos dado. Tiene en cuenta la verosimilitud y el número de parámetros. A mayor verosimilitud y menor número de parámetros, el modelo se considera relativamente mejor con respecto a otro modelo (principio de parsimonia).

INTRODUCCIÓN AL ANÁLISIS FILOGENÉTICO EN R

El análisis filogenético es un área extremadamente activa dentro del entorno de R. De hecho, actualmente existen más de 100 paquetes que permiten realizar distintos tipos de análisis filogenéticos, que pueden consultarse en https://cran.r-project.org/web/views/Phylogenetics.html. En este capítulo se presentarán dos muy utilizados, ape (Paradis y Schliep 2018) y phangorn (Schliep 2011). Para una mayor profundidad de los temas abordados en esta sección se recomienda la lectura de Paradis (2012). También existen excelentes programas específicos del método de parsimonia fuera de R, tales como PAUP (Swofford 2002) y TNT (Goloboff *et al.* 2008).

El análisis filogenético por MV también cuenta con diversos software, como MEGA (Kumar *et al.* 1993), PAML (Yang 1997), TREE-PUZZLE (Schmidt *et al.* 2002), PHYML (Guindon y Gascuel 2003), PHYLIP (Felsenstein 2004) y RAxML (Stamatakis *et al.* 2004). En el caso del análisis filogenético bayesiano es importante resaltar que está poco desarrollado dentro de R, y cuenta con software ampliamente utilizados por fuera de este entorno, como MrBayes (Huelsenbeck y Ronquist 2001) y BEAST (Drummond y Rambaut 2007), motivo por el cual, el análisis filogenético bayesiano no va a ser desarrollado en R en este libro.

Pasos previos

Para los métodos que aplicaremos a continuación, tomaremos secuencias alineadas del gen mitocondrial citocromo *b*, del género de pequeños loros neotropicales *Brotogeris* (Psittacidae) conformado por ocho especies (Ribas *et al.* 2009, Fig. 8.14A). Estas secuencias pueden descargarse del sitio web de GenBank (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch), utilizando en la búsqueda el número de acceso reportado en un determinado estudio. A modo de ejemplo, buscamos el número de acceso FJ652848 que corresponde a *B. tirica* en Ribas *et al.* (2009). En los resultados de la búsqueda se listan todas las secuencias que son similares genéticamente, por lo que en general no es necesario buscar número por número (Fig. 8.15).

Fig. 8.14. (A) Catita Chiriri (*Brotogeris chiriri*); (B) Loro Maitaca (*Pionus maximiliani*). Fotografías: Palacio, FX.

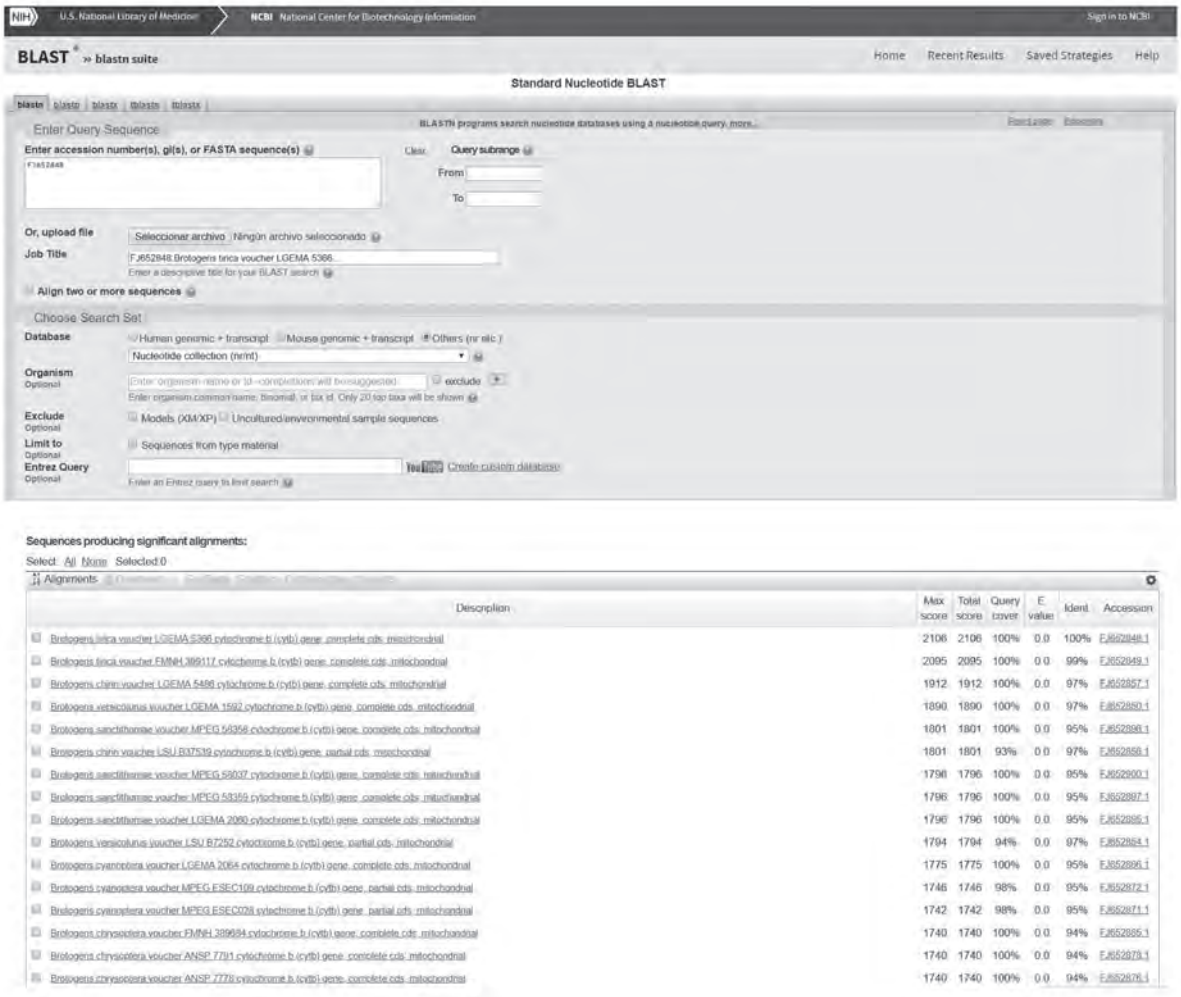


Fig. 8.15. Plataforma de GenBank para la búsqueda y descarga de secuencias de nucléotidos, proteínas y genomas. A modo de ejemplo, se muestra la búsqueda del número de acceso FJ652848 correspondiente a *Brotogeris tirica*.

Como grupo externo (*outgrup*) seleccionamos una especie de loro de un género afín a *Brotogeris*, *Pionus maximiliani* (Fig. 8.14B). Debe tenerse en cuenta que cada secuencia corresponde a un único individuo, por lo tanto, en la filogenia resultante las terminales corresponderán a individuos. Como nuestro objetivo es didáctico, se seleccionará una única secuencia por especie. Una vez seleccionada la secuencia debemos ingresar en *Secuence ID*, donde se encuentra toda la información correspondiente al *voucher*. Para poder descargar la secuencia ingresamos en *FASTA*. En bioinformática, este es un formato basado en texto utilizado para representar secuencias de ADN, ARN y aminoácidos. La primera línea es fundamental y corresponde al nombre de la secuencia antecedido por un signo >. Éste sirve para reconocer el formato y el nombre de la secuencia cuando la importemos en algún software. Lo siguiente consiste en copiar y pegar todo el texto (nombre de la secuencia más la secuencia en sí misma) en un archivo de texto (.txt). Para cada secuencia repetimos el procedimiento y pegamos debajo de la secuencia anterior, separando con un *Enter*. De esta forma, obtendremos un único archivo con todas las secuencias separadas por un espacio (ver el archivo "loros genbank.txt" disponible en https://fundacionazara.org.ar/analisis-multivariado-para-datos-biologicos/). Eventualmente, las secuencias pueden no estar alineadas, para lo cual se requiere previamente alinearlas mediante algún software.

Una vez preparado el archivo podemos importarlo en R con la función `read.FASTA()` del paquete ape.

```
> library(ape)
> loros <- read.FASTA("C:/R datos/loros genbank.txt")
```

Si exploramos el objeto veremos el número de secuencias, su longitud, las etiquetas de las primeras secuencias, así como la proporción de bases (A, G, C, T) en el conjunto de secuencias.

```
> loros
9 DNA sequences in binary format stored in a list.

All sequences of same length: 1140

Labels:
FJ652848.1 Brotogeris tirica
FJ652857.1 Brotogeris chiriri
FJ652850.1 Brotogeris versicolurus
FJ652896.1 Brotogeris sanctithomae
FJ652866.1 Brotogeris cyanoptera
FJ652885.1 Brotogeris chrysoptera
...

Base composition:
    a     c     g     t
0.285 0.360 0.126 0.229
```

Para construir filogenias, previamente es necesario convertir el conjunto de secuencias a otro formato, denominado phyDat, con el paquete phangorn. Debemos especificar si el tipo de dato (argumento `type`) es ADN, aminoácido u otro.

```
> library(phangorn)
> loros_phyDat <- phyDat(loros, type = "DNA")
> names(loros_phyDat)
[1] "FJ652848.1 Brotogeris tirica"       "FJ652857.1 Brotogeris chiriri"
[3] "FJ652850.1 Brotogeris versicolurus" "FJ652896.1 Brotogeris sanctithomae"
```

```
[5] "FJ652866.1 Brotogeris cyanoptera"    "FJ652885.1 Brotogeris chrysoptera"
[7] "FJ652903.1 Brotogeris jugularis"     "FJ652862.1 Brotogeris pyrrhoptera"
[9] "EF517622.1 Pionus maximiliani"
```

Además, vamos a modificar las etiquetas de las secuencias, de forma que sólo quede el nombre científico de las especies (este paso puede obviarse, en el árbol suelen dejarse las etiquetas de las secuencias). Para esto, quitamos los 12 primeros caracteres que corresponden al número de acceso.

```
> names(loros_phyDat) <- substring(names(loros), 12)
> names(loros_phyDat)
[1] "Brotogeris tirica" "Brotogeris chiriri" "Brotogeris versicolurus"
[4] "Brotogeris sanctithomae" "Brotogeris cyanoptera" "Brotogeris chrysoptera"
[7] "Brotogeris jugularis" "Brotogeris pyrrhoptera" "Pionus maximiliani"
```

Parsimonia

En los casos en los que hay un número reducido de taxones ($N \leq 10$), es posible encontrar el árbol más parsimonioso mediante el método de *branch and bound* (Hendy y Penny 1982). Recuerde que hay que seleccionar el tipo de parsimonia. La función `bab()` admite los métodos de Fitch y Sankoff.

```
> bab.fitch <- bab(data = loros_phyDat, method ="fitch")
> bab.sankoff <- bab(data = loros_phyDat, method ="sankoff")
```

Los objetos generados son de tipo multiPhylo, es decir que contienen múltiples filogenias. En este caso, dado que hay un único árbol más parsimonioso, cada objeto contiene un único árbol. Si escribimos el nombre de alguno de los dos objetos (`bab.fitch` ó `bab.sankoff`) obtenemos el número de terminales, el número de nodos, la información sobre si está enraizado o no, y si la filogenia incluye las longitudes de las ramas.

```
> class(bab.fitch)
[1] "multiPhylo"
> bab.fitch[[1]]

Phylogenetic tree with 9 tips and 7 internal nodes.

Tip labels:
      Pionus maximiliani, Brotogeris tirica, Brotogeris chiriri, Brotogeris versi-
colurus, Brotogeris sanctithomae, Brotogeris cyanoptera, ...

Unrooted; no branch lengths.
```

Observe que la función contiene un árbol filogenético no enraizado. El siguiente paso consiste en graficar la filogenia con el fin de ver las relaciones entre los taxones. El comando más simple para graficar una filogenia es el comando `plot()` que permite representar varios tipos de árboles (no enraizado, filograma, cladograma, árbol en abanico; Fig. 8.16). El término filograma utilizado en R se corresponde con un árbol filogenético representado por ángulos rectos en las dicotomías y politomías. Sin embargo, debemos enraizarlo utilizando la función `root()` y especificando el *outgroup* (*P. maximiliani*).

```
> bab.fitch_raiz <- root(bab.fitch, outgroup = "Pionus maximiliani",
+                        resolve.root = TRUE)
```

```
> bab.sankoff_raiz <- root(bab.fitch, outgroup = "Pionus maximiliani",
+                          resolve.root = TRUE)
> plot(bab.fitch, type = "unrooted", edge.width = 2)
> plot(bab.fitch_raiz, type = "phylogram", edge.width = 2)
> plot(bab.fitch_raiz, type = "fan", edge.width = 2)
> plot(bab.fitch_raiz, type = "cladogram", edge.width = 2)
```

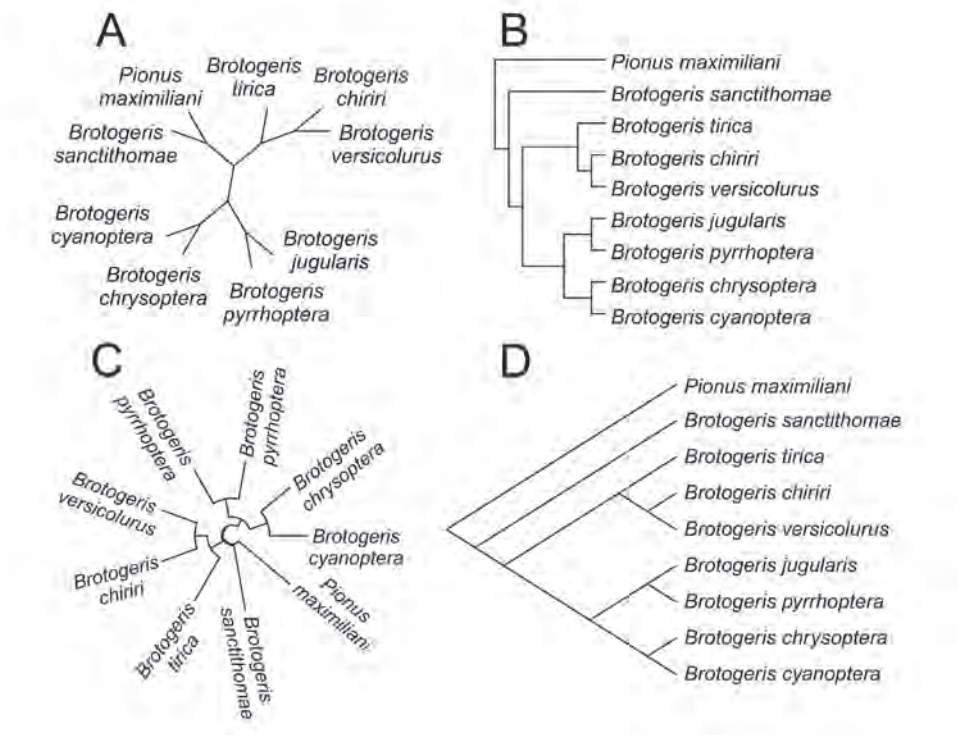


Fig. 8.16. Distintas formas de graficar filogenias, obtenidas mediante el método de *branch and bound* y parsimonia de Fitch sobre la base de secuencias de ADN de nueve especies de loros. (A) Árbol no enraizado; (B) filograma enraizado convencional; (C) árbol enraizado en abanico; (D) cladograma enraizado.

En los casos en los que hay más de 10 taxones debemos utilizar métodos heurísticos, como el método de *ratchet* (Nixon 1999; Fig. 8.17). Si bien la filogenia del ejemplo sólo contiene nueve especies, se muestra el método con fines didácticos.

```
> pratchet.fitch <- pratchet(loros_phyDat, method = "fitch")
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
[1] "Best pscore so far: 290"
> plot(pratchet.fitch, edge.width = 2)
> pratchet.fitch_raiz <- root(pratchet.fitch, outgroup = "Pionus maximiliani",
+                             resolve.root = TRUE)
> plot(pratchet.fitch_raiz, edge.width = 2)
```

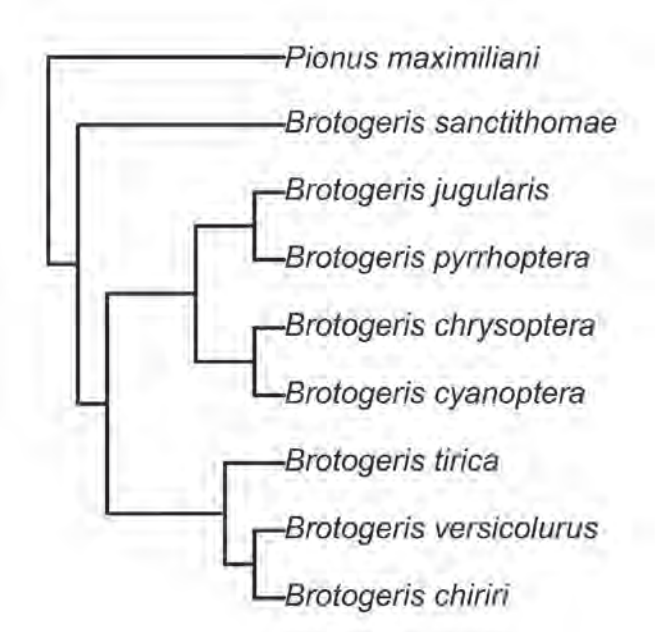


Fig. 8.17. Filogenia obtenida mediante el método de *ratchet* y parsimonia de Fitch sobre la base de secuencias de ADN de nueve especies de loros.

En este caso, ambos métodos dan como resultado la misma topología (observe que las ramas de las Figs. 8.16 y 8.17 sólo están rotadas). La función `parsimony()` devuelve la longitud del árbol (*score* de parsimonia), representado por el número mínimo de pasos.

```
> parsimony(bab.fitch, data = loros_phyDat, method = "fitch")
[1] 290
> parsimony(bab.fitch, data = loros_phyDat, method = "sankoff")
[1] 290
```

También podemos calcular el índice de consistencia (CI = mínimo número de cambios / longitud del árbol) y el índice de retención (RI = [máximo número de cambios – cambios observados] / [máximo número de cambios – mínimo número de cambios]).

```
> CI(tree = bab.fitch[[1]], data = loros_phyDat)
[1] 0.7655172
> RI(tree = bab.fitch[[1]], data = loros_phyDat)
[1] 0.6222222
```

Dado que los métodos heurísticos no garantizan encontrar el árbol más parsimonioso, podemos aplicar métodos de permutación de ramas para intentar encontrar árboles más cortos. La función `optim.parsimony()` permite implementar los métodos de NNI (*nearest neighbor interchange*) y SPR (*subtree pruning and regrafting*).

```
> optimPars <- optim.parsimony(tree = pratchet.fitch_raiz,
+                              data = loros_phyDat, rearrangements = "NNI")
Final p-score 290 after 0 nni operations
> treePars_raiz <- root(optimPars, outgroup = "Pionus maximiliani",
+                       resolve.root = TRUE)
```

En este caso puntual y debido al reducido número de taxones, los métodos de permutación de ramas no encuentran otro árbol con menor número de cambios, aunque en la mayoría de los casos se suelen obtener varios árboles igualmente parsimoniosos. En estas situaciones se calcula un árbol de consenso, generalmente de mayoría del 50%, que puede ser obtenido con la función `consensus()` sobre el conjunto de árboles, y se debe especificar qué proporción de árboles presentan un determinado clado (argumento `p = 0.5`).

A continuación, podemos mapear u optimizar los caracteres sobre el árbol final. Para esto, vamos a utilizar el paquete BiocManager (Morgan 2018) que pertenece al proyecto Bioconductor, software libre de bioinformática (https://www.bioconductor.org/), que a su vez contiene varios paquetes. Debe tenerse en cuenta que BiocManager sólo está disponible para versiones de R 3.5.0 o posteriores. Para poder graficar los caracteres sobre la filogenia, debemos descargar el paquete seqLogo (Bembom 2018) perteneciente a BiocManager. Para mapear caracteres y reconstruir los estados ancestrales vamos a aplicar el método ACCTRAN. Se debe tener en cuenta que todas las reconstrucciones ancestrales para parsimonia se basan en el método de Fitch, y hasta el momento sólo se permiten árboles bifurcados.

```
> library(BiocManager)
> BiocManager::install("seqLogo")
> library(seqLogo)
> bab.fitch.acctran <- acctran(tree = bab.fitch_raiz,
+                              data = loros_phyDat)[[1]]
> anc.acctran <- ancestral.pars(tree = bab.fitch.acctran, data = loros_phyDat,
+                               type = "ACCTRAN")
```

Para optimizar un carácter sobre el árbol obtenido debemos especificar el número de columna (argumento `i`) correspondiente (sitio de ADN en este caso). A modo de ejemplo se muestra el mapeo de los sitios 5 y 11 (Fig. 8.18). Debe tenerse en cuenta que cuando en un nodo se representa más de un estado, la estimación no es única. Los estados ancestrales se estiman por MV. Los colores para representar las bases siguen el orden A, C, G, T.

```
> plotAnc(tree = bab.fitch_raiz, anc.acctran, i = 5,
+         col = c("black", "white", "gray", "white"))
> plotAnc(tree = bab.fitch_raiz, anc.acctran, i = 11,
+         col = c("black", "white", "gray", "white"))
```

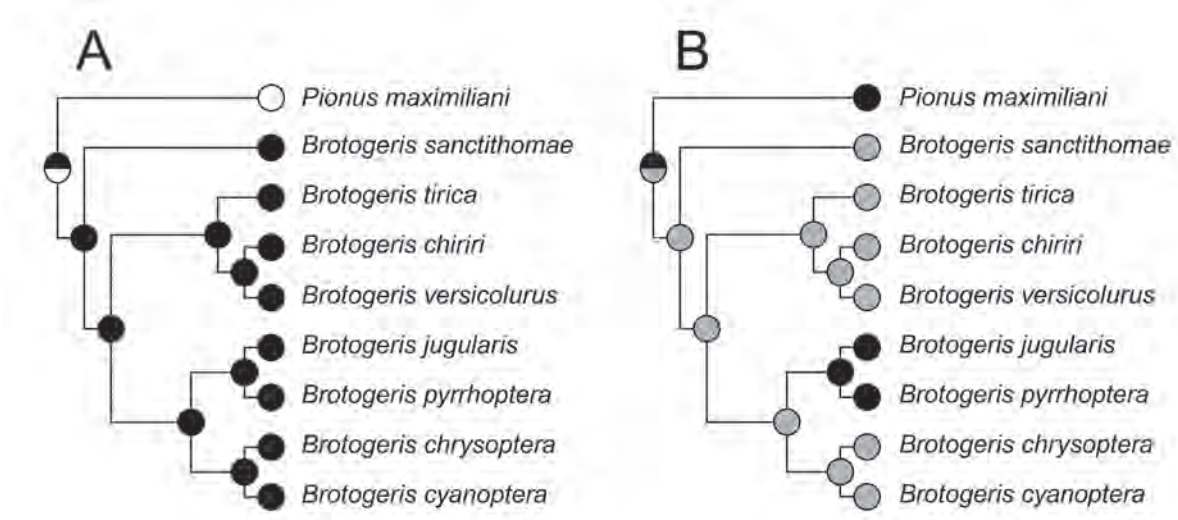


Fig. 8.18. Mapeo de caracteres utilizando el método de *branch and bound* con parsimonia de Fitch y método ACCTRAN. Los círculos representan gráficos de torta donde se muestra la probabilidad de cada base nitrogenada (negro: adenina, blanco: citosina, gris: guanina). (A) sitio 5; (B) sitio 11.

Por último, vamos a aplicar el método de *bootstrap* para evaluar el soporte de las ramas (en este caso con el método de *ratchet*). Para esto, debemos especificar el número de réplicas *bootstrap* (argumento `bs`). Luego enraizamos cada pseudo-réplica extrayendo el *ougroup*.

```
> arboles.boot <- bootstrap.phyDat(loros_phyDat, FUN = pratchet,
+                                  method = "fitch", bs = 1000)
> boot_raiz <- root(arboles.boot, outgroup = "Pionus maximiliani",
+                   resolve.root = TRUE)
```

El soporte de las ramas calcula en qué proporción un clado de la filogenia final está presente en la serie de los *N* árboles generados.

```
> soportes.boot <- prop.clades(phy = pratchet.fitch_raiz, boot_raiz)
```

La filogenia final, con la longitud y soporte de las ramas (Fig. 8.19A) puede graficarse con la función `plot()`. La función `nodelabels()` permite ajustar el tamaño y la posición de los valores de soporte. Los valores de soporte representan cuántas veces aparecen los clados en el conjunto de 1000 árboles. Si se quiere expresar este valor como un porcentaje, simplemente dividimos el objeto `soportes.boot` por 10.

```
> plot(pratchet.fitch_raiz, type = "phylogram", show.node.label = TRUE)
> nodelabels(soportes.boot/10, adj = c(1.1, -0.3), frame = "none")
```

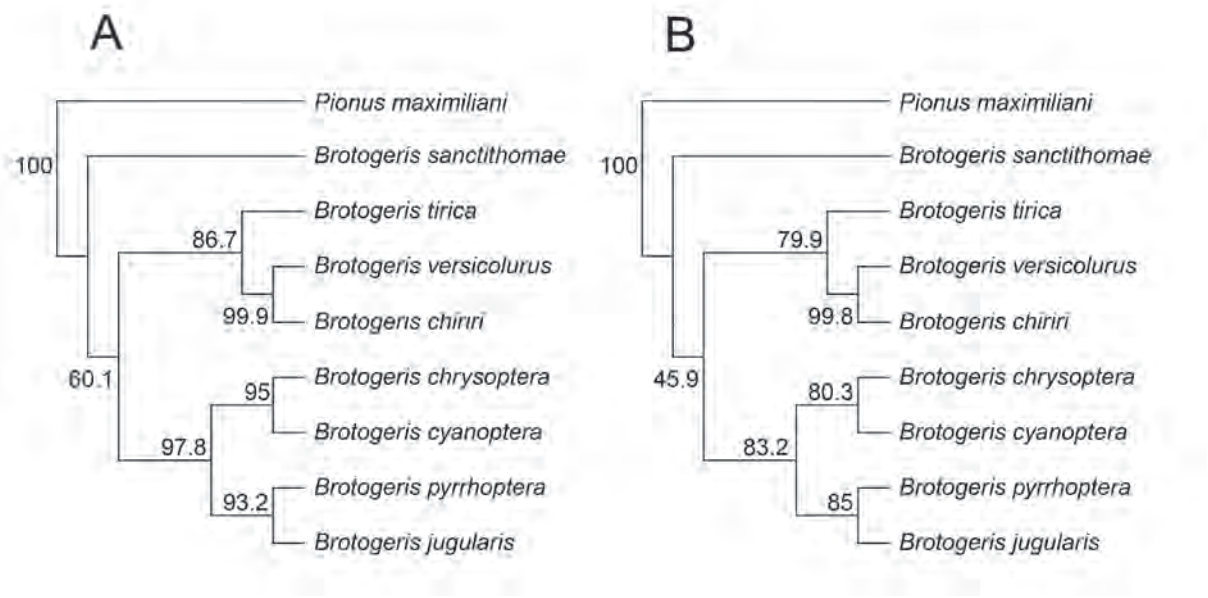


Fig. 8.19. Filogenia obtenida mediante el método de *ratchet* y parsimonia de Fitch sobre la base de secuencias de ADN de nueve especies de loros. Se muestran los soportes de las ramas obtenidos por 1000 pseudo-réplicas *bootstrap* (A) y *jackknife* (B).

Debido a que los paquetes ape y phangorn no tienen una función para calcular los valores de soporte *jackknife*, vamos a construir nuestra propia función para calcular estos valores, que denominamos `jackknife.phyDat()`. Este procedimiento se puede obviar sin mayores inconvenientes. A los fines prácticos, copiamos y pegamos las líneas de código en la consola, que representa una función con tres argumentos: la matriz de datos (`data`), la proporción de la matriz que debe muestrarse en cada pseudo-réplica (`p`) y el número de pseudo-réplicas (`replicates`). Por defecto, los valores de los dos últimos argumentos son 0,5 y 100, respectivamente. Recuerde que el *bootstrap* realiza un muestreo con reemplazo y genera matrices del mismo tamaño que la original, mientras que el *jackknife* realiza un muestreo sin reemplazo y genera matrices de menor tamaño que la original. Para el lector interesado en programación en R se recomienda la lectura de Matloff (2011), Teetor (2011) y Burns (2012).

```
> jackknife.phyDat <- function(data, p = 0.5, replicates = 100){
+ num.caract <- length(data[[1]]) # número de caracteres en la matriz
+ num.taxa <- length(data) # número de taxones en la matriz
+ n.muestras <- round(p*num.caract, 0) # número de caracteres a extraer
+ arboles <- NULL
+
+ for (i in 1:replicates){
+  subset.data <- subset(data, select = sample(1:num.caract, # muestreo sin
+                        size = n.muestras, replace = FALSE)) # reemplazo
+ arboles[[i]] <- pratchet(subset.data, method = "fitch") # parsimonia
+ }
+ arboles
+ }
```

Una vez creada nuestra función para *jackknife*, podemos aplicarla a nuestra matriz y seguir los mismos pasos que en el *bootstrap*, esto es, extraer *N* árboles, enraizarlos, calcular los valores de soporte y graficar la filogenia final junto con estos valores (Fig. 8.19B). En este caso vamos a utilizar 1000 pseudo-réplicas.

```
> arboles.jack <- jackknife.phyDat(data = loros_phyDat, replicates = 1000)
> class(arboles.jack) <- "multiPhylo"
> jack_raiz <- root(arboles.jack, outgroup = "Pionus maximiliani",
+                   resolve.root = TRUE)
> soportes.jack <- prop.clades(phy = pratchet.fitch_raiz, jack_raiz)
> plot(pratchet.fitch_raiz, type = "phylogram", show.node.label = TRUE)
> nodelabels(soportes.jack/10, adj = c(1.1, -0.3), frame = "none")
```

Neighbor-joining

Alternativamente, es posible construir un árbol basado en distancias genéticas. Para esto, es necesario calcular una matriz de distancia (MD) entre las secuencias –función `dist.ml()`– y elegir la distancia genética a utilizar. Con fines didácticos tomaremos el modelo JC69, que es el modelo más simple de evolución (igual tasa de sustitución y frecuencia de bases).

```
> dna_dist <- dist.ml(loros_phyDat, model = "JC69")
```

Una vez obtenida la MD, es posible obtener un árbol filogenético no enraizado mediante *neighbor-joining*.

```
> loros_NJ <- NJ(dna_dist)
> class(loros_NJ)
[1] "phylo"
> loros_NJ

Phylogenetic tree with 9 tips and 7 internal nodes.

Tip labels:
      FJ652848.1 Brotogeris tirica , FJ652857.1 Brotogeris chiriri, FJ652850.1 Brotogeris versicolurus, FJ652896.1 Brotogeris sanctithomae , FJ652866.1 Brotogeris cyanoptera, FJ652885.1 Brotogeris chrysoptera, ...

Unrooted; includes branch lengths.
```

A continuación, graficamos la filogenia (Fig. 8.20).

```
plot(loros_NJ, type = "unrooted", edge.width = 2)
```

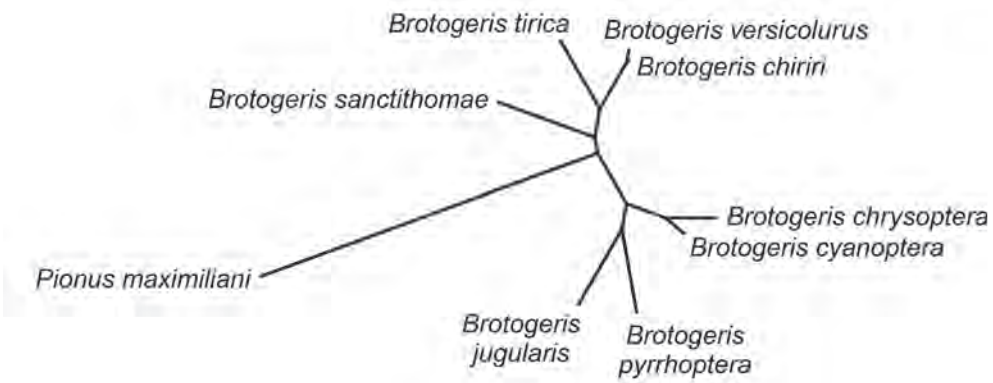


Fig. 8.20. Filogenia no enraizada obtenida por *neighbor-joining* (distancia JC69) sobre la base de secuencias de ADN de nueve especies de loros.

Ahora se debe enraizar la filogenia (Fig. 8.21). El argumento `resolve.root = TRUE` indica que la raíz debe representar un nodo bifurcado (esto es importante en parsimonia, donde uno de sus supuestos es que el *ingroup* debe ser monofilético).

```
> loros_NJ.raiz <- root(loros_NJ, outgroup = "Pionus maximiliani",
+                       resolve.root = TRUE)
> plot(loros_NJ.raiz, type = "phylogram", edge.width = 2)
```

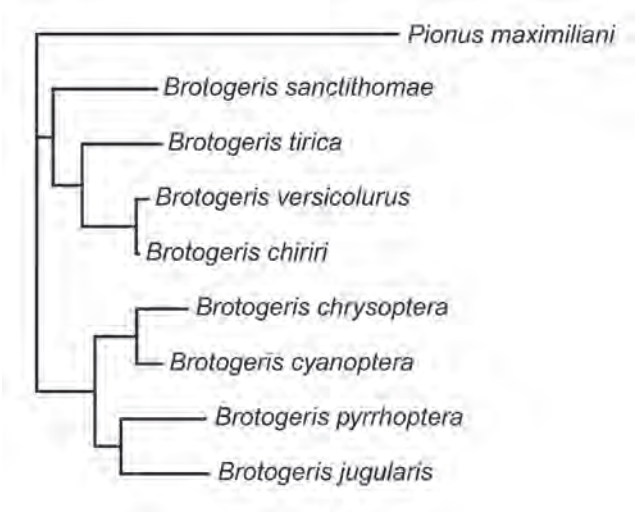


Fig. 8.21. Filogenia enraizada obtenida mediante *neighbor-joining* (distancia JC69) sobre la base de secuencias de ADN de nueve especies de loros. Se incluyen las longitudes de las ramas.

Si bien el árbol está enraizado –corroborar con la función `is.rooted()`–, esto no se refleja en la Figura 8.21. Si quitamos las longitudes de las ramas y representamos sólo la topología vemos que sí está enraizado (Fig. 8.22). Esto significa que la cantidad de cambios en las secuencias de ADN a nivel de la raíz es prácticamente nula para el gen utilizado y el modelo de sustitución JC69, y por lo tanto no se ve reflejada en la filogenia que incluye las longitudes de las ramas.

```
> is.rooted(loros_NJ.raiz)
[1] TRUE
> plot(loros_NJ.raiz, use.edge.length = FALSE, edge.width = 2)
```

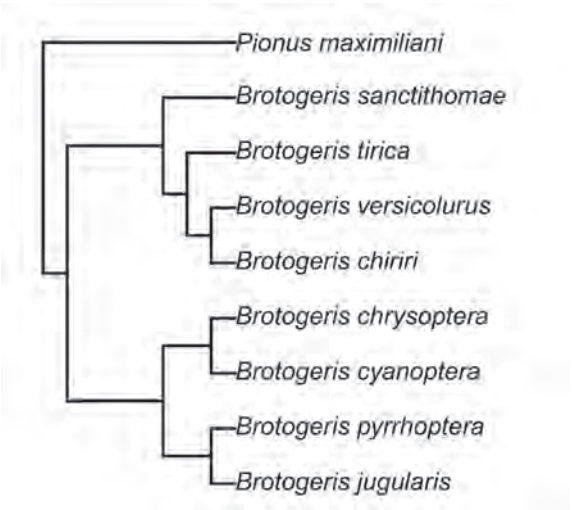


Fig. 8.22. Filogenia obtenida mediante *neighbor-joining* (distancia JC69) sobre la base de secuencias de ADN de nueve especies de loros. No se incluyen las longitudes de las ramas.

Máxima verosimilitud

El concepto de MV aplicado a las filogenias implica encontrar aquel árbol que maximiza la probabilidad de obtener las secuencias observadas. Si bien suena confuso, esto corresponde a la verosimilitud

de un árbol dadas las secuencias *L*(árbol | datos) = *P*(datos | árbol). Para entender mejor este concepto podemos calcular la probabilidad de obtener las secuencias observadas, considerando un árbol determinado, por ejemplo el obtenido mediante *neighbor-joining* –función `pml()`–. La función `mlphylo()` del paquete ape también permite construir filogenias mediante MV (Paradis y Schliep 2018).

```
> ML.NJ <- pml(tree = loros_NJ, data = loros_phyDat)
> logLik(ML.NJ)
'log Lik.' -3230.951 (df=15)
```

Esta verosimilitud suele obtenerse en términos logarítmicos (log-verosimilitud). Sin embargo, este valor por sí sólo no nos dice nada, sino que cobra sentido en comparación con la verosimilitud de otro árbol para encontrar aquel que mejor explique los datos. A modo de ejemplo, podemos comparar este modelo con el asumido por el modelo JC69.

```
> ML.JC <- optim.pml(ML.NJ, model = "JC")
optimize edge weights: -3230.951 --> -3226.213
optimize edge weights: -3226.213 --> -3226.213
optimize edge weights: -3226.213 --> -3226.213
optimize edge weights: -3226.213 --> -3226.213
> logLik(ML.JC)
'log Lik.' -3226.213 (df=15)
```

Recuerde que un mayor valor de log-verosimilitud indica un mejor ajuste, porque a medida que una probabilidad se aproxima a 1, el logaritmo de este valor también aumenta. Por lo tanto, en nuestro caso el modelo JC69 ajusta ligeramente mejor que el obtenido por NJ.

Existe una enorme cantidad de modelos de evolución de secuencias (pueden consultarse en la ayuda, y se corresponden a los descritos en Paradis 2006 y Posada 2008). La función `modelTest()` permite comparar 24 modelos de evolución de secuencias (Tabla 8.3).

```
> modelos <- modelTest(loros_phyDat)
[1] "JC+I"
[1] "JC+G"
[1] "JC+G+I"
[1] "F81+I"
[1] "F81+G"
[1] "F81+G+I"
[1] "K80+I"
[1] "K80+G"
[1] "K80+G+I"
[1] "HKY+I"
[1] "HKY+G"
[1] "HKY+G+I"
[1] "SYM+I"
[1] "SYM+G"
[1] "SYM+G+I"
[1] "GTR+I"
[1] "GTR+G"
[1] "GTR+G+I"
```

Tabla 8.3. Tabla resumen de la función `modelTest()`. Se muestran los modelos (Model), los grados de libertad (df), la log-verosimilitud (logLik), el criterio de información de Akaike (AIC), el AIC corregido para muestras pequeñas (AICc), el criterio de información bayesiano (BIC) y los pesos de cada modelo (AICw y AICcw). La definición de cada modelo puede consultarse en la ayuda de la función, y corresponde a la nomenclatura detallada en Posada (2008).

Model	df	logLik	AIC	AICw	AICc	AICcw	BIC
JC	15	–3226.2	6482.4	0.0	6482.9	0.0	6558.0
JC+I	16	–3168.0	6368.1	0.0	6368.6	0.0	6448.7
JC+G	16	–3167.9	6367.9	0.0	6368.3	0.0	6448.5
JC+G+I	17	–3168.1	6370.3	0.0	6370.8	0.0	6455.9
F81	18	–3146.4	6328.8	0.0	6329.5	0.0	6419.5
F81+I	19	–3082.0	6202.0	0.0	6202.7	0.0	6297.7
F81+G	19	–3082.1	6202.1	0.0	6202.8	0.0	6297.9
F81+G+I	20	–3082.8	6205.5	0.0	6206.3	0.0	6306.3
K80	16	–3080.8	6193.6	0.0	6194.1	0.0	6274.2
K80+I	17	–3007.7	6049.5	0.0	6050.0	0.0	6135.1
K80+G	17	–3007.8	6049.6	0.0	6050.1	0.0	6135.2
K80+G+I	18	–3007.9	6051.9	0.0	6052.5	0.0	6142.6
HKY	19	–2998.1	6034.2	0.0	6034.9	0.0	6130.0
HKY+I	20	–2889.1	5818.2	0.1	5818.9	0.1	5919.0
HKY+G	20	–2897.0	5833.9	0.0	5834.7	0.0	5934.7
HKY+G+I	21	–2888.9	5819.8	0.0	5820.6	0.0	5925.6
SYM	20	–3050.8	6141.6	0.0	6142.4	0.0	6242.4
SYM+I	21	–2986.2	6014.4	0.0	6015.2	0.0	6120.2
SYM+G	21	–2985.7	6013.4	0.0	6014.3	0.0	6119.2
SYM+G+I	22	–2985.1	6014.2	0.0	6015.1	0.0	6125.0
GTR	23	–2978.9	6003.8	0.0	6004.7	0.0	6119.6
GTR+I	24	–2883.2	5814.4	0.6	5815.4	0.6	5935.3
GTR+G	24	–2889.3	5826.7	0.0	5827.7	0.0	5947.6
GTR+G+I	25	–2882.8	5815.5	0.3	5816.7	0.3	5941.5

Un problema importante al utilizar la verosimilitud para comparar modelos, es que ésta no tiene en cuenta el número de parámetros del modelo de evolución de secuencias, y a medida que este número aumenta el ajuste tiende a ser mejor. Como alternativa surgen los denominados criterios de información, que tienen en cuenta el grado de ajuste del modelo a través de la verosimilitud, pero penalizan por el número de parámetros que se van incorporando (Burnham y Anderson 2004). De esta forma, un buen modelo es aquel que se ajusta bien a los datos con un número reducido de parámetros (principio de parsimonia). Uno de los más ampliamente utilizados, tanto en ecología como en análisis filogenético, es el criterio de información de Akaike (Akaike 1973, Posada y Crandall 2001, Posada y Buckley 2004, Symonds y Moussalli 2011):

$$\mathrm{AIC} = -2\log(L) + 2k$$

Donde k es el número de parámetros, de esta forma, una mayor verosimilitud (ajuste) tiende a reducir el valor de AIC, mientras que un mayor número de parámetros tiende a aumentarlo. Por lo tanto, un menor valor de AIC de un modelo con respecto a otro modelo indica un mejor ajuste.

Para identificar el mejor modelo en la salida de R seleccionamos aquella fila con el menor valor de AIC. También se reportan los pesos de Akaike (AICw y AICcw), que representan la probabilidad de que un modelo

dado sea el mejor en el conjunto de modelos analizados o candidatos (Posada y Buckley 2004, Symonds y Moussalli 2011). Observe que este modelo también tiene el mayor peso entre los modelos candidatos.

```
>  modelos[which.min(modelos$AIC),]
    Model df    logLik      AIC      AICw    AICc     AICcw       BIC
22 GTR+I 24 -2883.186 5814.372 0.5634443 5815.448 0.5601247 5935.303
```

Otras alternativas al AIC son el AIC corregido para pequeñas muestras (AICc) y el criterio de información bayesiano (BIC), que penalizan aún más por el número de parámetros en el modelo (Posada y Buckley 2004). Basándonos en el menor valor de AIC, ajustamos el modelo de tiempo reversible (GTR, *General Time Reversible*). Éste es el más flexible, ya que asume que todas las tasas de sustitución son diferentes y que las frecuencias de bases pueden ser diferentes (Tavaré 1986, Paradis *et al.* 2004, Gatto *et al.* 2006). También especificamos que queremos optimizar la topología del árbol (`optNni = TRUE`), las frecuencias de las bases (`optBf = TRUE`) y las tasas de todas las sustituciones posibles (`optQ = TRUE`). La variante más compleja (GTR+I+Γ) permite optimizar la proporción de sitios invariantes (`optInv = TRUE`) y modelar la variación en las tasas de sustitución entre sitios (`optGamma = TRUE`).

```
> ML.GTR <- optim.pml(ML.NJ, model = "GTR", optNni = TRUE, optBf = TRUE,
+                     optQ = TRUE)
optimize edge weights:  -3230.951 --> -3226.213
optimize base frequencies:  -3226.213 --> -3146.52
optimize rate matrix:  -3146.52 --> -2979.741
optimize edge weights:  -2979.741 --> -2979.377
optimize topology:  -2979.377 --> -2979.343
optimize topology:  -2979.343 --> -2979.343
1
optimize base frequencies: -2979.343 --> -2978.95
optimize rate matrix: -2978.95 --> -2978.871
optimize edge weights: -2978.871 --> -2978.87
optimize topology: -2978.87 --> -2978.87
0
optimize base frequencies: -2978.87 --> -2978.859
optimize rate matrix: -2978.859 --> -2978.857
optimize edge weights: -2978.857 --> -2978.857
optimize base frequencies: -2978.857 --> -2978.857
optimize rate matrix: -2978.857 --> -2978.857
optimize edge weights: -2978.857 --> -2978.857
```

Este nuevo modelo nos arroja la matriz de transición entre bases, así como las frecuencias de las bases.

```
> ML.GTR

 loglikelihood: -2978.857

unconstrained loglikelihood: -2670.98

Rate matrix:
        a           c           g        t
a   0.000 3.438620e+03 2.106189e+04 1858.173
```

```
c  3438.620 0.000000e+00 2.092466e-02 16829.040
g 21061.890 2.092466e-02 0.000000e+00     1.000
t  1858.173 1.682904e+04 1.000000e+00     0.000

Base frequencies:
0.2776524 0.3540032 0.1308949 0.2374494
```

Recuerde que en MV hay dos tipos de parámetros que deben estimarse: (1) aquellos puramente numéricos (tasas de sustitución, longitudes de las ramas) y (2) la topología del árbol. Para el primer tipo de parámetro se utilizan métodos numéricos relativamente rápidos, mientras que para la topología se utilizan métodos heurísticos más laboriosos que exploran el espacio de árboles (Paradis 2012). En este ejemplo podemos ver que las transiciones (A ↔ G, C ↔ T) son mucho más frecuentes que las transversiones, lo cual es consistente con lo esperado en términos biológicos.

Al igual que con parsimonia, se puede graficar simultáneamente el árbol junto con los caracteres. Primero, reconstruimos los estados ancestrales con la función `ancestral.pml()`.

```
> anc.ML <- ancestral.pml(ML.GTR, type = "ml")
```

Simultáneamente, podemos aplicar *bootstrap* para evaluar el soporte de las ramas, aunque en este caso utilizamos la función `bootstrap.pml()` por haber utilizado MV. Especificamos el número de réplicas *bootstrap* con el argumento `bs`.

```
> arboles.boot.ml <- bootstrap.pml(ML.GTR, bs = 1000)
> soportes <- prop.clades(phy = ML.GTR$tree, arboles.boot.ml)
```

Así, graficamos la filogenia resultante con las longitudes y soporte de las ramas. A fines comparativos mapeamos los sitios 5 y 11 (Fig. 8.23).

```
> plotAnc(tree = ML.GTR, anc.ML, i = 5,
+         col = c("black", "white", "gray", "white"))
> nodelabels(soportes/10, adj = c(1.1, -0.3), frame = "none")
> plotAnc(tree = ML.GTR, anc.ML, 11,
+         col = c("black", "white", "gray", "white"))
> nodelabels(soportes/10, adj = c(1.1, -0.3), frame = "none")
```

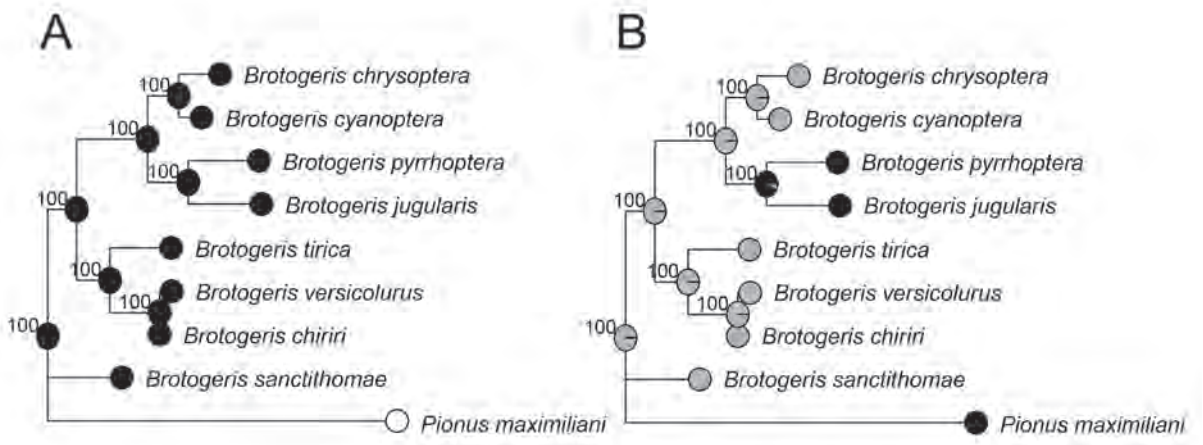


Fig. 8.23. Filogenia obtenida por MV (modelo *General Time Reversible*) sobre la base de secuencias de ADN de nueve especies de loros. Los círculos representan gráficos de torta donde se muestra la probabilidad de cada base nitrogenada (negro: adenina, blanco: citosina, gris: guanina). Se muestran los soportes de las ramas obtenidos por 1000 pseudo-réplicas *bootstrap*. (A) mapeo del sitio 5; (B) mapeo del sitio 11.

EPÍLOGO
LA COMPLEJIDAD: UN SIGNO DE NUESTRO TIEMPO

En este libro hemos presentado una variedad de métodos del análisis multivariado utilizados en la Biología contemporánea. Nuestro objetivo ha sido transmitir nuestra certidumbre de que el análisis multivariado es una herramienta indispensable en la tarea del biólogo de nuestro tiempo.

Si se nos pidiese caracterizar a nuestro tiempo en sólo una palabra estaríamos tentados por decir "complejidad".

Las relaciones biológicas son vistas como sistemas altamente complejos, que con cada avance tecnológico en el estudio de ellas, adquieren ante el investigador aún mayor complejidad. Esto implica la necesidad de considerar simultáneamente numerosas variables para desentrañar esas relaciones.

El análisis multivariado es la disciplina que permite analizar cuantitativamente las relaciones biológicas, utilizando numerosas variables al mismo tiempo.

La gran cantidad de información acumulada en las bases de datos y el incesante progreso de la tecnología computacional, favorecen el uso del análisis multivariado.

Los análisis filogenéticos, tan útiles y difundidos en nuestro tiempo, son considerados en este libro como parte del universo de espacios multidimensionales que el análisis multivariado genera. Por otro lado, un libro sobre el análisis multivariado no puede prescindir de poner a los métodos dentro del contexto del, muy utilizado mundialmente, programa computacional R. Por ello a cada método presentado en este libro se lo coloca en ese contexto.

Con nuestro libro intentamos responder dos preguntas que a menudo los investigadores se formulan: ¿qué es el análisis multivariado? y ¿qué puede el análisis multivariado hacer por mí? No sabemos si hemos logrado nuestro objetivo, pero guardamos la esperanza que, cuando menos, hayamos despertado el interés del lector por el análisis multivariado como una herramienta útil en sus investigaciones.

Hay dos consideraciones que hemos tratado de transmitir implícitamente a través de los ejemplos presentados en el libro, pero que deseamos explicitar en este epílogo:

- el análisis multivariado no es un sucedáneo de la creatividad, sino un estímulo para generarla; y
- el análisis multivariado no implica simplemente la habilidad de calcular, sino algo más importante: la capacidad de razonar cuantitativamente.

No es un buen síntoma de las instituciones o de los individuos tratar de vaticinar el futuro, pues estamos inmersos en un intenso y agitado presente que requiere creatividad, imaginación, talento, inteligencia y esfuerzo para enfrentarlo. El análisis multivariado ha sido y es un aporte singular para que, con esos requerimientos, el biólogo moderno avance en la comprensión de la complejidad de las relaciones biológicas.

REFERENCIAS BIBLIOGRÁFICAS

Abdi H, LJ Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4): 433-459.

Adachi J, M Hasegawa. 1996. *MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood (No. 28).* Institute of Statistical Mathematics. Tokio.

Adams EN III. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Biology*, 21(4), 390-397.

Aggarwal CC, A Hinneburg, DA Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. En Van den Bussche J, V Vianu (eds.) *International conference on database theory.* Pp. 420-434. Springer. Berlín.

Agnarsson I, JA Miller. 2008. Is ACCTRAN better than DELTRAN? *Cladistics*, 24(6): 1032-1038.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. En Petrov, BN, F Caski (eds.) *Proceedings of the second international symposium on information theory*. Pp. 267-281. Akademiai Kiado. Budapest.

Aldrich J. 1997. RA Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3): 162-176.

Alfaro ME, S Zoller, F Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, 20(2): 255-266.

Ameghino F. 1884. *Filogenia*. Ediciones Anaconda. Buenos Aires.

Apodaca MJ, L Katinas, E Guerrero. 2019a. Hidden areas of endemism: small units in the southeastern Neotropics. *Systematic Biology*, 17(5): 425-438.

Apodaca MJ, JD McInerney, OE Sala, L Katinas, JV Crisci. 2019b. A concept map of evolutionary biology to promote meaningful learning in biology. *The American Biology Teacher*, 81(2): 79-87.

Arendt J, D Reznick. 2007. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology and Evolution*, 23(1): 26-32.

Baker AN, AN Smith, FB Pichler. 2002. Geographical variation in Hector's dolphin: recognition of new subspecies of *Cephalorhynchus hectori*. *Journal of the Royal Society of New Zealand*, 32(4): 713-727.

Ball GH. 1965. Data analysis in the social sciences: what about the details? *Proceedings, Fall Joint Computer Conference*, 533-559.

Ball GH, DJ Hall. 1965. *ISODATA, a novel method of data analysis and pattern classification*. Stanford research institute Menlo Park. California.

Battey CJ, EB Linck, KL Epperly, C French, DL Slager, Jr PW Sykes, J Klicka. 2018. A migratory divide in the Painted Bunting (*Passerina ciris*). *American Naturalist*, 191(2): 259-268.

Bayes T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370-418.

Beaton D, CRC Fatt, H Abdi. 2014. An exposition of multivariate analysis with the singular value decomposition in R. *Computational Statistics and Data Analysis*, 72: 176-189.

Bellman R. 1957. *Dynamic programming*. Princeton University Press. Nueva Jersey.

Bembom O. 2018. seqLogo: sequence logos for DNA sequence alignments. Paquete de R versión 1.48.0. https://rdrr.io/bioc/seqLogo/

Benzécri JP. 1969. Statistical analysis as a tool to make patterns emerge from data. En Watanabe S. (ed.) *Methodologies of pattern recognition.* Pp. 35-74. Academic Press. Cambridge.

Benzécri JP. 1980. Introduction a l'analyse des correspondances d'apres un exemple de données medicales. *Les cahiers de l'analyse des données*, 5: 283-310.

Bertalanffy L. 1987. *Teoría general de sistemas*. Fondo de Cultura Económica. Madrid.

Blackith RE, RA Reyment. 1971. *Multivariate morphometrics*. Whitefriars Press. Londres.

Bookstein FL. 2017. A newly noticed formula enforces fundamental limits on geometric morphometric analyses. *Evolutionary Biology*, 44(4): 522-541.

Boulesteix AL. 2005. A note on between-group PCA. *International Journal of Pure and Applied Mathematics*, 19: 359-366.

Bouveyron C, S Girard, C Schmid. 2007a. High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52(1): 502-519.

Bouveyron C, S Girard, C Schmid. 2007b. High-dimensional discriminant analysis. *Communications in Statistics. Theory and Methods*, 36(14): 2607-2623.

Box GE, GC Tiao. 2011. *Bayesian inference in statistical analysis*. John Wiley and Sons. Nueva York.

Bray JR, JT Curtis.1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4): 325-349.

Bulut H. 2019. MVTests: multivariate hypothesis tests. Paquete de R versión 1.1. https://CRAN.R-project.org/package=MVTests

Bunge M. 1969. *La investigación científica*. Ediciones Ariel. Barcelona.

Burnham K, DR Anderson. 2004. *Model selection and multi-model inference*. Springer-Verlag. Nueva York.

Burns P. 2012. *The R inferno*. Lulu Press. Carolina del Norte.

Cailliez F. 1983. The analytical solution of the additive constant problem. *Psychometrika*, 48(2): 305-308.

Cain AJ, GA Harrison. 1958. An analysis of the taxonomist's judgement of affinity. *Proceedings of the Zoological Society of London*, 131: 85.

Camin JH, RR. Sokal. 1965. A method for deducing branching sequences in phylogeny. *Evolution*, 19: 311-326.

Cao Y, WP Williams, AW Bark. 1997. Similarity measure bias in river benthic Aufwuchs community analysis. *Water Environment Research*, 69(1): 95-106.

Cardini A, P O'Higgins, FJ Rohlf. 2019. Seeing distinct groups where there are none: spurious patterns from between-group PCA. *Evolutionary Biology*, 46(4): 303-316.

Cattell RB. 1952. *Factor analysis: an introduction and manual for the psychologist and social scientist*. Harper. Oxford.

Cattell RB. 1966. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2): 245-276.

Cattell RB, S Vogelmann. 1977. A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research,* 12(3): 289-325.

Cavalli-Sforza LL, AW Edwards. 1967. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3): 550-570.

Cawley GC, NL Talbot. 2003. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11): 2585-2592.

Charrad M, N Ghazzali, V Boiteau, A Niknafs, MM Charrad. 2014. Package 'NbClust'. *Journal of Statistical Software*, 61: 1-36.

Chen MH, L Kuo, PO Lewis. 2014. *Bayesian phylogenetics: methods, algorithms, and applications.* Chapman and Hall/CRC. Boca Raton.

Cheng Q, J Han. 2004. Morphological variations and discriminant analysis of two populations of *Coilia ectenes*. *Journal of Lake Science*, 16(4): 356-364.

Chessel D, AB Dufour, J Thioulouse. 2004. The ade4 package-I-One-table methods. *R news*, 4(1): 5-10.

Choi SS, SH Cha, CC Tappert. 2010. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1): 43-48.

Cigliano MM, MS Fernández, AA Lanteri. 2005. Métodos cuantitativos. En Lanteri AA, MM Cigliano (eds.) *Sistemática biológica: fundamentos teóricos y ejercitaciones.* Pp. 137-153. Editorial de la Universidad Nacional de La Plata. La Plata.

Clarke KR. 1993. Nonparametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, 18: 117-143.

Cliff N. 1988. The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, 103(2): 276.

Clifford HT, W Stephenson. 1975. *An introduction to numerical classification*. Academic Press. Nueva York.

Cohen M, E Nagel. 1971. *Introducción a la lógica y al método científico*. Amorrortu Editores. Buenos Aires.

Cormack RM. 1971. A review of classification. *Journal of the Royal Statistical Society Series A (General)*: 321-367.

Cottam G, FG Goff, RH Whittaker. 1978. Wisconsin comparative ordination. En Whittaker RH (ed.) *Ordination of plant communities*. Pp. 185-213. Springer. Dordrecht.

Cramér H. 1946. *Mathematical of statistics*. Princeton University Press. Princeton.

Crawley MJ. 2012. *The R book*. John Wiley and Sons. Hoboken.

Crisci JV. 1974. A numerical-taxonomic study of the subtribe Nassauviinae (Compositae, Mutisieae). *Journal of the Arnold Arboretum*, 55(4): 568-610.

Crisci JV, MJ Apodaca, L Katinas. 2019. El fin de la botánica. *Revista del Museo de La Plata*, 4(1): 41-50.

Crisci JV, JH Hunziker, RA Palacios, CA Naranjo. 1979. A numerical-taxonomic study of the genus *Bulnesia* (Zygophyllaceae): cluster analysis, ordination and simulation of evolutionary trees. *American Journal of Botany*, 66(2): 133-140.

Crisci JV, MF López Armengol. 1983. *Introducción a la teoría y práctica de la taxonomía numérica.* Organización de los Estados Americanos. Washington.

Crisci JV, TF Stuessy. 1980. Determining primitive character states for phylogenetic reconstruction. *Systematic Botany*, 5(2):112-135.

Culhane AC, G Perriere, EC Considine, TG Cotter, DG Higgins. 2002. Between-group analysis of microarray data. *Bioinformatics*, 18(12): 1600-1608.

Czekanowski J. 1909. *Zur differentialdiagnose der neandertalgruppe*. Friedrich Vieweg and Sohn. Berlín.

da Silva TE, DFR Alves, ACA Barros-Alves, FG Taddei, A Fransozo. 2018. Morphometric differences between two exotic invasive freshwater caridean species (genus *Macrobrachium*). *Invertebrate Reproduction and Development*. doi: 0.1080/07924259.2018.1505668

Darwin C. 1859. *On the origin of species*. Murray. Londres.

de Pinna MC. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics*, 7: 367-394.

Dechaume-Moncharmont FX, K Monceau, F Cezilly. 2011. Sexing birds using discriminant function analysis: a critical appraisal. *Auk*, 128(1): 78-86.

Dice LR. 1945. Measures of the amount of ecological association between species. *Ecology*, 26: 297-302.

Dollo L. 1893. Les lois de l'evolution. *Bulletin de la Societé Belge de Geologié, Paléontologie et Hydrologie*, 7: 164-166.

Donoho DL. 2000. High-dimensional data analysis: the curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1(2000): 1-33.

D'Orazio M. 2019. StatMatch: statistical matching or data fusion. Paquete de R versión 1.3.0. https://CRAN.R-project.org/package=StatMatch

Douady CJ, F Delsuc, Y Boucher, WF Doolittle, EJ Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, 20(2): 248-254.

Drummond AJ, A Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1), 214.

Duncan T. 1984. Willi Hennig, character compatibility, Wagner parsimony, and the "Dendrogrammaceae" revisited. *Taxon*, 33(4): 698-704.

du Toit SHC, AGW Steyn, RH Stumpf. 1986. *Graphical exploratory data analysis*. Springer. Nueva York.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution: International Journal of Organic Evolution*, 63(1): 1-19.

Edwards AW, LL Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. Phenetic and phylogenetic classification. *Systematic Association Publication*, 67-76.

Efron B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7: 1-26.

Efron B. 2013. Bayes' theorem in the 21st century. *Science*, 340(6137): 1177-1178.

Erixon P, Svennblad B, Britton T, B Oxelman. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, 52(5): 665-673.

Evans GA. 2000. Designer science and the "omic" revolution. *Nature Biotechnology*, 18(2): 127.

Farris JS. 1969a. On the cophenetic correlation coefficient. *Systematic Zoology*, 18(3): 279-285.

Farris JS. 1969b. A successive approximations approach to character weighting. *Systematic Zoology*, 18(4): 374-385.

Farris JS. 1970. Method for computing Wagner trees. *Systematic Zoology*, 19(1): 83-92.

Farris JS. 1977. Phylogenetic analysis under Dollo's law. *Systematic Zoology*, 26(1): 77-88.

Farris JS. 1983. The logical basis of phylogenetic analysis. En Platnick NI, VA Funk (eds.) *Advances in cladistics*, Vol. 2. Pp 7-36. Columbia University Press. Nueva York.

Farris JS. 1985. Distance data revisited. *Cladistics*, 1(1): 67-86.

Farris JS. 1989. The retention index and the rescaled consistency index. *Cladistics*, 5(4): 417-419.

Farris JS, AG Kluge. 1985. Parsimony, synapomorphy, and explanatory power: a reply to Duncan. *Taxon*, 34(1): 130-135.

Farris JS, AG Kluge, MJ Eckardt. 1970. A numerical approach to phylogenetic systematics. *Systematic Zoology*, 19(2): 172-189.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17: 368-376.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the boostrap. *Evolution*, 39(4): 783-791.

Felsenstein J. 1993. PHYLIP (phylogeny inference package), version 3.5. http://evolution.genetics.washington.edu/phylip.html

Felsenstein J. 2004. *Inferring phylogenies*. Sinauer associates. Sunderland.

Fernández MS, MM Cigliano, AA Lanteri. 2005. Sistemática filogenética: argumentación hennigiana. En Lanteri AA, MM Cigliano (eds.) *Sistemática biológica: fundamentos teóricos y ejercitaciones.* Pp. 123-134. Editorial de la Universidad Nacional de La Plata. La Plata.

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179-188.

Fisher RA. 1940. The precision of discriminant functions. *Annals of Eugenics,* 10: 422-429.

Fisher RA. 1958. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284): 789-798.

Fitch WM. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. *Systematic Biology*, 20(4): 406-416.

Fitch WM, E Margoliash. 1967. Construction of phylogenetic trees. *Science*, 155(3760): 279-284.

Forey PL, CJ Humphries, IJ Kitching, RW Scotland, DJ Siebert, DM Williams. 1992. *Cladistics. A practical course in systematics*. Oxford University Press. Nueva York.

Fraley C, AE Raftery. 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41(8): 578-588.

Frontier S. 1976. Decrease of eigenvalues in principal component analysis-comparison with broken stick model. *Journal of Experimental Marine Biology and Ecology*, 25(1): 67-75.

Fuchs DV, VS Berríos, D Montalti. 2017. Morphometric differences between sexes in the white-faced ibis (*Plegadis chihi*). *Wilson Journal of Ornithology*, 129(2): 317-322.

Futuyma D, M Kirkpatrick. 2017. *Evolution*. Sinauer. Sunderland.

Gagné SA, R Proulx. 2009. Accurate delineation of biogeographical regions depends on the use of an appropriate distance measure. *Journal of Biogeography*, 36(3): 561-562.

Galton F. 1877. Typical laws of heredity. *Proceedings of the Royal Institution*, 8: 282-301.

Gatto L, D Catanzaro, MC Milinkovitch. 2006. Assessing the applicability of the GTR nucleotide substitution model through simulations. *Evolutionary Bioinformatics*, 2: 145-155.

Gauch Jr HG. 1982. *Multivariate analysis in community ecology*. Cambridge University Press. Cambridge.

Gelman A, DB Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4): 457-472.

Goloboff PA. 1998. *Principios básicos de cladística*. Sociedad Argentina de Botánica. Buenos Aires.

Goloboff PA, JS Farris, KC Nixon. 2008. TNT, a free program for phylogenetic analysis. *Cladistics*, 24(5): 774-786.

Goloboff PA, M Pittman, D Pol, X Xu. 2018. Morphological data sets fit a common mechanism much more poorly than DNA sequences and call into question the Mkv model. *Systematic Biology*, 68(3): 494-504.

Goslee SC, DL Urban. 2007. The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22(7): 1-19.

Gower JC. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53 (3/4): 325-338.

Gower JC. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.

Gower JC. 1982. Euclidean distance geometry. *Mathematical Scientist*, 7(1): 1-14.

Graham RL, LR Foulds. 1982. Unlikelihoods that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Mathematical Biosciences*, 60: 133-142.

Green PT, CM Theobald. 1989. Sexing birds by discriminant analysis: further considerations. *Ibis*, 131(3): 442-447.

Greenacre M. 2008. *La práctica del análisis de correspondencias*. Fundación BBVA. Madrid.

Greenacre M, R Primicerio 2014. *Multivariate analysis of ecological data*. Fundación BBVA. Madrid.

Gu Z, R Eils, M Schlesner. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics,* 32(18): 2847-2849.

Guindon S, O Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), 696-704.

Guttman L. 1954. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2): 149-161.

Haeckel E. 1866. *Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Descendenz-Theorie.* Reimer. Berlín.

Hall BK. 2007. Homoplasy and homology: dichotomy or continuum? *Journal of Human Evolution*, 52(5): 473-479.

Hallgrímsson B, BK Hall. 2005. Variation and variability: central concepts in biology. En Hallgrímsson B, BK Hall (eds.) *Variation*. Pp. 1-7. Academic Press. Boston.

Hamann U. 1961. Merkmalsbestand und verwandtschaftsbeziehungen der farinosae: ein beitrag zum system der monokotyledonen. *Willdenowia*, 639-768.

Hardy MA. 1993. *Regression with dummy variables*. Sage. Newbury Park.

Hartigan JA. 1975. *Clustering algorithms*. John Wiley and Sons. Nueva York.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57: 97-109.

Heikinheimo H, Fortelius M, Eronen J, H Mannila. 2007. Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, 34(6): 1053-1064.

Hendy MD, D Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, 59(2): 277-290.

Hennig C. 2018. fpc: flexible procedures for clustering. Paquete de R versión 2.1-11.1. https://CRAN.R-project.org/package=fpc

Hennig W. 1950. *Grundzuge einer theorie der phylogenetischen tystematik*. Zentralverlag. Berlín.

Hennig W. 1968. *Elementos de una sistemática filogenética*. EUDEBA. Buenos Aires.

Herrera CM. 2009. *Multiplicity in unity: plant subindividual variation and interactions with animals*. University of Chicago Press. Chicago.

Herrera CM. 2017. The ecology of subindividual variability in plants: patterns, processes, and prospects. *Web Ecology*, 17(2): 51-64.

Higgins D, P Lemey. 2009. Multiple sequence alignment. En Lemey P, M Salemi, AM Vandamme (eds.) *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Pp. 68-108. Cambridge University Press. Cambridge.

Hirschfeld HO. 1935. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4): 520-524.

Højsgaard S, U Halekoh. 2018. doBy: groupwise statistics, LSmeans, linear contrasts, utilities. Paquete de R versión 4.6-2. https://CRAN.R-project.org/package=doBy

Holder M, PO Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, 4(4): 275.

Holmes S. 2003. Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, 18(2): 241-255.

Horn HS. 1966. Measurement of "overlap" in comparative ecological studies. *American Naturalist*, 100(914): 419-424.

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6): 417.

Hubálek Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews*, 57(4): 669-689.

Huelsenbeck JP, KA Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28(1): 437-466.

Huelsenbeck JP, F Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8): 754-755.

Huelsenbeck JP, F Ronquist, R Nielsen, JP Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550): 2310-2314.

Husson F, S Lê, J Pagès. 2017. *Exploratory multivariate analysis by example using R*. and Hall/CRC. Nueva York.

Ihaka R, R Gentleman. 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3): 299-314.

Indykiewicz P, P Minias, J Kowalski, P Podlaszczuk. 2019. Shortcomings of discriminant functions: a case study of sex identification in the Black-Headed Gull. *Ardeola*, 66(2): 361-372.

Jaccard P. 1900. Contribution au problème de l'immigration post-glaciare de la flore alpine. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 36: 87-130.

Jackson DA. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8): 2204-2214.

Jain AK. 2010. Data clustering: 50 years beyond *K*-means. *Pattern Recognition Letters*, 31(8): 651-666.

Jain AK, RC Dubes. 1988. *Algorithms for clustering data*. Prentice Hall. Nueva Jersey.

Jain AK, MN Murty, PJ Flynn. 1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3): 264-323.

James FC, CE McCulloch. 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annual Review of Ecology and Systematics*, 21(1): 129-166.

James G, D Witten, T Hastie, R Tibshirani. 2013. *An introduction to statistical learning.* Springer. Nueva York.

Jeffreys H 1935. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, 31: 203-222.

Jolliffe IT. 2002. *Principal component analysis*. Springer. Nueva York.

Jolliffe IT, J Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065): 20150202.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24: 1403-1405.

Jukes TH, CR Cantor. 1969. Evolution of protein molecules. En Munro HN (ed.) *Mammalian protein metabolism*. Pp. 21-123. Academic Press. Nueva York.

Kaiser HF. 1961. A note on Guttman's lower bound for the number of common factors. *British Journal of Statistical Psychology*, 14(1): 1-2.

Kane A. 2012. Determining the number of clusters for a k-means clustering algortihm. *Indian Journal of Computer Science and Engineering*, 3(5): 670-672.

Kass RE, AE Raftery. 1995. Bayes factors. *Journal of the American Statistical Association*, 90(430): 773-795.

Kassambara A. 2017a. *Practical guide to cluster analysis in R: unsupervised machine learning*. STHDA. Marsella.

Kassambara A. 2017b. *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*. STHDA. Marsella.

Kassambara A, F Mundt. 2017. factoextra: extract and visualize the results of multivariate data analyses. Paquete de R versión 1.0.5. https://CRAN.R-project.org/package=factoextra

Kaufman L, PJ Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons. Hoboken.

Kelchner SA, MA Thomas. 2007. Model use in phylogenetics: nine key questions. *Trends in Ecology and Evolution*, 22(2): 87-94.

Kemsley EK. 1996. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems*, 33(1): 47-61.

Kendall DG. 1971. Seriation from abundance matrices. En Hodson FR, DG Kendal, P Tautu (eds.) *Mathematics in the archaeological and historical sciences*. Pp. 215-252. Edinburgh University Press. Edimburgo.

Kenkel NC, L Orlóci. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. *Ecology*, 67(4): 919-928.

Kluge AG, JS Farris. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Biology*, 18(1): 1-32.

Kneller GF. 1978. *Science as a human endeavor*. Columbia University Press. Nueva York.

Kodinariya TM, PR Makwana. 2013. Review on determining number of cluster in K-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6): 90-95.

Koonin EV, KS Makarova, L Aravind. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Reviews in Microbiology*, 55(1): 709-742.

Krebs CJ. 1999. *Ecological methodology*. Harper and Row. Nueva York.

Kreft H, W Jetz. 2010. A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, 37(11): 2029-2053.

Kruskal JB. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1): 1-27.

Kruskal JB. 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2): 115-129.

Kruskal JB. 1977. Multidimensional scaling and other methods for discovering structure. En Enslein K, A Ralston, HS Wilf (eds.) *Statistical methods for digital computers*. Pp. 296-339. Wiley. Nueva York.

Kruskal JB, M Wish. 1978. *Multidimensional scaling, quantitative applications in the social sciences*. Sage Publications. Beverly Hills.

Kuhn M. 2019. caret: classification and regression training. Paquete de R versión 6.0-83. https://CRAN.R-project.org/package=caret

Kuiper FK, L Fisher. 1975. 391: a Monte Carlo comparison of six clustering procedures. *Biometrics*, 31: 777-783.

Kumar S, K Tamura, M Nei. 1993. MEGA - Molecular evolutionary genetics analysis software for microcomputers. *Computer Applications in the Biosciences*, 10:189-191.

Kulczynski S. 1928. Die Pflanzenassoziationen der Pieninen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles*, 1927: 57-203.

Laliberté E, P Legendre, B Shipley. 2014. FD: measuring functional diversity from multiple traits, and other tools for functional ecology. Paquete de R versión 1.0-12. https://cran.r-project.org/package=FD.

Lance GN, WT Williams. 1967. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Computer Journal*, 9(4): 373-380.

Lankester ER. 1870. II. On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Annals and Magazine of Natural History*, 6(31): 34-43.

Larget B, DL Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16(6): 750-759.

Lê S, J Josse, F Husson. 2008. FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1-18.

Legendre P, L Legendre. 1998. *Numerical ecology: developments in environmental modelling*. Elsevier. Amsterdam.

Legendre P. 1990. Quantitative methods and biogeographic analysis. En Garbary DJ, GR South (eds.) *Evolutionary biogeography of the marine algae of the North Atlantic*. Pp. 9-34. Springer. Berlín.

Leisch F, K Hornik, BD Ripley. 2017. mda: Mixture and Flexible Discriminant Analysis. Paquete de R versión 0.4-10. https://CRAN.R-project.org/package=mda

Lemey P, M Salemi, AM Vandamme. 2009. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press. Cambridge.

Lemmon AR, EC Moriarty. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology*, 53(2): 265-277.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6): 913-925.

Li H. 2015. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and its Application*, 2: 73-94.

Li S, DK Pearl, H Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, 95(450): 493-508.

Li WH, D Graur. 1991. *Fundamentals of molecular evolution*. Sinauer. Sunderland.

Lingoes JC. 1971. Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, 36(2): 195–203.

Lloyd SP. 1982. Least squares quantization in PCM. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 28: 129-137.

Logan M. 2011. *Biostatistical design and analysis using R: a practical guide*. John Wiley and Sons. Nueva Jersey.

MacArthur RH. 1957. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences*, 43(3): 293-295.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14): 281-297.

Maechler M, P Rousseeuw, A Struyf, M Hubert, K Hornik. 2018. Cluster: cluster analysis basics and extensions. Paquete de R versión 2.0.7-1. https://cran.r-project.org/package=cluster

Mahalanobis PC. 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 12(1936): 49-55.

Manel S, JM Dias, SJ Ormerod. 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, 120(2-3): 337-347.

Marchenko VA, LA Pastur. 1967. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4): 457-483.

Marramà G, J Kriwet. 2017. Principal component and discriminant analyses as powerful tools to support taxonomic identification and their use for functional and phylogenetic signal detection of isolated fossil shark teeth. *PloS ONE*, 12(11): e0188806.

Matloff N. 2011. *The art of R programming: a tour of statistical software design*. No Starch Press. San Francisco.

Mau B, MA Newton. 1997. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 6(1): 122-131.

Mayr E. 1969. *Principles of systematic zoology*. McGraw-Hill. Nueva York.

Mehta T, M Tanik, DB Allison. 2004. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature Genetics*, 36(9): 943-947.

Metropolis N, AW Rosenbluth, MN Rosenbluth, AH Teller, E Teller. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6): 1087-1092.

Meyer D, C Buchta. 2019. proxy: distance and similarity measures. Paquete de R versión 0.4-23. https://CRAN.R-project.org/package=proxy

Michener CD, RR Sokal. 1957. A quantitative approach to a problem in classification. *Evolution*, 11(2): 130-162.

Minchin PR. 1987. An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, 69: 89-107.

Miyamoto MM. 1985. Consensus cladograms and general classifications. *Cladistics*, 1(2): 186-189.

Moline PM, HP Linder. 2006. Input data, analytical methods and biogeography of *Elegia* (Restionaceae). *Journal of Biogeography*, 33(1): 47-62.

Montalti D, M Graña Grilli, RE Maragliano, G Cassini. 2012. The reliability of morphometric discriminant functions in determining the sex of Chilean flamingos *Phoenicopterus chilensis*. *Current Zoology*, 58(6): 851-855.

Morgan M. 2018. BiocManager: access the Bioconductor project package repository. Paquete de R versión 1.30.4. https://CRAN.R-project.org/package=BiocManager

Mueller LD, FJ Ayala. 1982. Estimation and interpretation of genetic distance in empirical studies. *Genetical Research*, 40: 127-137.

Müller F. 1864. *Für Darwin*. Wilhelm Engelmann. Leipzig.

Müller K, H Wickham. 2019. tibble: simple data frames. Paquete de R versión 2.1.3. https://CRAN.R-project.org/package=tibble

Murtagh F, P Legendre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, 31(3): 274-295.

Nascimento FF, M dos Reis, Z Yang. 2017. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology and Evolution*, 1(10): 1446.

Nelson GJ, NI Platnick. 1981. *Systematics and biogeography*. Columbia University Press. Nueva York.

Nenadic O, M Greenacre. 2007. Correspondence analysis in R, with two-and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3): 1-13.

Neyman J. 1971. Molecular studies of evolution: a source of novel statistical problems. En Gupta SS, J Yackel (eds.) *Statistical decision theory and related topics*. Pp. 1-27. Academic Press. Nueva York.

Nixon KC. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 19: 407-414.

Normark BB, AA Lanteri. 1998. Incongruence between morphological and mitochondrial-DNA characters suggests hybrid origins of parthenogenetic weevil lineages (genus *Aramigus*). *Systematic Biology*, 47(3): 475-494.

Novara LJ. 2012. Zygophyllaceae R.Br. Flora del Valle de Lerma. Aportes Botánicos de Salta. Vol. I N° 17. Facultad de Ciencias Naturales. Universidad Nacional de Salta.

Odum EP. 1950. Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology*, 31(4): 587-605.

Oksanen JF, G Blanchet, M Friendly, R Kindt, P Legendre, D McGlinn, PR Minchin, RB O'Hara, GL Simpson, P Solymos, MHH Stevens, E Szoecs, H Wagner. 2018. vegan: community ecology package. Paquete de R versión 2.5-3. https://CRAN.R-project.org/package=vegan

Olden JD, DA Jackson. 2002. A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology*, 47(10): 1976-1995.

Ordano M, J Fornoni, K Boege, CA Domínguez. 2008. The adaptive value of phenotypic floral integration. *New Phytologist*, 179(4): 1183-1192.

Orlóci L. 1975. *Multivariate analysis in vegetation research*. Dr. W.J. Junk Publishers. La Haya.

Owen R. 1843. *Lectures on the comparative anatomy and physiology of the invertebrate animals*. Longman. Londres.

Page RD. 1996. On consensus, confidence, and "total evidence". *Cladistics*, 12(1): 83-92.

Palacio FX, D Montalti. 2013. Seasonal variation and effect of non-native invasive vegetation on two bird communities in northeast of Buenos Aires province, Argentina. *Ornitología Neotropical*, 24: 157-168.

Palacio FX, JM Girini, M Ordano. 2017. Linking the hierarchical decision-making process of fruit choice and the phenotypic selection strength on fruit traits by birds. *Journal of Plant Ecology*, 10(4): 713:720.

Palacio FX, M Lacoretz, M Ordano. 2014. Bird-mediated selection on fruit display traits in *Celtis ehrenbergiana* (Cannabaceae). *Evolutionary Ecology Research*, 16(1): 51-62.

Palacios RA, Hunziker JH. 1984. Revisión taxonómica del género *Bulnesia* (Zygophyllaceae). *Darwiniana*, 25(1-4): 299-320.

Paliy O, V Shankar. 2016. Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, 25(5): 1032-1057.

Paradis E. 2002. *R for beginners*. https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

Paradis E. 2012. *Analysis of phylogenetics and evolution with R*. Springer. Nueva York.

Paradis E, J Claude, K Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20: 289-290.

Paradis E, K Schliep. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3): 526-528.

Patterson C. 1988. Homology in classical and molecular biology. *Molecular Biology Evolution*, 5(6): 603-625.

Pawlowsky-Glahn V, A Buccianti. 2011. *Compositional data analysis*. John Wiley and Sons. Londres.

Pearson K. 1895. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352): 240-242.

Pearson K. 1900. I. Mathematical contributions to the theory of evolution-VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London*, 195(262-273): 1-47.

Pearson K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559-572.

Peres-Neto PR, DA Jackson, KM Somers. 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology*, 84(9): 2347-2363.

Peres-Neto PR, DA Jackson, KM Somers. 2005. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics Data Analysis*, 49(4): 974-997.

Pigliucci M. 2003. Phenotypic integration: studying the ecology and evolution of complex phenotypes. *Ecology Letters*, 6(3): 265-272.

Pigliucci M, K Preston. 2004. *Phenotypic integration: studying the ecology and evolution of complex phenotypes*. Oxford University Press. Oxford.

Piro A, DV Fuchs, D Montalti. 2018. Morphometric differences between sexes of two subspecies of Black-crowned Night-Heron (*Nycticorax nycticorax*) using discriminant function analysis. *Waterbirds*, 41(1), 87-93.

Podani J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon*, 48(2): 331-340.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, 25(7): 1253-1256.

Posada D, TR Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5): 793-808.

Posada D, KA Crandall. 2001. Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, 50(4): 580-601.

Press WH, SA Teukolsky, WT Vetterling, BP Flannery. 2007. *Numerical recipes: the art of scientific computing*. Cambridge University Press. Cambridge.

Prim RC. 1957. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6): 1389-1401.

Qiao Z, L Zhou, JZ Huang. 2009. Sparse linear discriminant analysis with applications to high dimensional low sample size data. *International Journal of Applied Mathematics*, 39(1): 48-50.

Quackenbush J. 2007. Extracting biology from high-dimensional biological data. *Journal of Experimental Biology*, 210(9): 1507-1517.

Quenouille MH. 1956. Notes on bias in estimation. *Biometrika*, 43(3/4): 353-360.

Quinn GP, MJ Keough. 2002. *Experimental design and data analysis for biologists*. Cambridge University Press. Cambridge.

R Core Team. 2018. R: a language and environment for statistical computing. https://www.r-project.org/

RStudio. 2012. RStudio: integrated development environment for R. http://www.rstudio.org/

Rannala B, Z Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43(3): 304-311.

Rannala B, Z Yang. 2008. Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics*, 9: 217-231.

Rannala B, T Zhu, Z Yang. 2011. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Molecular Biology and Evolution*, 29(1): 325-335.

Rao CR. 1948. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society*. *Series B (Methodological)*, 10(2): 159-203.

Rao CR. 1952. *Advanced statistical methods in biometric research*. John Wiley and Sons. Nueva York.

Ribas CC, CY Miyaki, J Cracraft. 2009. Phylogenetic relationships, diversification and biogeography in Neotropical *Brotogeris* parakeets. *Journal of Biogeography*, 36(9): 1712-1729.

Richman MB. 1988. A cautionary note concerning a commonly applied eigenanalysis procedure. *Tellus*, 40(1): 50-58.

Roberts DW. 2016. labdsv: ordination and multivariate analysis for ecology. Paquete de R versión 1.8-0. https://CRAN.R-project.org/package=labdsv

Robidoux S, S Pritchard. 2014. Hierarchical clustering analysis of reading aloud data: a new technique for evaluating the performance of computational models. *Frontiers in Psychology*, 5: 1-7.

Roch S. 2010. Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science*, 327(5971): 1376-1379.

Rogers DJ, TT Tanimoto. 1960. A computer program for classifying plants. *Science*, 132(3434): 1115-1118.

Rohlf FJ. 1970. Adaptive hierarchical clustering schemes. *Systematic Biology*, 19(1): 58-82.

Rohlf FJ. 1972. An empirical comparison of three ordination techniques in numerical taxonomy. *Systematic Zoology*, 21(3): 271-280.

Rohlf FJ, LF Marcus. 1993. A revolution morphometrics. *Trends in Ecology and Evolution*, 8(4): 129-132.

Romesburg HC. 1984. *Cluster analysis for researchers*. Life Time Learning Publication. Belmont.

Ronquist F. 2004. Bayesian inference of character evolution. *Trends in Ecology and Evolution*, 19(9): 475-481.

Ronquist F, P van der Mark, JP Huelsenbeck. 2009. Bayesian phylogenetic analysis using Mr BAYES. En Lemey P, M Salemi, AM Vandamme (eds.) *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Pp. 210-266. Cambridge University Press. Cambridge.

Rosa D. 1918. *L'Ologénèse: nouvelle théorie de l'évolution et de la distribution géographique des êtres vivants*. Bemporad and Figlio Editori. Florencia.

Rueda M, Rodríguez MÁ, BA Hawkins. 2010. Towards a biogeographic regionalization of the European biota. *Journal of Biogeography*, 37(11): 2067-2076.

Russell PF, TR Rao. 1940. On habitat and association of species of anopheline larvae in south-eastern Madras. *Journal of the Malaria Institute of India*, 3(1): 153-178.

Saitou N, M Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4): 406-425.

Salas C. 2008. ¿Por qué comprar un programa estadístico si existe R? *Ecología Austral*, 18(2): 223-231.

Sánchez G. 2013. DiscriMiner: tools of the trade for discriminant analysis. Paquete de R versión 0.1-29. https://CRAN.R-project.org/package=DiscriMiner

Sanderson MJ, L Hufford. 1996. *Homoplasy*. Academic Press. San Diego.

Sankoff D. 1975. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1): 35-42.

Saraçli S, N Doğan, İ Doğan. 2013. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013(1): 1-8.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4): 592-593.

Schmera D, J Podani. 2018. Through the jungle of methods quantifying multiple-site resemblance. *Ecological Informatics*, 44: 1-6.

Schmidt HA, K Strimmer, M Vingron, A von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3): 502-504.

Schmidt HA, A von Haeseler. 2009. Phylogenetic inference using maximum likelihood methods. En Lemey P, M Salemi, AM Vandamme (eds.) *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Pp. 181-209. Cambridge University Press. Cambridge.

Schroeder MP, A González-Pérez, N López-Bigas. 2013. Visualizing multidimensional cancer genomics data. *Genome Medicine*, 5(1): 9.

Schuh RT. 2000. *Biological systematics: principles and applications*. Cornell University Press. Ithaca.

Schuh RT, AVZ Brower. 2010. *Biological systematics: principles and applications*. Cornell University Press. Ithaca.

Sharma S. 1996. *Applied multivariate techniques*. John Wiley and Sons. Nueva York.

Shepard RN. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27(3): 219-246.

Shepard RN. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3(2): 287-315.

Shi GR. 1993. Multivariate data analysis in palaeoecology and palaeobiogeography-a review. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 105(3-4): 199-234.

Simmons MP, KM Pickett, M Miya. 2004. How meaningful are Bayesian support values? *Molecular Biology and Evolution*, 21(1): 188-199.

Simpson GG. 1943. Mammals and the nature of continents. *American Journal of Science*, 241: 1-31.

Singh N, K Harvati, JJ Hubliny, CP Klingenberg. 2012. Morphological evolution through integration: a quantitative study of cranial integration in *Homo, Pan, Gorilla* and *Pongo*. *Journal of Human Evolution*, 62(1): 155-164.

Smith MR. 2019. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biology Letters*, 15(2): 20180632.

Sneath PH, RR Sokal. 1973. *Numerical taxonomy. The principles and practice of numerical classification*. WH Freeman Company. San Francisco.

Soetaert K. 2017. plot3D: plotting multi-dimensional data. Paquete de R versión 1.1.1. https://CRAN.R-project.org/package=plot3D

Sokal RR, CD Michener. 1958. A statistical method for evaluating systematic relationship. *University of Kansas Science Bulletin*, 28: 1409-1438.

Sokal RR, FJ Rohlf. 1962. The comparison of dendrograms by objective methods. *Taxon*, 11: 33-40.

Sokal RR, PH Sneath. 1963. *Principles of numerical taxonomy*. WH Freeman Company. San Francisco.

Sokal RR. 1961. Distance as a measure of taxonomic similarity. *Systematic Zoology*, 10(2): 70-79.

Sørensen TA. 1948. A method of stablishing groups of equal amplitude in plant sociology based of similarity of species content, and its application to analyses of vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter*, 5: 1-34.

Spence NA, PJ Taylor. 1970. Quantitative methods in regional taxonomy. *Progress in Geography*, 2: 1-64.

Stamatakis A, T Ludwig, H Meier. 2004. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4): 456-463.

Steinhaus H. 1956. Sur la division des corp materiels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III,* 12(4): 801-804.

Sugar CA, GM James. 2003. Finding the number of clusters in a dataset: an information-theoretic approach. *Journal of the American Statistical Association*, 98(463): 750-763.

Suzuki Y, GV Glazko, M Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 99(25): 16138-16143.

Suzuki R, H Shimodaira. 2015. pvclust: hierarchical clustering with P-values via multiscale bootstrap resampling. Paquete de R versión 2.0-0. https://CRAN.R-project.org/package=pvclust

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). https://paup.phylosolutions.com

Swofford DL, WP Maddison. 1987. Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, 87: 199-229.

Symonds MR, A Moussalli. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, 65(1): 13-21.

Tarca AL, VJ Carey, XW Chen, R Romero, S Drăghici. 2007. Machine learning and its applications to biology. *PLoS Computational Biology*, 3(6): e116.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2): 57-86.

Teetor P. 2011. *R cookbook: proven recipes for data analysis, statistics, and graphics*. O'Reilly Media. Sebastopol.

ter Braak CJ. 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics*, 41(4): 859-873.

Terentjev PV. 1931. Biometrische Untersuchungen über die morpho-logischen Merkmale von *Rana* ridibunda Pall: (Amphibia, Salientia). *Biometrika*, 23: 23-51.

Todeschini R, V Consonni, H Xiang, J Holliday, M Buscema, P Willett. 2012. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11): 2884-2901.

Tofilski A. 2008. Using geometric morphometrics and standard morphometry to discriminate three honeybee subspecies. *Apidologie*, 39(5): 558-563.

Venables WN, BD Ripley. 2002. *Modern applied statistics with S-PLUS*. Springer. Nueva York.

Wagner Jr WH. 1961. Problems in the classification of ferns. *Recent Advances in Botany*, 1: 841-844.

Walker M. 1968. *El pensamiento científico*. Grijalbo. México D.F.

Walsh J, AI Kovach, KJ Babbitt, KM O'brien. 2012. Fine-scale population structure and asymmetrical dispersal in an obligate salt-marsh passerine, the Saltmarsh Sparrow (*Ammodramus caudacutus*). *Auk*, 129(2): 247-258.

Ward Jr JH. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301): 236-244.

Warnes GR, B Bolker, L Bonebakker, R Gentleman, HAL Wolfgang, T Lumley, M Maechler, A Magnusson, S Moeller, M Schwartz, B Venables. 2019. gplots: various R programming tools for plotting data. Paquete de R versión 3.0.1.1. https://CRAN.R-project.org/package=gplots

Wartenberg D, S Ferson, FJ Rohlf. 1987. Putting things in order: a critique of detrended correspondence analysis. *American Naturalist*, 129(3): 434-448.

Weihs C, U Ligges, K Luebke, N Raabe. 2005. klaR analyzing German business cycles. En Baier D, Decker R, Schmidt-Thieme L (eds.) *Data analysis and decision support*. Pp. 335-343. Springer. Berlin.

Weinstein JN. 2008. A postgenomic visual icon. *Science*, 319(5871): 1772-1773.

Whelan S, P Liò, N Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics*, 17(5): 262-272.

Whitlock M, D Schluter. 2015. *The analysis of biological data*. Roberts and Co. Publishers. Greenwood Village.

Whittaker RH. 1973. *Ordination and classification of communities*. Junk. La Haya.

Wickham H, M Averick, J Bryan, W Chang, L McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilk, K Woo, H Yutani. 2019. Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43): 1686.

Wickham H, G Grolemund. 2016. *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media. Sebastopol.

Wiley EO. 1975. Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary taxonomists. *Systematic Zoology*, 24(2): 233-243.

Wiley EO. 1981. *Phylogenetics*. John Wiley and Sons. Nueva York.

Wilkinson M. 1995. More on reduced consensus methods. *Systematic Biology*, 44(3): 435-439.

Wilkinson L, M Friendly. 2009. The history of the cluster heat map. *American Statistician*, 63(2): 179-184.

Wilks SS. 1932. Certain generalizations in the analysis of variance. *Biometrika*, 24: 471-494.

Williams BK. 1983. Some observations of the use of discriminant analysis in ecology. *Ecology*, 64(5): 1283-1291.

Williams WT, MB Dale. 1965. Fundamental problems in numerical taxonomy. *Advance Botanical Research*, 2: 35-75.

Wishart D. 1969. Mode analysis. En Cole AJ (ed.) *Numerical taxonomy*. Pp. 282-308. Academic Press. Nueva York.

Wishart D. 2005. Number of clusters. En Everitt B, DD Howell (eds.) *Encyclopedia of statistics in behavioral science*. Pp. 1442-1446. John Wiley and Sons. Chichester.

Wolda H. 1981. Similarity indices, sample size and diversity. *Oecologia*, 50(3): 296-302.

Wong TT. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9): 2839-2846.

Wright A, D Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* 9: e109210.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5): 555-556.

Yang Z, B Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution*, 14(7): 717-724.

Yang Z, B Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology*, 54(3): 455-470.

Yang Z, B Rannala. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5): 303.

Yendle PW, HJ MacFie. 1989. Discriminant principal components analysis. *Journal of Chemometrics*, 3(4): 589-600.

Zar JH. 1999. *Biostatistical analysis*. Prentice Hall. Nueva Jersey.

INDICE DE AUTORES

INDICE TEMÁTICO

SOBRE LOS AUTORES

Facundo Xavier PALACIO
Es Licenciado en Biología de la Universidad Nacional de La Plata (UNLP), Doctor en Ciencias Biológicas de la Universidad Nacional de Tucumán y Diplomado en bioestadística básica aplicada, mediada con entorno R, de la Universidad Nacional de Córdoba. Sus temas de investigación son la ecología de aves, la evolución y la aplicación del análisis multivariado a datos biológicos. Es docente en la Cátedra de Matemática y de Elementos de Matemática de la Facultad de Ciencias Naturales y Museo (UNLP) e investigador del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) en la División Zoología Vertebrados del Museo de La Plata.

María José APODACA
Es Licenciada en Biología y Doctora en Ciencias Naturales de la Universidad Nacional de La Plata (UNLP). Sus temas de investigación son la biogeografía, la evolución biológica y la aplicación del análisis multivariado a estas disciplinas. Es docente e investigadora en la Cátedra de Biogeografía de la Facultad de Ciencias Naturales y Museo (UNLP) e investigadora del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), con lugar de trabajo en la División Plantas Vasculares del Museo de La Plata.

Jorge Víctor CRISCI
Es Licenciado en Botánica y Doctor en Ciencias Naturales de la Universidad Nacional de La Plata (UNLP). Sus temas de investigación son la biogeografía, la evolución biológica, la sistemática y la educación. Es Profesor Emérito de la UNLP, Jefe de la División Plantas Vasculares del Museo de La Plata, investigador del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y Académico de Número de la Academia Nacional de Agronomía y Veterinaria.

AZARA

FUNDACIÓN DE HISTORIA NATURAL

La Fundación Azara, creada el 13 de noviembre del año 2000, es una institución no gubernamental y sin fines de lucro dedicada a las ciencias naturales y antropológicas. Tiene por misión contribuir al estudio y la conservación del patrimonio natural y cultural del país, y también desarrolla actividades en otros países como Paraguay, Bolivia, Chile, Brasil, Colombia, Cuba y España.

Desde el ámbito de la Fundación Azara un grupo de investigadores y naturalistas sigue aún hoy en el siglo XXI descubriendo especies –tanto fósiles como vivientes– nuevas para la ciencia, y en otros casos especies cuya existencia se desconocía para nuestro país.

Desde su creación la Fundación Azara contribuyó con más de cien proyectos de investigación y conservación; participó como editora o auspiciante en más de doscientos libros sobre ciencia y naturaleza; produjo ciclos documentales; promovió la creación de reservas naturales y la implementación de otras; trabajó en el rescate y manejo de la vida silvestre; promovió la investigación y la divulgación de la ciencia en el marco de las universidades argentinas de gestión privada; asesoró en la confección de distintas normativas ambientales; organizó congresos, cursos y casi un centenar de conferencias.

En el año 2004 creó los Congresos Nacionales de Conservación de la Biodiversidad, que desde entonces se realizan cada dos años. Desde el año 2005 comaneja el Centro de Rescate, Rehabilitación y Recría de Fauna Silvestre "Güirá Oga", vecino al Parque Nacional Iguazú, en la provincia de Misiones. En sus colecciones científicas –abiertas a la consulta de investigadores nacionales y extranjeros que lo deseen– se atesoran más de 200.000 piezas. Actualmente tiene actividad en varias provincias argentinas: Misiones, Corrientes, Entre Ríos, Chaco, Catamarca, San Juan, La Pampa, Buenos Aires, Río Negro, Neuquén y Santa Cruz. La importante producción científica de la institución es el reflejo del trabajo de más de setenta científicos y naturalistas de campo nucleados en ella, algunos de los cuales son referentes de su especialidad.

La Fundación recibió apoyo y distinciones de instituciones tales como: Field Museum de Chicago, National Geographic Society, Consejo Superior de Investigaciones Científicas de España, Fundación Atapuerca, Museo de la Evolución de Burgos, The Rufford Foundation, entre muchas otras.

www.fundacionazara.org.ar
www.facebook.com/fundacionazara

Este libro es una introducción al análisis multivariado, incluyendo la reconstrucción filogenética. El análisis multivariado es una metodología que intenta encontrar patrones de similitud entre objetos (por ejemplo, entre especies, entre cuadrantes geográficos, entre localidades) en base a un conjunto de variables (por ejemplo, atributos de las especies, presencia de especímenes en cuadrantes geográficos, parámetros ambientales de una localidad). Estos patrones permiten formar grupos cuyos objetos son más similares entre sí que con los objetos de otros grupos, así como contrastar hipótesis sobre las relaciones entre los objetos, explicar la causalidad de los agrupamientos, y predecir objetos y variables todavía no utilizados o descubiertos.

El libro se estructura en cinco grandes ejes temáticos distribuidos en ocho capítulos: (1) generación de los datos (objetos x variables), (2) medidas de similitud entre objetos, (3) análisis de agrupamientos, (4) técnicas de ordenación y (5) análisis filogenéticos.

El objetivo es que el investigador pueda reconocer un problema que requiera la aplicación del análisis multivariado, decidir cuál técnica es la más apropiada para sus datos y utilizar el programa computacional R para aplicarla.

vmeditores.com.ar

AZARA
FUNDACIÓN DE HISTORIA NATURAL

umai
Universidad
Maimónides