

## Evolución y Filogenia de Arthropoda

Sección I: Conceptos y métodos en el estudio de la filogenia

# CLADISMO: LA RECONSTRUCCIÓN FILOGENÉTICA BASADA EN PARSIMONIA

Miquel Àngel Arnedo

Center for Conservation Research y Training, University of Hawai'i  
3050 Maile Way, Gilmore 409, Honolulu, HI 96822  
Dirección actual: Division of Insect Biology, ESPM  
201 Wellman Hall, University of California-Berkeley, Berkeley, CA-94720-3112  
miquelan@nature.berkeley.edu

### Resumen

La metodología denominada cladismo o cladística se ha convertido en el paradigma de la reconstrucción filogenética. En el presente artículo se revisan y comentan sus principios básicos y su justificación filosófica. El cladismo se caracteriza frente a otras escuelas sistemáticas por distinguir las homologías derivadas (sinapomorfías) de las primitivas (plesiomorfías), y utilizar solo las primeras como caracteres válidos para la agrupación de taxones. La forma en que estas sinapomorfías son reconocidas, distingue las principales aproximaciones actuales a la reconstrucción filogenética. La parsimonia es un criterio para la selección de hipótesis filogenéticas alternativas, bajo el cual la hipótesis preferible es la que minimiza el número de transformaciones adicionales de los caracteres. El análisis cladístico consiste de una serie de etapas sucesivas, que incluyen: la selección de los taxones de estudio, la construcción de la matriz de caracteres, la búsqueda de los árboles más parsimoniosos y la evaluación de la fiabilidad de la hipótesis filogenética escogida. Se presta una especial atención a dos temáticas que son actualmente objeto de un acalorado debate: la ponderación de los caracteres y el análisis de datos de distinta naturaleza.

**Palabras clave:** Reconstrucción filogenética, Terminología cladista, Ponderación de los caracteres, Búsquedas de árboles, Fiabilidad de los cladogramas, Árboles de consenso, Análisis combinado.

### Cladistics: Phylogenetic reconstruction based on parsimony

#### Abstract

The cladistic framework has become the dominant paradigm for phylogenetic reconstruction. Here, I review and discuss its basic tenets and philosophical justification. Cladistics is characterized from other systematic schools by separating derived homologies (synapomorphies) from primitive (plesiomorphies), and by using only the former as valid characters for grouping taxa. Current approximations to phylogenetic reconstruction can be distinguished by their criteria for recognizing synapomorphies. Under parsimony the preferred hypothesis minimizes the number of character transformations.

Cladistic analyses consist of a series of subsequent stages that include taxon sampling, construction of the character matrix, search for the most parsimonious tree and assessment of the reliability of the selected phylogenetic hypothesis. Special attention is paid to hotly debated issues such as character weighting and the analysis of characters from different sources.

**Key words:** Cladistics, Phylogenetic reconstruction, Cladistic terminology, Character weighting, Tree search, Cladogram reliability, Consensus trees, Combined analyses.

## 1. CONCEPTOS BÁSICOS

Incluso a los ojos del profano, la naturaleza exhibe una cierta ordenación jerárquica. La existencia de dicha jerarquización en la diversidad biológica, es el resultado del proceso evolutivo, es decir, de la descendencia con modificación (Eldredge y Cracraft, 1980). La reconstrucción o inferencia filogenética es el proceso por el cual mediante la aplicación de un conjunto determinado de técnicas y metodologías, obtenemos una hipótesis sobre las relaciones genealógicas o evolutivas entre los organismos objeto de estudio. Dicha hipótesis de relación se representa en forma de diagrama arborescente, al que se denomina árbol filogenético o filogenia. El árbol filogenético se compone de dos elementos principales: los nodos y los internodos o ramas (Fig. 1). Existen dos tipos de nodos: los

nodos terminales, a los cuales únicamente llega una rama y que representan los taxones de interés, y los nodos internos, en los cuales confluyen más de una rama y que representan los antepasados hipotéticos de los nodos que de ellos derivan. La multiplicación de ramas en los nodos internos se interpreta como un fenómeno de especiación, o formación de dos taxones a partir de un tercero ancestral. Las ramas, las líneas que conectan los nodos, representan los cambios producidos entre dos nodos a lo largo del tiempo, es decir, la evolución. Las ramas pueden ser terminales, si conectan un nodo interno con un nodo terminal, internas, si conectan dos internodos, o raíces, si tan solo se unen a un nodo. Las raíces representan el origen o el inicio evolutivo del grupo de taxones estudiado y

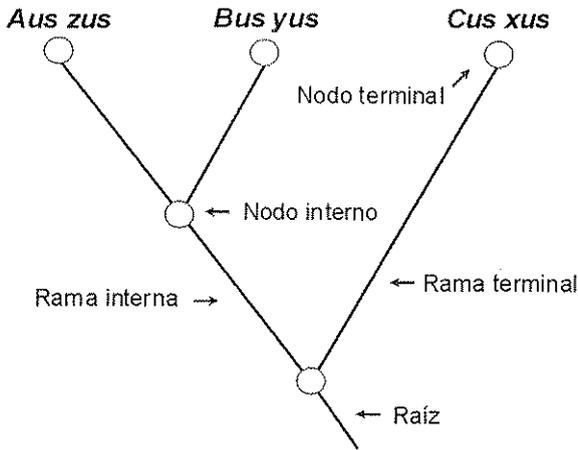


Fig. 1.- Árbol filogenético donde se indican los principales elementos que lo componen.

su función es la de dar direccionalidad al árbol, permitiendo así referirse a nodos anteriores o antepasados y nodos posteriores o descendientes. Si el árbol carece de raíz, se le denomina red, o simplemente árbol sin raíz. Por otra parte, el grupo interno o *ingroup*, es el conjunto de taxones estudiados por el investigador, el grupo externo o *outgroup* es cualquiera de los taxones utilizados en un análisis no incluidos entre los taxones de estudio. El grupo hermano o *sister group* es el grupo genealógicamente más cercano al grupo considerado o, en otras palabras, el grupo que comparte un antepasado común más reciente con el mismo.

La metodología denominada cladística o cladismo se ha convertido en el paradigma de la inferencia filogenética (Kluge y Wolf, 1993). El cladismo fue originalmente introducido en biología por el entomólogo alemán Willi Hennig, quién en el año 1950 publicó el libro *Grundzüge einer Theorie der phylogenetischen Systematik*, el cual pasaría a marcar un hito en la historia de la biología comparativa y la ciencia evolutiva (Janvier, 1984). Sin embargo, esta obra permaneció prácticamente ignorada hasta su traducción al inglés, aparecida el año 1966 bajo el título *Phylogenetic systematics*. Las ideas de Hennig fueron formuladas en el contexto de una pequeña revolución dentro de la taxonomía, cuyo principal objetivo era dotarla de una mayor objetividad. La taxonomía tradicional había sido acusada de no ser una auténtica ciencia debido precisamente a la gran subjetividad de los criterios que utilizaba (Quicke, 1993). Paralelamente a la aparición de la sistemática filogenética, fue propuesta otra metodología con unos objetivos similares pero con una filosofía radicalmente diferente: la fenética o taxonomía numérica (Sneath y Sokal, 1973). Finalmente, se puede definir una tercera escuela taxonómica, denominada taxonomía

evolutiva o ecléctica (Mayr y Ashlock, 1991; Simpson, 1961), resultado de la adopción de los principios de la teoría sintética de la evolución a la taxonomía. La aportación de Hennig a esta "revolución taxonómica" fue la de postular el principio de que las clasificaciones debían reflejar exactamente las relaciones entre los taxones objeto de dicha clasificación y, más concretamente, el orden de ramificación observado en sus filogenias. Algunos taxónomos evolutivos adujeron que la ramificación era solo uno de los aspectos de la filogenia, y que el otro, la anagénesis (es decir, las transformaciones o cambios acumulados entre dos ramificaciones), se ignoraba completamente. E. Mayr propuso la denominación de cladismo (del griego *Klados*: rama) para esta escuela taxonómica, para distinguirla de la, de acuerdo con el propio Mayr, auténticamente filogenética que incluiría los dos componentes. Evidentemente, reclamar que la clasificación debía reflejar fielmente la filogenia requería la proposición de una metodología que permitiera recuperar las relaciones filogenéticas. Ha sido precisamente este aspecto del cladismo el que con más fuerza se ha desarrollado, convirtiéndose en la actualidad en la metodología prácticamente universal para la reconstrucción filogenética.

La forma más simple de caracterizar el cladismo frente a otras metodologías taxonómicas es a través del tratamiento que cada una de ellas hace de los caracteres. Un carácter es una característica, una parte observable, un atributo de un organismo, que puede ser adecuadamente descrito o definido (Wiley et al., 1991). En fenética, todos los caracteres son tratados como iguales, y se habla de similitud global (*overall similarity*). El conjunto de caracteres considerado y a través de la aplicación de una determinada función de similitud (o, en su caso, de disimilitud), es substituido por un único valor que pretende expresar la proximidad (o distancia) entre los diferentes taxones. En taxonomía evolutiva se distinguen dos tipos de caracteres: los homólogos y los homoplásicos. Dos caracteres son homólogos si están relacionados evolutivamente, de tal forma que uno proviene de la transformación del otro o que los dos provienen de la transformación de un tercero. Al conjunto de caracteres homólogos se le denomina serie de transformación (Wiley et al., 1991). Cabe destacar que, actualmente, la citada nomenclatura está bastante en desuso y que, en general, se tiende a referirse a la serie de transformación como el carácter, y a los caracteres como los estados del carácter. Por otra parte, dos caracteres son homoplásicos si no están relacionados evolutivamente, es decir, si provienen de transformaciones independientes de un carácter anterior. Pueden distinguirse tres tipos diferentes de homoplasias (Fig. 2): convergencias o transformaciones de un mismo carácter en líneas evolutivas muy distantes, paralelismos o transformaciones del mismo carácter en linajes muy cercanos y reversiones, transformaciones de un

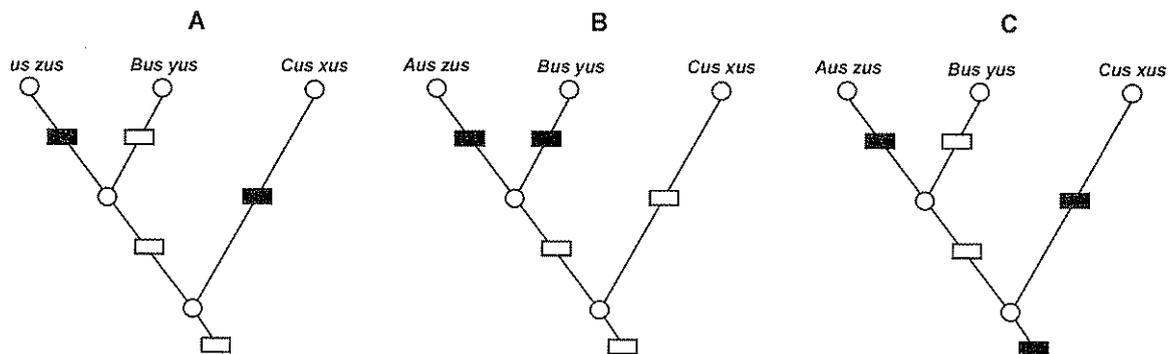


Fig. 2.- Representación gráfica de los principales tipos de caracteres homoplásicos: A. Convergencia, B. Paralelismo, C. Reversión.

carácter a una estado anterior evolutivamente. Los caracteres basados en secuencias nucleotídicas presentan una forma adicional de homoplasia. Dado que tan solo existen cinco caracteres (o estados) posibles (A, C, G, T, -), la probabilidad de compartir el mismo carácter por cambios independientes es extremadamente alta. Si se compara una determinada posición nucleotídica (=serie de transformación o carácter) en taxones muy divergentes o con tasas de cambio muy elevadas, la superposición de cambios (*multiple hits*) habrá borrado cualquier información filogenética de las mismas. La existencia de sustituciones múltiples en un determinado conjunto de posiciones puede ponerse de manifiesto a través de la relación existente entre el número de cambios y el grado de divergencia entre taxones. A partir de cierto valor de divergencia, el número de cambios ya no aumenta debido a que estos se dan en posiciones que ya habían cambiado anteriormente y que, por tanto, ya habían sido contabilizados. A este fenómeno se le denomina saturación. Finalmente, en cladismo se distinguen dos tipos diferentes de caracteres homólogos: las plesiomorfias, o caracteres plesiomórficos, y las apomorfias, o caracteres apomórficos. Si dos caracteres son homólogos, es decir, uno proviene de la transformación evolutiva del otro, el primero constituye una apomorfia y el segundo una plesiomorfia. Si el carácter es compartido por más de un taxón, se le denomina sinapomorfia o simplisimorfia, según los casos. El término autapomorfia define un carácter derivado y exclusivo de uno de los taxones. En la Fig. 3 se representan los diferentes tipos de caracteres comentados. Así por ejemplo el carácter círculo gris es homoplásico, dado que es el producto de la transformación independiente del carácter círculo blanco en dos líneas diferentes. Sin embargo, círculo gris y círculo blanco son homólogos, dado que uno proviene de la transformación del otro. De la misma manera, todos los caracteres con forma de cuadrado y todos los caracteres con forma de triángulo son homólogos entre ellos, es decir, forman series de transformación, o, en la nomenclatura más utilizada, son estados del mismo carácter. Ahora bien, en el caso de los triángulos, los triángulos blancos son plesiomórficos respecto a los grises, que, por tanto, son apomórficos. Igualmente, los cuadrados grises son derivados, sinapomorfos, de los cuadrados blancos, simplisimorfos, a pesar de que al mismo tiempo, son plesiomórficos respecto al cuadrado negro, que de ellos deriva, siendo, por tanto, una autapomorfia.

La importancia de distinguir entre los diferentes tipos de caracteres comentados es que cada uno de ellos define un determinado tipo de agrupación. Así, un conjunto de taxones que comparte una apomorfia, es decir una sinapomorfia, constituye un grupo monofilético (grupo A, Fig. 3). Un conjunto de taxones con una plesiomorfia común, una simplisimorfia, forma un grupo denominado parafilético (grupo B, Fig. 3), y, finalmente, un conjunto de taxones agrupados en base a la presencia de un carácter homoplásico, forman un grupo polifilético (grupo C, Fig. 3). Las autapomorfias, obviamente, no definen agrupaciones. Los grupos parafiléticos y los polifiléticos son agrupaciones artificiales. Solo los grupos monofiléticos, o clados, constituyen grupos naturales y, por tanto, son los únicos que tienen importancia filogenética (Janvier, 1984). La incapacidad de las aproximaciones fenética y evolutiva o ecléctica de detectar la existencia de sinapomorfias y, por tanto, su imposibilidad de reconocer grupos monofiléticos, las desacredita como metodologías de reconstrucción filogenética.

La reconstrucción de la filogenia de un conjunto de organismos es equivalente al proceso de recuperar todos y cada uno de los grupos monofiléticos formados por éstos, lo

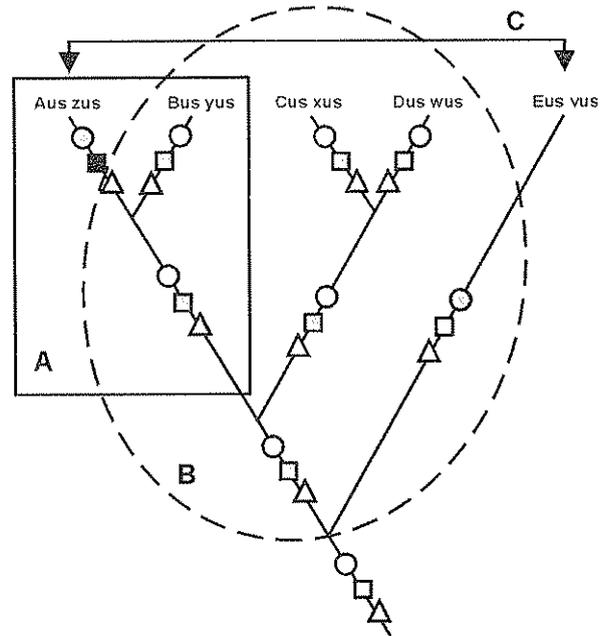


Fig. 3.- Filogenia hipotética de cinco especies, donde se muestran los tipos de caracteres y agrupaciones que éstos definen: A. Grupo monofilético, B. Grupo parafilético, C. Grupo polifilético.

que a su vez, supone establecer la sinapomorfias que los definen. Por tanto, los caracteres representan las evidencias que permiten recuperar las relaciones filogenéticas de los taxones. Así pues, la cuestión de como recuperar las relaciones filogenéticas queda reducida a evaluar un conjunto de caracteres para distinguir aquellos que representan sinapomorfias. Al proceso de selección y evaluación de los caracteres de los taxones estudiados se le denomina análisis cladístico. El resultado de este análisis es la elaboración de un cladograma, que no es más que un diagrama en forma de árbol donde se representan la distribución de los distintos caracteres y los grupos que definen. La interpretación de las ramificaciones como fenómenos de especiación y de los nodos internos como antepasados, convierte al cladograma en una filogenia. Finalmente, el establecimiento de los posibles procesos, así como la inferencia del contexto biogeográfico o temporal que generaron dicha filogenia constituyen el escenario. La gran mayoría de autores tienden a sinonimizar cladograma con filogenia y, de la misma manera, análisis cladístico con análisis filogenético. Para comprender mejor la diferencia entre cladograma y filogenia, basta destacar el hecho de que un mismo cladograma puede representar diferentes filogenias, tal y como se representa en la Fig. 4. Las tres filogenias representadas son compatibles con el cladograma. En la filogenia A, la especie *Cus xus* es un antepasado de *Bux yus* que, a su vez, lo es de *Aus zus*. En la filogenia B, *Cus xus* es antepasado del ancestro común de *Aus zus* y *Bus yus*. Finalmente, en la filogenia C, *Aus zus* y *Bus yus* comparten un antepasado común que no lo es de *Cus xus*, a pesar de que también comparten un antepasado común con ésta. El escenario representado es compatible con la filogenia C, donde se hipotetiza un tiempo geológico de separación de *Cus xus* de las otras, así como un fenómeno de vicariancia como explicación al origen de *Aus zus* y *Bus yus*.

Ahora bien, cómo pueden distinguirse las sinapomorfias del resto de caracteres. La determinación de cuales son los caracteres sinapomórficos se lleva a cabo en tres pasos sucesivos. El primer paso consiste en establecer las series de transformación (= estados de los caracteres), en base a las

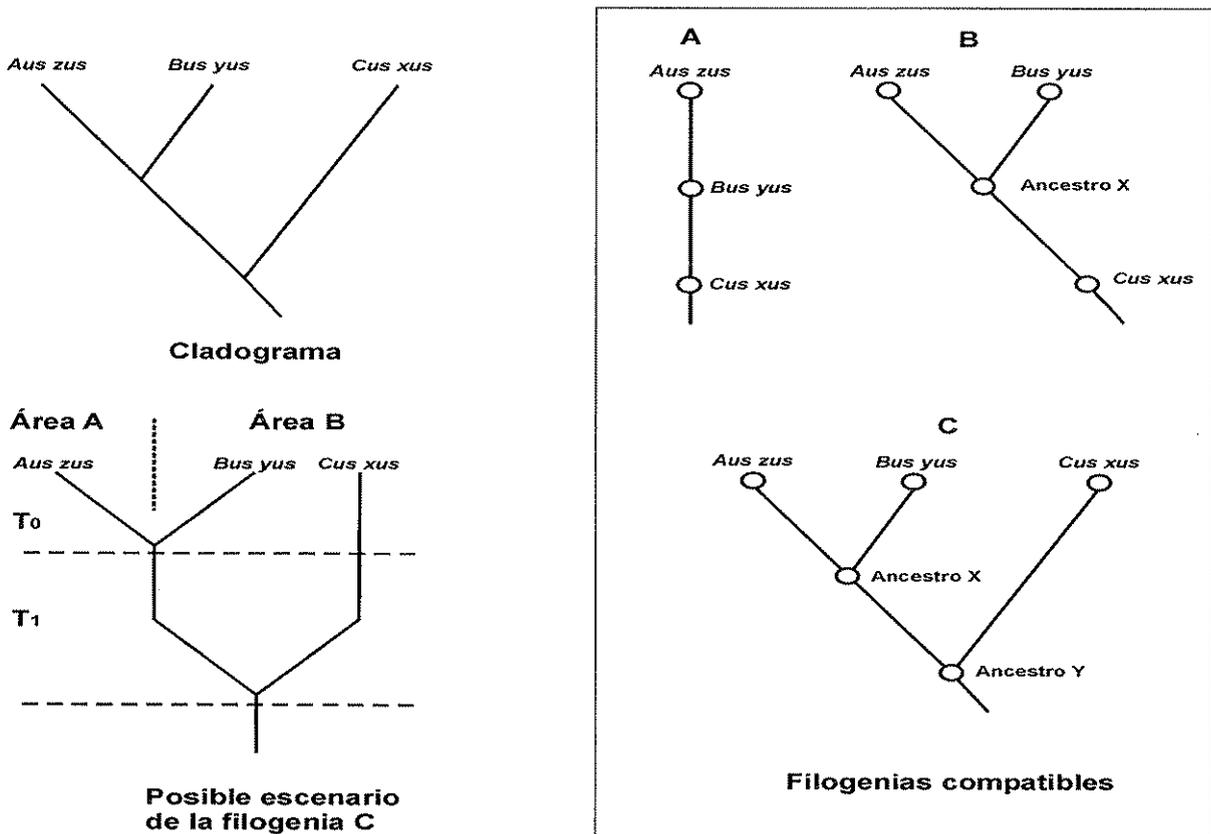


Fig. 4.- Diferencias entre cladograma, filogenia y escenario. Las tres filogenias representadas (A-C) son compatibles con el cladograma.

denominadas homologías primarias (De Pinna, 1991). Éstas, que corresponden con el concepto clásico de homología, han de considerarse como una hipótesis preliminar que ha de corroborarse o rechazarse en los pasos sucesivos. Diversos autores prefieren utilizar la palabra homología como sinónimo de sinapomorfia, y denominan a esta etapa “establecimiento de las hipótesis de agrupación” (Brower y Schawaroch, 1996). En la definición de las series de transformación (=caracteres) debe tenerse en cuenta el conocido como principio auxiliar de Hennig (Hennig, 1953, 1966), el cual formula que “nunca debe asumirse ni la convergencia ni la evolución paralela de un carácter, siempre asumir homología del mismo en ausencia de evidencia contraria”. Desde un punto de vista operacional, estas hipótesis preliminares de homología pueden definirse utilizando los criterios de Remane (1956): similitud en la posición, similitud en la microestructura y en el desarrollo y continuidad entre formas intermedias. El paso siguiente es determinar la polaridad de los caracteres (=estados de carácter), es decir, que caracteres (estados) son primitivos (=plesiomórficos) y cuales son derivados (=apomórficos). Hennig propuso cuatro reglas, una principal y tres accesorias, para establecer dicha polaridad:

1. *Regla del grupo externo* (esquema de argumentación o argumentación de los caracteres). De dos o más caracteres (=estados) homólogos establecidos en un grupo monofilético, aquel que también pueda encontrarse en su grupo hermano, corresponderá al estado plesiomórfico de dicho carácter, mientras el que se encuentre exclusivamente en el grupo interno será el apomórfico (Wiley et al., 1991).
2. *Regla de la progresión*: Hennig asumía que los taxones más primitivos del grupo estudiado deberían hallarse en

las zonas más cercanas al centro de origen de dicho grupo y que, por tanto, los caracteres (=estados) mostrados por estos serían los primitivos (Janvier, 1984):

3. *Regla de la correlación de las series de transformación*. La hipótesis de la polaridad de los caracteres de una serie de transformación (=estados de carácter), puede corroborarse mediante la congruencia con la distribución de otras series (=caracteres).
4. *Regla del criterio ontogenético*. A pesar de haber sido originalmente formulado como un criterio adicional, algunos autores han considerado este método como el más importante porque muestra una evidencia directa de la polaridad del carácter (=estado), mientras que los restantes representan evidencias más o menos indirectas (Nelson y Platnick, 1981; Patterson, 1982). Nelson (1978) reformuló esta regla como “dada una determinada serie de transformación de caracteres (estados de carácter), desde un carácter (estado) más general a uno menos general, el más general es el primitivo, y el menos general es el derivado”. Por contra, para otros autores (Patterson, 1982), es el orden de aparición de los caracteres (=estados) durante la ontogenia del organismo, lo que se corresponde con su polaridad.

Un criterio adicional que puede ser usado para el establecimiento de la polaridad es la secuencia estratigráfica (Gingerich, 1979), en la cual la polaridad se obtiene a partir de la observación de la secuencia de aparición de caracteres en el registro fósil.

El tercer y último paso consiste en diferenciar las auténticas homologías de las homoplasias. El principal criterio utilizado es la congruencia entre los caracteres, del cual la cual la regla 3 anteriormente comentada es una ampliación. La asunción básica es que las diferentes homolo-

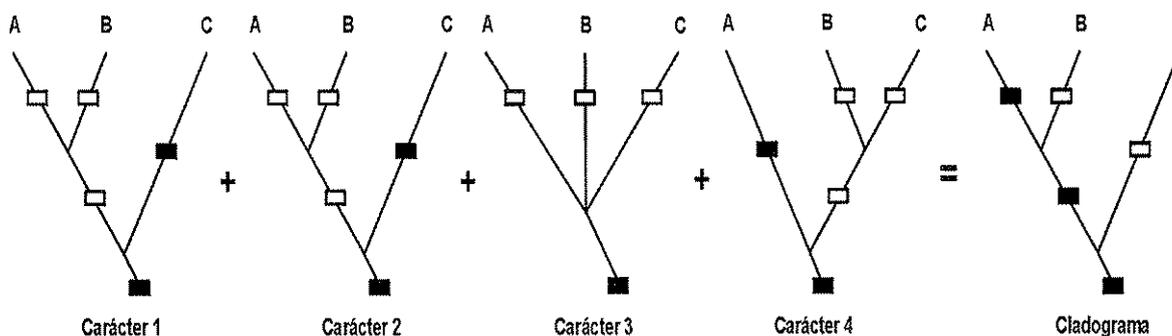


Fig. 5.- Test de congruencia entre los caracteres (apomorfias en blanco, plesiomorfias en negro) como criterio para la distinción entre homología y homoplasias.

gías deberían mostrar un patrón congruente entre ellas, debido a que son el resultado de un mismo proceso: la descendencia con modificación o evolución. Por contra, la distribución de las homoplasias debería ser incongruente con el resto de caracteres, ya que no guardan necesariamente relación alguna con la jerarquía resultante de las relaciones genealógicas. Un ejemplo de la aplicación de dicho test se muestra en la Fig. 5. Supongamos que tenemos tres taxones y cuatro series de transformación (=caracteres), en los cuales se ha establecido la polaridad de sus caracteres (=estados) y por tanto se sabe las agrupaciones que cada uno apoya o respalda. Al comparar las diferentes agrupaciones, se observa que dos de las series de transformación (caracteres 1 y 2) respaldan los mismos grupos, una tercera (carácter 3) no es informativa y la cuarta (carácter 4) es incongruente. La explicación más simple es que la cuarta serie (carácter 4) es homoplásica. Finalmente, se muestra la reinterpretación de la última serie de transformación (carácter 4) sobre el cladograma resultante.

**2. APROXIMACIONES ACTUALES A LA INFERENCIA FILOGENÉTICA**

Hasta el momento se han discutido los principios y la metodología de lo que podría denominarse cladismo Hennigiano o filogenética sistemática clásica. Sin embargo, el posterior desarrollo de alguno de estos principios, principalmente debido a su implementación en ordenadores y a la adquisición de nuevos tipos de caracteres, básicamente secuencias nucleotídicas de los ácidos nucleicos, ha dado lugar a la aparición de diferentes metodologías filogenéticas. Éstas, a pesar de compartir como base común el reconocimiento de los grupos monofiléticos como únicos indicadores de relaciones filogenéticas, difieren profundamente en la manera en que estos son establecidos. El punto de partida de todos estos métodos es el mismo: la matriz de caracteres (fig. 6), la cual resulta de asignar a cada taxón objeto de estudio el carácter (=estado) que presenta de la lista de series de transformación (=caracteres) previamente establecida, utilizando el principio auxiliar de Hennig (1966) y mediante la aplicación de los criterios clásicos de determinación (Remane, 1956).

Los principales métodos de análisis utilizados en la actualidad son: (1) el análisis de parsimonia, (2) el análisis de máxima verosimilitud (*maximum likelihood*) y (3) el conjunto de análisis basados en matrices de distancias.

**2.1. Análisis de parsimonia**

La utilización de la parsimonia en filogenia fue originalmente introducida como criterio de construcción de árboles en algoritmos computacionales (Camin y Sokal, 1965; Edwards y Cavalli-Sforza, 1963; Fitch, 1971; Kluge y Farris, 1969) y posteriormente justificada desde una óptica filosófica y epistemológica (Eldredge y Cracraft, 1980; Farris, 1969; Farris, 1974; Farris, 1983; Kluge, 1985; Sober, 1988; Wiley, 1981). La asunción básica es que la existencia de compatibilidad entre las agrupaciones derivadas de las distintas series de transformación (=caracteres) es el resultado de la existencia de una genealogía, es decir, la presencia de antepasados comunes. Ahora bien, la homoplasia no puede ser explicada en estos términos y algún proceso distinto debe ser invocado para explicar su existencia. A su vez, no hay ninguna razón que permita suponer que estos procesos distintos a la propia descendencia con modificación, deban estar relacionados entre ellos y, por tanto, las homoplasias no tienen porque tener una distribución compatible ni entre ellas, ni con el resto de caracteres. Un principio básico en ciencia es el denominado "navaja de Ockham", el cual establece que: "...dadas distintas soluciones posibles a un problema, la solución más económica es la preferible" (Scotland, 1992). En el presente contexto, la hipótesis más parsimoniosa es aquella que explica la totalidad de los datos, con el menor numero posible de invocaciones de procesos alternativos *a posteriori* (Wiley, 1981). Un error frecuente es el suponer que la parsimonia requiere que la evolución deba haber transcurrido a través de los caminos más cortos posibles. La implementación de la parsimonia en reconstrucción filogenética no tienen nada que ver con esta afirmación, de la misma forma que tampoco implica que las

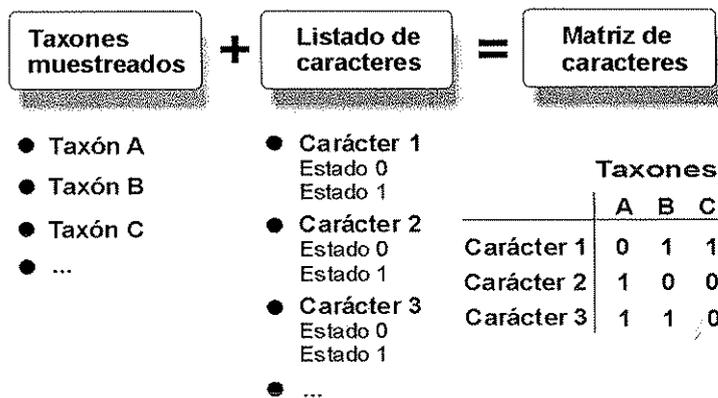


Fig. 6.- Ejemplo de matriz de caracteres y elementos que la componen.

homoplasias hayan de ser necesariamente fenómenos poco frecuentes (Farris, 1983). Cuando se minimizan el número de homoplasias necesarias para explicar un árbol filogenético, se está maximizando al mismo tiempo el número de congruencias entre los caracteres que, se asume, son el resultado de la existencia de relaciones genealógicas entre los mismos. Operacionalmente, la utilización del criterio de parsimonia puede interpretarse de la siguiente manera: si los caracteres de una determinada serie de transformación (=estados) son homólogos, el número de transformaciones (cambios) entre ellos es el mínimo de los posibles, el cual es igual al número total de caracteres (=estados) de la serie menos uno, dado que su presencia es el resultado de compartir un antepasado común. Por el contrario, si alguno de los caracteres de la serie (=estados) es homoplásico, esto implicará que existe al menos una transformación adicional y, por tanto, el número de transformaciones (cambios) será superior, al menos en uno, al mínimo. Por tanto, al minimizar el número de cambios totales de las series de transformación (=caracteres), se está minimizando el número de caracteres (=estados) homoplásicos. La implementación computacional del criterio de parsimonia es extraordinariamente simple, ya que aporta un criterio de evaluación de los diferentes árboles: el cladograma más parsimonioso es aquel con el menor número de transformaciones entre los caracteres (=estados). Una de las ventajas de la incorporación del criterio de parsimonia al análisis cladístico y su implementación computacional, es que hace innecesaria la polarización de los caracteres *a priori*, derivándose ésta del propio análisis. Esta propiedad fue originalmente destacada por Farris (1982) quien caracterizó explícitamente la polaridad en términos de la situación de la raíz en el árbol. El razonamiento es el siguiente. La aplicación del criterio de parsimonia supone contabilizar el número de transformaciones entre caracteres (=estados) en el árbol a evaluar, lo cual a su vez requiere asignar un determinado carácter (=estado) a los nodos internos correspondientes a los antepasados hipotéticos. Esto se lleva a cabo de manera que el carácter (=estado) asignado minimice el número de transformaciones. A este proceso de reconstrucción de los caracteres (=estados) en los nodos internos se le denomina optimización. Dado que es el propio análisis el que asigna dichos caracteres (=estados), si se fija la raíz del árbol queda automáticamente determinada la dirección de la transformación de los mismos y, por tanto, su plesiomorfía relativa (Farris, 1982). Consecuentemente, lo único necesario para polarizar los caracteres (y, por tanto, distinguir apomorfias de plesiomorfias) es situar la raíz del árbol. Esto puede llevarse a cabo añadiendo a la matriz uno o más taxones externos (*outgroups*), analizando todos los taxones conjuntamente, y situando la raíz en el internodo que separe a los taxones externos de aquellos que forman parte del grupo estudiado. Esta proceder permite a su vez evaluar la monofilia del grupo interno, a través de la adición de diversos *outgroups* y comprobando si la asignación de la raíz es compatible con la misma. La relación entre polaridad, *outgroups* y establecimiento de la raíz ha sido recientemente sistematizada por Nixon y Carpenter (1993). Esta propiedad no es exclusiva de la parsimonia, y es aplicable a cualquier metodología que permita la asignación de la raíz en cualquier nodo del árbol, sin cambiar el valor del mismo bajo el criterio de optimización o algoritmo de construcción escogido.

A diferencia de otros métodos, la parsimonia es agnóstica en relación a los detalles del proceso evolutivo, reflejando simplemente el apoyo empírico de las diferentes hipótesis (Farris, 1983). Como se comentará posteriormente, ciertas aproximaciones a la inferencia filogenética requieren el establecimiento explícito de un modelo evolutivo. Sin

embargo, algunos autores han señalado que la parsimonia asume, implícitamente, la existencia de una serie de condiciones: (1) la independencia de los caracteres, (2) la independencia de los taxones, es decir, la ausencia de hibridación, introgresión o transferencia lateral de información, y (3) tasas similares de cambios a lo largo de las ramas del árbol (Chippindale y Wiens, 1994). La necesidad de independencia de los caracteres deriva directamente de la concepción hipotético-deductiva de la inferencia filogenética. La congruencia entre caracteres es la base del apoyo de una determinada hipótesis sobre las restantes, si esta congruencia es el producto de la dependencia o relación condicional de los caracteres entonces la misma queda en entredicho. La segunda condición es el resultado de que cuando las relaciones existentes entre los taxones con tocogenéticas (reticulares), las sinapomorfias dejan de ser indicadores inequívocos de descendencia de un antepasado común (Nixon y Wheeler, 1990). La tercera condición ha sido propuesta por autores que defienden una aproximación estadística al problema de la reconstrucción filogenética y está basada en el comportamiento de la parsimonia en simulaciones por ordenador utilizando datos moleculares y bajo determinados modelos evolutivos. Felsenstein (1978) demostró computacionalmente con cuatro taxones que si la tasa de cambios en las ramas terminales es mucho mayor que la existente en el resto, la parsimonia une sistemáticamente estos taxones, independientemente de sus verdaderas relaciones genealógicas. A esta situación se la denomina "atracción de las ramas largas" (*long branch attraction*). De dichos resultados se deriva que la parsimonia asume una cantidad de cambios similar entre las ramas. Investigaciones posteriores han establecido que, para que el árbol real pueda ser recuperado utilizando parsimonia, ha de existir una cantidad similar y constante de cambios entre los caracteres (=estados), entre las series de transformación (=caracteres) y entre los taxones (Yang, 1996). Sin embargo, debe destacarse que la existencia de este efecto en casos reales no ha podido ser demostrado, y que simulaciones utilizando distintas condiciones de tasas evolutivas demuestran que existe una cierta región del espacio de parámetros en el cual son los métodos basados en modelos y no la parsimonia los que resultan inconsistentes (*long branch repulsion*, Sidall, 1998).

Otras críticas han sido formuladas en contra de la aplicación de la parsimonia como criterio para reconstrucción filogenética. Se ha propuesto que no existe ninguna razón para suponer que, en evolución, la divergencia de los caracteres sea siempre más común que los fenómenos que producen homoplasia (Panchen, 1979). Como se ha explicado anteriormente, la aplicación del principio de parsimonia no implica que se asuma que la homoplasia sea un fenómeno infrecuente, simplemente que ésta no tiene porque ser congruente con la distribución del resto de caracteres. Evidentemente, pueden existir situaciones en las que existan un conjunto de homoplasias que sean congruentes entre ellas debido a que sean el producto de un proceso común, como por ejemplo la adaptación a un medio ambiente concreto. Por ejemplo, si se reconstruye una filogenia de vertebrados, pueden existir caracteres, como la ausencia de pelo y la presencia de aletas, con una distribución congruente y que una sistemáticamente peces con delfines. Sin embargo, la adición de más caracteres, o incluso la reinterpretación de los mismos (p.ej. la estructura anatómica de las aletas), acabará por demostrar la incongruencia de los mismos. Al fin y al cabo, la adaptación a un determinado medio puede transformar los caracteres, pero no borrar la existencia de lo heredado o crearlos *de novo*. De todas maneras, la mayoría de críticas provienen de aquellos que entienden la reconstrucción filogenética como un proble-

ma estadístico. Una propiedad importante de un estimador estadístico es la denominada consistencia. Un estimador estadístico es consistente si tiende hacia el valor verdadero de determinado parámetro a medida que se aumenta la cantidad de observaciones (Felsenstein, 1988b). Se ha demostrado a través de simulaciones de datos moleculares que la parsimonia puede ser inconsistente en determinadas situaciones (Felsenstein, 1978; Penny et al., 1992). La respuesta a estas críticas ha sido la de destacar la naturaleza teórica de los modelos asumidos para generar dichos datos y la ausencia de ejemplos reales que demuestren dicho comportamiento (Farris, 1983). Por otra parte, Yang (1996) ha destacado que el criterio de parsimonia utiliza solo una parte de la información disponible y que otros métodos, y más concretamente la máxima verosimilitud, utilizan de manera más eficiente dichas observaciones. Esto es debido a que en parsimonia solo se utilizan aquellos caracteres que apoyan ciertas agrupaciones de taxones, mientras que en distancias o máxima verosimilitud los caracteres invariables o autapomórficos también pueden ser informativos. Sin embargo, debe tenerse en cuenta que el precio a pagar, dicho aumento de información, es la adopción explícita de un determinado modelo evolutivo.

## 2.2. Análisis de máxima verosimilitud (*maximum likelihood*)

Este representa un cambio radical en la aproximación a la inferencia filogenética, la cual pasa a ser considerada como un problema esencialmente estadístico. Dicha postura es absolutamente irreconciliable con la anteriormente expuesta, dado que supone un cambio de paradigma. El uso de parsimonia en la reconstrucción filogenética ha sido justificada utilizando una perspectiva hipotético-deductiva (Popper, 1968a, 1968b). En sistemática, ésta se basa en que dado que es imposible conocer cual es la filogenia real de los taxones, todo lo que podemos hacer es adoptar un criterio (parsimonia) que permita evaluar los méritos relativos de las hipótesis propuestas (cladograma), de forma que la hipótesis escogida sea la que exhiba una mayor congruencia en la distribución de las sinapomorfias, las cuales constituyen los tests de falsificación de la hipótesis (Eldredge y Cracraft, 1980) (ver Panchen, 1992; para una crítica a la validez de dicha conexión).

La utilización de máxima verosimilitud en filogenia fue introducida por Cavalli-Sforza y Edwards (1967) y, posteriormente, desarrollada y popularizada por Felsenstein (1981). La máxima verosimilitud es el método más general para la obtención de estimaciones estadísticas. A partir de la adopción de un cierto modelo de evolución (M) y de la observación de los caracteres de determinados taxones (D), la verosimilitud de un árbol (T) es la probabilidad (P) de las observaciones (caracteres), dado dicho árbol y el modelo asumido,  $P(D; T, M)$ , y considerada como función de los árboles (Felsenstein, 1988b). Dado un cierto árbol, la probabilidad del conjunto de datos ha de sumar 1. Sin embargo, cuando los datos se mantienen constantes y se evalúan distintos árboles, la suma de estos no tiene porque ser uno, por lo cual son consideradas verosimilitudes en lugar de probabilidades. El análisis consiste pues en hallar el árbol, o árboles, con un mayor valor de verosimilitud. En general, en la evaluación de cada uno de los árboles no solamente se tiene en cuenta la ramificación, sino también la longitud de las ramas y, en muchos casos, la estimación de parte de los parámetros del modelo adoptado (Yang, 1996). Un corolario del método es la existencia de un modelo de evolución subyacente a los datos. Si bien dichos modelos

han sido desarrollados para datos moleculares, éstos son inexistentes, y probablemente imposibles, para datos morfológicos, de forma que esta aproximación queda limitada a la utilización de datos moleculares. Ésta es una de las principales críticas que se le pueden hacer a esta metodología. No obstante, la crítica más importante y la más frecuentemente esgrimida en su contra, es la necesidad de adoptar explícitamente un modelo evolutivo, el cual no solo es absolutamente indemostrable, sino que, en el mejor de los casos, representa una simplificación de la realidad con efectos imprevisibles sobre el resultado final (Carpenter, 1992). Los defensores de ésta metodología esgrimen los resultados provenientes de simulaciones, que demuestran que, en muchos casos, se recupera el árbol correcto independientemente que el modelo escogido sea erróneo (Nei, 1996; Yang, 1996). Por otra parte, cabe destacar que si bien es cierto que el árbol con un máximo valor de verosimilitud, dado un cierto espacio de probabilidad de los parámetros, es el mejor estimador de la filogenia, deja de serlo el momento en que los árboles comparados pertenecen a distintos espacios de probabilidad (Nei, 1996). Esta situación se da cuando los parámetros del modelo no permanecen constantes en las comparaciones entre secuencias, es decir, cuando el proceso evolutivo no es constante, lo cual puede considerarse como extremadamente probable.

## 2.3. Análisis basado en matrices de distancias

Estos métodos constituyen el último remanente del feneticismo en sistemática. Se basan en la premisa de que la distancia entre taxones, entendida como una medida de su disimilitud, se relaciona directamente con su relación filogenética. Desgraciadamente, dicha asunción es únicamente válida en ausencia de homoplasia. Precisamente, y como se ha comentado anteriormente, la no distinción entre homologías y homoplasias determina que las metodologías feneticistas no sean aptas para la reconstrucción filogenética, al menos por lo que respecta a los datos morfológicos. Sin embargo, la posibilidad de utilizar modelos evolutivos a partir de los cuales puedan ser corregidas las estimaciones de las distancias en función de la homoplasia existente, les ha reservado un lugar en la inferencia filogenética basada en datos moleculares. A diferencia de las aproximaciones anteriores, basadas en caracteres discretos, estos métodos transforman, a través de la aplicación de cierta función, la matriz de caracteres en una nueva matriz, la matriz de distancias, en donde cada una de las celdas representa una medida del grado de disimilitud de los dos taxones implicados. La función para calcular la distancia entre los distintos taxones corrige, a su vez, el valor teniendo en cuenta la homoplasia existente, calculada mediante la adopción de un modelo evolutivo de los datos (Williams, 1992). El siguiente paso consiste en la construcción del árbol, la cual puede llevarse a cabo mediante dos métodos distintos: (1) métodos algorítmicos, los cuales establecen un determinado protocolo para ir uniendo taxones en función de sus distancias, p. ej. UPGMA o *Neighbor joining* (Saitou y Nei, 1987), (2) métodos basados en un criterio de optimización, los cuales permiten evaluar diferentes topologías p. ej. el método de los mínimos cuadrados (Cavalli-Sforza y Edwards, 1967) o el de la mínima evolución (Rzhetsky y Nei, 1992).

Dado que los métodos filogenéticos basados en matrices de distancia también requieren la adopción explícita de un modelo evolutivo, adolecen de similares limitaciones que los métodos de máxima verosimilitud.

### 3. ETAPAS DEL ANÁLISIS CLADÍSTICO BASADO EN PARSIMONIA

A continuación se comentan los pasos sucesivos en los que consiste el análisis de parsimonia, con las diferentes opciones existentes en cada caso y su problemática particular asociada. A su vez, se explican las diferencias entre la utilización de datos morfológicos o moleculares, entre los cuales únicamente se consideraran las secuencias nucleotídicas. Finalmente, se discuten las diferentes aproximaciones propuestas para tratar datos mixtos y, más concretamente, datos morfológicos y moleculares.

#### 3.1. Muestreo taxonómico

El primer estadio del análisis filogenético es la selección de los taxones que serán objeto de estudio. Si bien en algunos casos la selección de los taxones que constituyen el grupo interno puede resultar trivial, en otros, ya sea debido a la cantidad de taxones implicados, los límites difusos del grupo, etc., ésta puede no serlo en absoluto. En este sentido, se ha demostrado a través de simulaciones por ordenador que la adición de taxones a una matriz de datos resulta en un aumento de la precisión del cladograma (Wheeler, 1992). Es la selección del grupo externo la que suele plantear mayores dificultades. A pesar de que, contrariamente a lo que parece ser la idea más extendida, el grupo externo no tiene por que ser necesariamente el grupo hermano del grupo de estudio (Nixon y Carpenter, 1993), el conocimiento previo de las relaciones externas de algunos organismos es lo suficientemente pobre como para que la selección de éstos resulte problemática. Por otra parte, cuando se utilizan caracteres correspondientes a secuencias nucleotídicas, la adopción de un grupo externo muy divergente puede resultar en una asignación al azar de la raíz del grupo interno (*random outgroup*), generalmente en la rama interna que acumula un mayor número de cambios (Wheeler, 1990b). Hay casos en la que la función del grupo externo puede no ser tanto la de enraizar el árbol resultante como la de testar el posible estatus monofilético del mismo. La severidad de dicho test será función directa de la exhaustividad de la representación del grupo externo.

#### 3.2. Muestro de caracteres

Los caracteres representan la evidencia a partir de la cual se derivan las relaciones filogenéticas, la ausencia de dicha evidencia conduce a la incapacidad de caracterizar parte de los posibles grupos monofiléticos existente en el estudio. El efecto final sobre el árbol será la aparición de las denominadas politomías, nodos de los que se generan más de dos ramas. En cladismo se asume que las ramificaciones en el cladograma son siempre dicotómicas, dado que el proceso de especiación o cladogénesis da como resultado dos especies nuevas. Por otra parte, en un contexto hipotético-deductivo, las hipótesis más explícitas son siempre preferibles a las más ambiguas, ya que representan afirmaciones más susceptibles de ser rebatidas, es decir, más falsificables (Wiley, 1981). Además, un cladograma dicotómico puede resolver la ambigüedad o el conflicto en favor de la evidencia (Miyamoto, 1985). El tipo de politomía resultado de la ausencia de evidencia provocada por un muestreo pobre, o al menos no suficiente, de los caracteres se le denomina suave (*soft polytomy*). Ésta también puede ser debida a la presencia de conflicto entre los grupos apoyados por dos o más caracteres. Por el contrario, la denominada politomía dura (*hard polytomies*), es aquella en

que la ausencia de apoyo de una determinada agrupación se interpreta como un fenómeno real de especiación múltiple y, por tanto, son consideradas hipótesis filogenéticas legítimas. Sin embargo, en general se tiende a preferir la interpretación suave a la dura, ya que: "la experiencia demuestra que una politomía hoy ha desaparecido mañana" (Coddington y Scharff, 1996).

En general, la adición de nuevos caracteres a una matriz de datos resulta en un aumento de la resolución filogenética (Wheeler, 1992), por tanto, en principio, cuantos más caracteres mejor. Por otra parte, el número de caracteres necesario para resolver las relaciones entre los taxones dependerá directamente de su capacidad para agrupar dichos taxones, a la vez que de su congruencia. Un carácter que no apoya ninguna agrupación de taxones es no informativo. Entre éstos se distinguen los caracteres invariables, los cuales son monomórficos en los taxones estudiados, y los autapomórficos, que son exclusivos de un solo taxón. Cuando se utilizan datos morfológicos, dichos caracteres no informativos pueden ser detectados y evitados o eliminados antes de incluso construir la matriz. Por el contrario, cuando se trabaja con datos moleculares, éstos no pueden ser detectados hasta que se obtiene dicha matriz. En estos casos, la cantidad de caracteres informativos dependerá en gran medida de la tasa de cambio de la secuencia nucleotídica analizada, de aquí la importancia de la selección del fragmento de secuencia a analizar, el cual deberá mostrar un nivel de variabilidad adecuado para resolver las relaciones entre los taxones de estudio. Afortunadamente, los conocimientos acumulados a través del estudio de los procesos de evolución molecular proveen una información más o menos precisa de la utilidad de distintos genes para resolver diferentes niveles taxonómicos (Brower y DeSalle, 1994; Simon et al., 1994).

Finalmente, cuanto más diversa sea la procedencia de los caracteres, p. ej. morfológicos, diferentes genes moleculares, comportamentales, etc., mayor será el poder explicativo del árbol obtenido, ya que éste se habrá derivado a partir de una mayor parte de la evidencia (Kluge y Wolf, 1993). El tema del análisis combinado de distintos tipos de caracteres se aborda en más profundidad en apartados posteriores.

#### 3.3. Construcción de la matriz de caracteres: definición y codificación de los caracteres

Esta es, probablemente, la etapa más importante del análisis filogenético, ya que, por un lado, constituye la conexión entre las observaciones y el propio análisis y, por el otro, tiene una fuerte influencia sobre los resultados finales (Pleijel, 1995). Básicamente consiste en dos etapas: el establecimiento de las homologías primarias (De Pinna, 1991), el grado de corroboración de las cuales determinará el cladograma, distinguiendo las auténticas homologías (homologías secundarias) de las homoplasias, y de su codificación. A pesar de la importancia de este paso, en la actualidad parece ponerse una mayor atención en los resultados de la manipulación de la matriz de caracteres con diferentes programas de ordenador, enfatizando cualquier elucubración derivada de dicha matriz por encima de las observaciones que son el origen de la misma (Patterson y Johnson, 1997), olvidando que si los caracteres han sido erróneamente identificados o codificados todo de lo que ellos se desprenda será simplemente falso.

La delimitación de los caracteres se realiza a dos niveles: (1) su identidad topográfica y (2) la identidad de sus estados (Brower y Schawaroch, 1996). En el primero, se establecen el conjunto de características comparables entre los taxones. En el segundo, haciendo uso de los criterios clásicos para la

delimitación de homologías potenciales (De Pinna, 1991; Patterson, 1982; Patterson, 1988; Remane, 1952), las diferentes manifestaciones de los caracteres se clasifican como iguales (mismo estado) o diferentes. Operacionalmente, estos dos pasos corresponden a: (1) definir las columnas de la matriz de caracteres y (2) asignar un valor a cada celda. Cabe destacar que caracteres o estados de caracteres no son categorías absolutas y lo que en un estudio es un carácter puede convertirse en un estado en otro estudio a una escala jerárquica diferente. Bajo la perspectiva del establecimiento de las homologías primarias, el hecho de trabajar con caracteres morfológicos o caracteres moleculares plantea problemas completamente opuestos. Así, en el estudio de la morfología la delimitación espacial del carácter es relativamente fácil, mientras que la identificación de los diferentes estados, es decir, decidir que manifestaciones del carácter son genuinamente distintas y cuales pueden considerarse las mismas puede resultar extremadamente compleja. Por otra parte, dado que la identificación de los estados implica una discretización de los datos, la utilización de caracteres continuos resulta problemática (Chappill, 1989; Felsenstein, 1988a; Stevens, 1991). Los caracteres provenientes de las secuencias nucleotídicas tan solo tienen cinco manifestaciones, representadas por las cuatro bases nucleotídicas (adenina, citosina, guanina y timina) y las deleciones-insersiones, por lo cual la definición de los estados resulta trivial. Por el contrario, la asignación de las posiciones que corresponden al mismo carácter puede plantear un auténtico desafío. El proceso por el que se establece las identidades topográficas de las posiciones nucleotídicas se denomina alineación de las secuencias. En aquellas situaciones en las que los fragmentos de la secuencia de los distintos taxones tienen el mismo número de nucleótidos, como suele ser el caso en los genes codificadores de proteínas, la alineación no resulta ser un problema. Sin embargo, existen situaciones en que la longitud de las mismas puede ser distinta, debido a la existencia de mutaciones tipo *indel*, es decir, la inserción o deleción de algún nucleótido. La alineación de este tipo de secuencias pasa por reconocer donde han tenido lugar dichas mutaciones, incorporando espacios (*gaps*) con el objeto de conservar la homología posicional. Esta es la situación habitual cuando se trabaja con genes estructurales, como los genes ribosómicos, o fragmentos no codificantes. La alineación puede llevarse a cabo de dos formas distintas: manual o automáticamente. La alineación manual se basa en el reconocimiento visual de algunos motivos más o menos conservados y en la minimización de la incorporación de *gaps*. Una forma de mejorar la calidad de las alineaciones manuales es la incorporación, en caso que exista, de información estructural. Así, por ejemplo, los genes ribosómicos tienen estructura secundaria en la molécula transcrita, con zonas complementarias (*stems*) y zonas de cadena sencilla (*loops*). El reconocimiento de las zonas complementarias en la secuencia puede resultar de una gran ayuda en la determinación de las posiciones homólogas (Kjer, 1995). Sin embargo, hay ocasiones en que la divergencia entre las secuencias es tan grande que la asignación de homología se ve seriamente comprometida. Llegados a este punto, la mejor opción es simplemente eliminar estas zonas del análisis posterior. La alineación manual ha sido criticada debido a su falta de objetividad y de repetitividad, tanto en la asignación de posiciones como en la eliminación de las mismas en el análisis (Gatesy et al., 1993). Los alineadores automáticos se basan en la optimización de una cierta función de las coincidencias entre nucleótidos, las sustituciones entre éstos y la introducción de *gaps*. La inclusión de éstos últimos ha de penalizarse, ya que de lo contrario cualquier par de secuencias podría ser perfectamente alineada insertando suficientes *gaps* (Hillis et al., 1996).

Se ha propuesto que la alineación de las secuencias debería formar parte del propio proceso de inferencia filogenética (Sankoff et al., 1973). Esta aproximación ha sido implementada en algunos algoritmos de alineación automática (Wheeler, 1996; Wheeler y Gladstein, 1994), los cuales generan alineamientos que globalmente minimizan las diferencias entre las secuencias. A pesar de que, en principio, los alineadores automáticos ofrecen un criterio objetivo de alineación, la necesidad de incorporar en los mismos valores de parámetros tales como el coste de la sustitución entre los nucleótidos, o la penalización por la adición de nuevos *gaps* o de aumentar la longitud de los preexistentes, introduce de nuevo una cierta subjetividad. Sin embargo, quizás el punto más oscuro sea el referente al elevado coste computacional que éstos suponen, la cual comporta que en algunos casos sea inviable obtener una alineación fiable.

La codificación de los caracteres consiste en la asignación de un símbolo alfanumérico a cada uno de los estados de carácter definidos previamente. Los estados de los caracteres morfológicos suelen codificarse con números, ya sea empezando por 0 o 1, mientras que para los nucleótidos se utilizan las iniciales de la base correspondiente (A, C, G y T) o el signo “-” cuando se trata de *gaps*. Atendiendo al número de estados, los caracteres pueden ser binarios, si tan solo tienen dos estados, o multiestados, si tienen tres o más. Evidentemente, los caracteres derivados de secuencias nucleotídicas serán siempre multiestado. La codificación, que a primera vista puede parecer simple o incluso trivial, puede convertirse en una operación harto compleja. Especialmente conflictivo resulta el tratamiento de los denominados datos inciertos (*missing data*). Éstos corresponden a situaciones donde no existe ningún estado que permita definir una determinada situación en alguno de los taxones de estudio. Esto puede deberse a tres razones (Platnick et al., 1991b): (1) el estado es inexistente, es decir, el taxón en cuestión tiene uno de los estados definidos, pero éste no es observable en el individuo concreto que se está estudiando, ya sea debido a una mal estado de conservación, porque se trata de un estado de desarrollo donde el carácter todavía no se ha manifestado, porque unas determinadas posiciones nucleotídicas no han podido ser secuenciadas, etc., (2) el carácter es inexistente, es decir, el carácter definido no existe en el taxón, p.ej. color de las plumas en mamíferos, y (3) el carácter es polimórfico, es decir, diferentes individuos pertenecientes al mismo taxón presentan estados distintos. A pesar que cada una de estas situaciones es ciertamente diferente, son agrupadas a menudo bajo la misma codificación, “interrogante”, con la esperanza de que al no serle asignado ninguno de los estados definidos, dicha entrada no afecte el resultado final del análisis. Para este carácter en concreto, el taxón podrá ser asignado indistintamente a cualesquier de los grupos definidos por los estados sin variar el número de pasos final del cladograma. Si bien el caso en que el estado sea desconocido esto no plantea mayores problemas, ya que puede ser evitada estudiando otro individuo o, en cualquier caso, la codificación como interrogante refleja exactamente lo que ocurre, esto no es así en las restantes situaciones. Asignar un interrogante cuando el carácter en cuestión es polimórfico en el taxón pasa por alto la existencia de una transformación entre estados (Nixon y Davis, 1991). Ello supone que el número de pasos del árbol final será subestimado y, por tanto, que dicho árbol podría no ser el más parsimonioso si el carácter hubiese sido codificado con un estado concreto. Las soluciones propuestas a éste dilema son: (1) intentar establecer a partir de fuentes de evidencia adicionales cual sería el estado derivado para el taxón polimórfico (Wiley et al., 1991) o (2) dividir el taxón en tantos nodos terminales como estados se observan.

Los caracteres inexistentes plantean una problemática distinta. Debido a que durante el análisis se asigna a cada uno de los nodos internos el estado del carácter más parsimonioso, al nodo que lleva a taxones que carecen del carácter se les asigna igualmente un estado, a pesar de que el mismo es imposible (Platnick et al., 1991b). Éste hecho puede conducir a la obtención de resultados no esperados ni deseados al analizar los datos. La razón es que, dependiendo del estado que haya sido asignado, el número de pasos final puede variar, descartando resoluciones por ser menos parsimoniosas cuando en realidad lo son igualmente o, lo que es lo mismo, escogiendo soluciones que pueden no ser las más parsimoniosas (Maddison, 1993). Conviene destacar que este efecto no se da ni cuando tan solo hay un taxón terminal con el carácter inaplicable, ni cuando todos los taxones inaplicables forman un clado exclusivo. En realidad el fenómeno comentado es un caso particular de un problema mucho más amplio que tiene que ver con la definición de los caracteres y de sus estados. El siguiente ejemplo ayudará a entender este punto (modificado de Pleijel, 1995). Dada la variabilidad mostrada en la Fig. 7, se podrían definir los caracteres hasta de cuatro maneras diferentes (A-D).

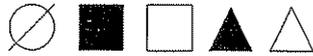


Fig. 7.- Diferentes manifestaciones de un carácter hipotético: ausente, cuadrado negro, cuadrado blanco, triángulo negro, triángulo blanco.

**A** (o codificación compuesta).

1. Carácter X: ausente (0); presente, cuadrado y negro (1); presente cuadrado y blanco (2); presente, triangular y negro (3); presente, triangular y blanco (4).

**B.**

1. Carácter X: ausente (0); presente, cuadrado (1); presente, triangular (2).
2. Carácter Y: ausente (0); presente, negro (1); presente blanco (2).

**C.**

1. Carácter X: ausente (0); presente (1).
2. Carácter Y: ausente (0); triangular (1).
3. Carácter Z: negro (0); blanco (1).

**D** (o codificación reductiva).

1. Carácter V: ausente (0); presente (1).
2. Carácter W cuadrado: ausente (0); presente (1)
3. Carácter X triangular: ausente (0); presente (1)
4. Carácter Y negro: ausente (0); presente (1)
5. Carácter Z blanco: ausente (0); presente (1)

Se puede observar fácilmente que el tipo de definición de carácter C introduce caracteres inexistentes en la matriz. Por tanto, la utilización de cualquier de los otros tipos evitaría el problema. Sin embargo, la elección de un tipo u otro de definición se basa en diferentes criterios, y todos ellos pueden inducir a algún tipo de error en el resultado. Entre los criterios que pueden utilizarse para escoger una forma u otra, dejando a parte los *missing data*, se destacan (Pleijel, 1995; Wilkinson, 1995a):

1. Independencia de los caracteres. Como se ha comentado anteriormente, y por las razones aducidas, la parsimonia asume que los caracteres son independientes. Bajo esta perspectiva, la codificación compuesta (A) es la que

minimiza la posibilidad de dependencia los caracteres. Por contra, en el tipo B y, de forma todavía más exagerada, en la codificación reductiva (D), el estado ausente se repite en algunos caracteres, cuando, obviamente, se trata de la misma característica. El efecto de esta multiplicación del estado es que los grupos apoyados por el mismo serán artificialmente favorecidos frente al resto

2. Independencia de la jerarquía. La homología de los caracteres es un concepto relativo que depende directamente de la escala filogenética que se este estudiando. Así, una determinada característica que a un cierto nivel está o no presente, puede tener diferentes manifestaciones si se observa a una escala más fina. La codificación reductiva (D) es la única que evita tener que redefinir los caracteres cada vez que se añaden a la matriz nuevos taxones con diferentes combinaciones de los mismos.
3. Recuperación y contraste de la información. Este criterio tiene que ver con la eficiencia de cada uno de los tipos de codificación para incorporar las observaciones sin pérdida de datos, a la vez que a la forma más apropiada para testar los mismos a través de la congruencia. Teniendo en cuenta, que la congruencia se aplica entre los caracteres y no entre los estados de los mismos que son fijados de antemano, las codificaciones A-C suponen una pérdida de la capacidad de contraste de los datos. Contrariamente, ésta se maximiza en la codificación reductiva, donde toda la variabilidad observada podrá ser contrastada. Por otro lado, la existencia de caracteres múltiples en las codificaciones A-C resulta en un menor grado de información de los mismos, ya que el apoyo a un grupo determinado por parte de alguno de los estados dependerá de la situación del resto de los mismos, lo cual no ocurre en la codificación reductiva.
4. Reconstrucción de los estados de los antepasados. En la codificación compuesta, tan solo uno de los estados definidos puede ser asignado a un nodo interno o antepasado. En cambio, si se utiliza la codificación reductiva, éste carácter puede separarse en varios, de forma que el antepasado puede presentar una combinación no observada en ninguno de los taxones terminales.

Como se puede comprobar, existen criterios a favor y en contra de cada uno de estos métodos de definición de los caracteres. Si bien las codificaciones más reductivas muestran ventajas en cuanto a contraste estabilidad y reconstrucción evolutiva, pueden afectar negativamente el análisis sobrevalorando algunos de los caracteres y los grupos que estos apoyan.

En el contexto de los datos moleculares, un problema similar al representado por los caracteres inaplicables o inexistentes se da en el tratamiento de los *gaps*. Así, se observa que en muchos análisis moleculares publicados los *gaps* son tratados como *missing data*, es decir, asignándoles un interrogante. Sin embargo, asumiendo que la incorporación de estos *gaps* a las secuencias ha sido el resultado de su alineación, es decir, del establecimiento de las hipótesis de homología posicional, parece ilógico no considerarlos como un quinto estado adicional a las cuatro bases nucleotídicas (Wheeler, 1993). Por otra parte, tratar los *gaps* como un quinto estado tiene el inconveniente de que cada una de las posiciones de una inserción-delección múltiple (que incluye más de un *gap*) es considerada como un carácter independiente, cuando ciertamente pueden no serlo (Maddison y Maddison, 1992), lo cual resultará en una sobrevaloración de los grupos apoyados por los mismos.

Dejando a parte las causa por las que pueden introducirse interrogantes en la matriz de caracteres, su presencia puede tener efectos negativos en el análisis filogenético (Nixon y Davis, 1991; Platnick et al., 1991b). Uno de estos efectos es la multiplicación del número de árboles obtenidos. Debido a la ausencia en los taxones incompletos de caracteres que pueden estar apoyando alguno de los cladogramas, la asignación de estos taxones a uno u otro grupo puede no variar el número de pasos final. A pesar de que a primera vista la solución más simple podría parecer la de eliminar los taxones incompletos, tal acción podría derivar en una alteración importante de las relaciones filogenéticas inferidas. Este sería el caso si, por ejemplo, el taxón eliminado presenta en los caracteres conocidos combinaciones de estados ausentes en los taxones restantes. (Wilkinson, 1995b).

**3.4. Medidas del grado de ajuste entre caracteres y árboles.**

Diferentes índices han sido descritos y utilizados en la literatura para reflejar la relación entre caracteres y árboles, en términos de ajuste de los primeros sobre los últimos.

El conocido como índice de consistencia (*ic*) (Kluge y Farris, 1969), es el cociente entre el número de pasos mínimo de un carácter (*m*), que es igual a su número de estados menos uno, y el número de pasos del carácter observado en el cladograma (*s*):

$$ic = m / s, IC = \sum m / \sum s$$

En el caso que el ajuste sea perfecto, el número de pasos observado será igual al mínimo teórico y el valor del índice será 1. Cuanto peor sea el encaje entre el carácter y el árbol más pasos adicionales mostrará el carácter sobre el árbol, y su índice bajará, aunque nunca llegará a ser 0. El índice de consistencia conjunto de un árbol (*IC*) es el cociente entre la suma de los valores particulares de *m* y *s* de cada carácter. Debido a que las situaciones en que el número de paso "reales" excede el mínimo teórico se explican por la presencia de homoplasia, el *IC* ha sido considerado como una medida de la misma, aunque la influencia de ciertas variables sobre el mismo comprometan su valor comparativo entre diferentes análisis. Así, el valor de *IC* está correlacionado con el número de taxones, de forma que las matrices con mayor número de taxones tienden a mostrar valores de *IC* menores; sin embargo, se ha desarrollado una fórmula empírica que permite calcular el valor de *IC* esperado dado un cierto número de taxones (*n*):

$$IC_{esperado} = 0,90 - 0,022n + 0,000213n^2, \text{ para } n < 60$$

(Sanderson y Donoghue, 1989)

El *IC* es también sensible a la proporción de caracteres ordenados de una matriz. Los caracteres ordenados, en general, encajan peor en la definición de las agrupaciones que los desordenados porque son más restrictivos. La presencia de *missing data* es otra fuente de distorsión, ya que éstos no incrementan nunca la inconsistencia de los caracteres. Finalmente, los caracteres autapomórficos o invariables incrementan artificialmente el *IC*, ya que no apoyan ninguna agrupación y, por tanto, siempre encajan perfectamente. Si bien cuando se trabaja con datos morfológicos dichos caracteres no suelen ya introducirse en las matrices, esto es inevitable cuando se utilizan datos moleculares, por lo que este tipo de caracteres debe eliminarse antes de computar el valor del *IC*. En otro orden de cosas, debe mencionarse que Klassen et al. (1991) han demostrado que, para que una matriz de datos sea filogenéticamente informativa, su *IC* debe ser mayor que el *IC* de una matriz aleatorizada del mismo tamaño (*n*). Este valor puede obtenerse de la siguiente fórmula:

$$IC_{random} = 2.937 \times n^{-0,9339} \text{ (Klassen et al., 1991)}$$

El denominado índice de retención (*ir*) (Farris, 1989) fue originalmente propuesto con el objetivo de evitar algunos de los problemas asociados al *IC* y expresa la proporción de similitudes en el árbol que son explicables como sinapomorfías. En este caso, la aproximación se realiza a través de la evaluación de la cantidad de sinapomorfías en la matriz. Así, el *ir* es el complementario del valor de homoplasia relativa presente en la matriz respecto a la máxima posible. La fórmula para su cálculo es:

$$ir = (g - s) / (g - m), IR = (\sum g - \sum s) / (\sum g - \sum m)$$

donde *g* es el número máximo de transformaciones posibles de un carácter en cualquier árbol, que, en la práctica, es el número de pasos dada una politomía total (Wiley et al., 1991). Al valor *g - m* se le denomina variabilidad informativa (Farris, 1991). Se puede observar que cuando *s* es igual a *m*, es decir, en ausencia de homoplasia, el *ir* es 1. Cuanto mayor sea la homoplasia del carácter, menor será su valor de *ir*. El índice de retención de un árbol (*IR*) es la suma de los valores *g*, *s* y *m* para cada carácter. Valores altos de *IR* indican que los cambios se acumulan preferentemente en las ramas internas, mientras que los valores bajos sugieren que los cambios se acumulan en ramas terminales (Siebert, 1992). El *IR* es pues, una buena medida del apoyo, basado en la evidencia, de los grupos y, además, no es sensible ni a la presencia de autapomorfías ni de caracteres invariables, ya que en ellos *g* es igual a *s* y, por tanto, el *ir* es 0. Finalmente, el índice reescalado de consistencia (*CR*) (Farris, 1989) es el resultado de multiplicar los dos índices anteriores. Su principal virtud es que permite discriminar diferencias en el valor de *g* en caracteres que de otra forma muestran niveles similares de homoplasia (Quicke, 1993).

**3.5. Ponderación de los caracteres y de los cambios entre estados**

Toda evidencia comparativa tiene un valor potencial en la inferencia filogenética. Ahora bien, ¿son todos los caracteres igualmente válidos como indicadores de las relaciones filogenéticas? o, de la misma forma, ¿son todas las transformaciones entre los estados igualmente importantes evolutivamente? La opinión más generalizada es que ni todos los caracteres aportan la misma evidencia, ni todas las transformaciones entre estados son equivalentes. (Farris, 1969; Wheeler, 1990a; Williams y Fitch, 1989). Acorde a esta premisa, será necesario introducir en el análisis las posibles consideraciones, hipótesis o asunciones, sobre el valor relativo tanto de los caracteres como de las transformaciones de sus estados. Con el objeto de facilitar la discusión, nos referiremos al valor relativo asignado a un determinado carácter frente a los restante como el peso (*weight*) del carácter, y al valor del cambio de un estado particular a otro como el coste de la transformación.

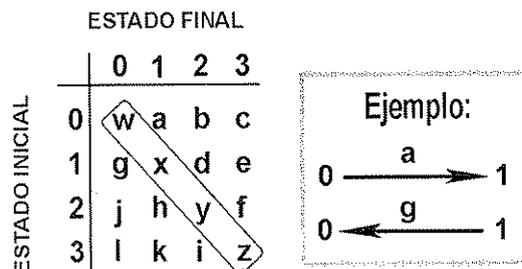


Fig. 8.- Ejemplo de matriz de costes.

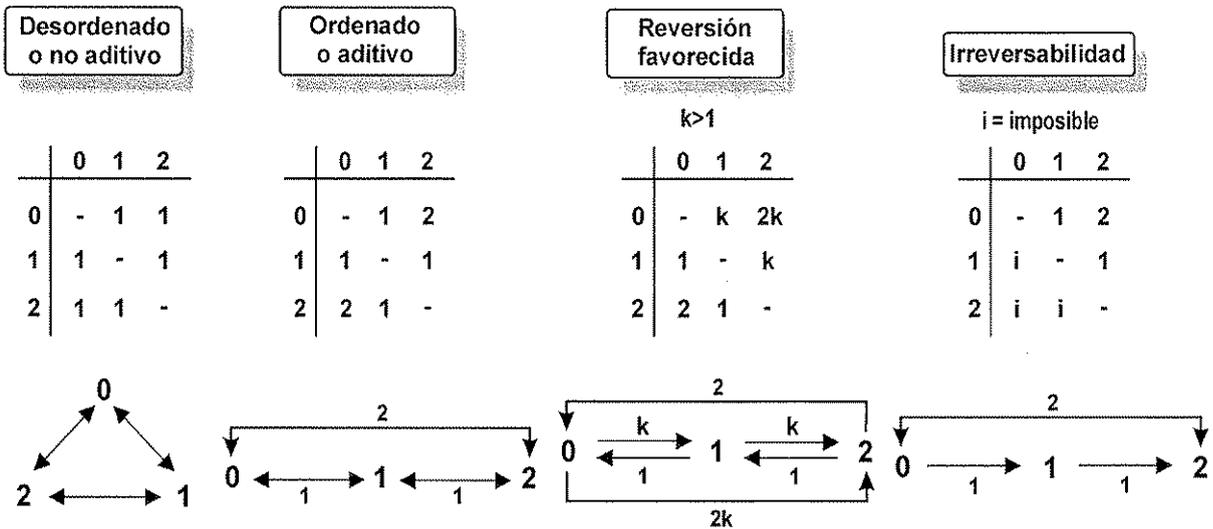


Fig. 9.- Algunas relaciones apriorísticas entre los estados de un carácter.

Conviene destacar que, algunos autores, a pesar de aceptar que no todos los caracteres tienen la misma importancia como indicadores de las relaciones filogenéticas, consideran que la ponderación diferencial de los caracteres es una forma de incorporar al análisis modelos evolutivos que representan asunciones que lo alejan de sus bases empíricas, y desaconsejan, por tanto, su utilización (Kluge, 1997a; Kluge, 1997b; Siebert, 1992). Contra esta opinión, se ha esgrimido que la no ponderación de los caracteres no es más que una forma particular de ponderación, la ponderación uniforme, la cual no deja de ser igualmente una asunción externa al análisis (Wheeler, 1986). En cualquier caso, el problema principal consiste en determinar el valor concreto de los pesos y/o costos a asignar. En este sentido, una justificación adicional de la utilización de una ponderación uniforme es la ausencia de evidencia objetiva para determinar el valor exacto de un determinado carácter o estado respecto a los restantes. La ponderación de caracteres y cambios de estado puede ser *a priori*, si se basa en asunciones previas al análisis, o *a posteriori*, si la evidencia se deriva a partir de un análisis precedente.

### 3.5.1. Coste de las transformaciones

Las relaciones entre los estados de un carácter, es decir, el coste de todos los cambios posibles entre los mismos, puede representarse en forma de una matriz de costes o matriz de pasos (*step-matrix*) (Sankoff y Rousseau, 1975). La Fig. 8 ilustra un ejemplo de matriz de costes. En filas se representa el estado inicial, previo al cambio, del carácter, y en columnas el estado final, resultado de la transformación. En las celdas figura el valor numérico concreto asignado a la transformación entre dos estados particulares. Estas matrices son simétricas si la transformación de un determinado estado *i* a un estado *j* tiene el mismo coste que el cambio inverso. En caso contrario, la matriz resultante será asimétrica. Bajo el criterio de parsimonia, tan solo los estados de carácter que definan grupos son informativos. Esto se refleja en la matriz de costos no asignando ningún valor a la transformación de un estado en sí mismo. Sin embargo, los autores que defienden la adopción explícita de un modelo evolutivo, proponen obtener la matriz de costes a partir de la asignación de probabilidades tanto del cambio como de mantenimiento del mismo estado, pudiendo, por tanto, asignar costes a la diagonal (Maddison y Maddison, 1992). Cabe mencionar, que las matrices de costes derivados de modelos probabilísticos convierten a la reconstrucción más parsimoniosa de los estados ancestrales, en la máxima probabilidad Bayesiana

de la estima de dichos estado (Maddison, 1990; Maddison y Maddison, 1992). A partir de esta propiedad y en el contexto de los datos moleculares, se ha sugerido la posibilidad de calcular la matriz de costos mediante la estima de máxima verosimilitud de los cambios entre estados (Yang, 1994). La principal ventaja de esta aproximación es que, computacionalmente, la implementación de la parsimonia es mucho más eficiente que la de la máxima verosimilitud, permitiendo búsquedas mucho más exhaustivas de los árboles óptimos (Swofford et al., 1996a).

A pesar que, contrariamente a lo anteriormente expuesto, se prefiera no adoptar de manera explícita un modelo probabilístico de evolución de los caracteres, todavía pueden ser incorporadas en el análisis ciertas restricciones sobre los cambios entre estados, basándose en conocimientos previos acerca de sus características. En el caso más simple, que correspondería al completo desconocimiento de las relaciones entre los estados, cualquier estado puede transformarse libremente en cualquier otro. Los caracteres tratados de esta manera se denominan desordenados o no aditivos. El algoritmo para el cálculo del número de pasos de un árbol cuando los caracteres son desordenados fue originalmente propuesto por Fitch (1971), y a su aplicación se le conoce con el nombre de parsimonia de Fitch. Otra relación entre los estados puede ser la representada por una serie de transformaciones lineales, de forma que el cambio entre estados contiguos sea menos costoso que entre estados que no lo son. Un ejemplo de aplicación de esta asunción sería la de un carácter definido como "pilosidad", donde los estados fuesen: "glabro", "poco piloso", "muy piloso". No sería descabellado plantear la posibilidad de que la transformación de "muy piloso" a "glabro", pase necesariamente por la presencia de un estado "poco piloso", y que, por tanto, la transformación de "muy piloso" a "glabro" sea el doble de costosa que las transformaciones de "muy piloso" a "poco piloso" o de "glabro" a "poco piloso". Los caracteres que incorporan este tipo de asunciones se denominan ordenados o aditivos. Los caracteres binarios representan el caso más simple de caracteres ordenados. Farris (1970) y Kluge y Farris (1969) propusieron el algoritmo para el cálculo del número de pasos de los árboles más parsimoniosos bajo ésta premisa, la cual se conoce con el nombre de parsimonia de Wagner. Dos casos particulares de ésta última, son las denominadas parsimonia de Dollo (Farris, 1977) y parsimonia de Camin-Sokal (Camin y Sokal, 1965). En el primer caso, a la restricción que los caracteres sean ordenados se le añade la de favorecer las reversiones de las transformaciones sobre su aparición independiente, lo cual se consigue haciendo más

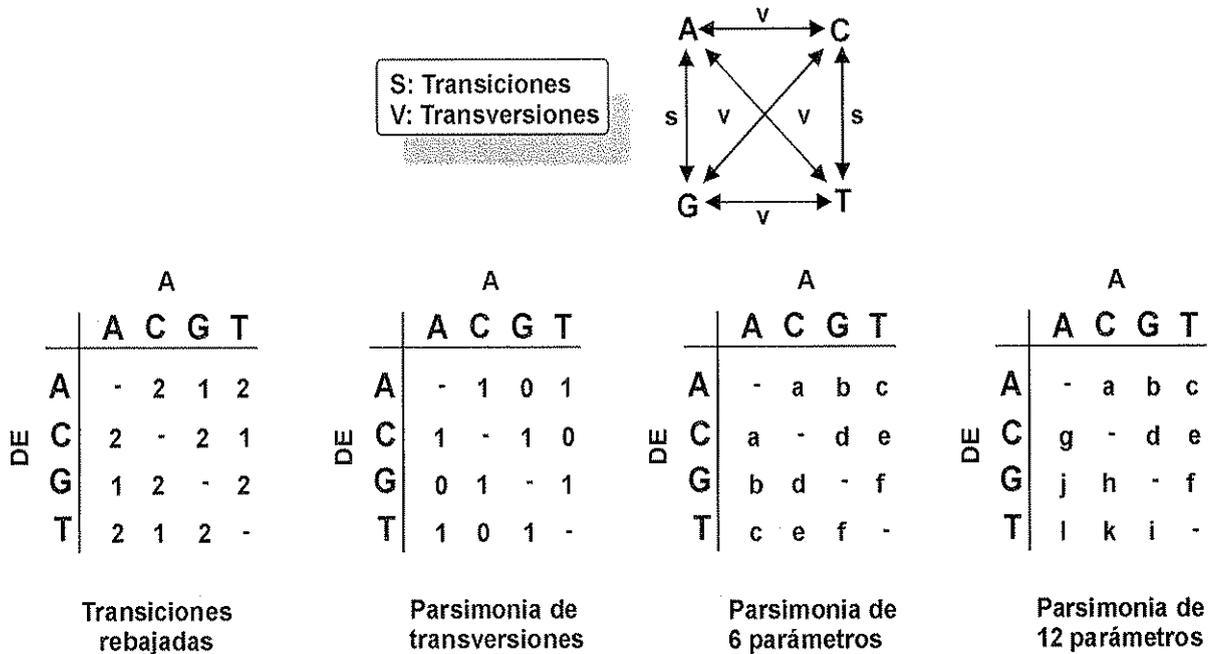


Fig. 11.- Tipos de transformaciones entre estados nucleotídicos y ejemplos de matrices de costos en datos moleculares.

costosas las transformaciones en un sentido que en el contrario (Fig. 9). El análisis de los caracteres derivados de las dianas de restricción, constituyen un ejemplo de aplicación de estas asunciones. La pérdida de una diana de restricción, que supone como mínimo el cambio de uno de sus nucleótidos, será generalmente más fácil de observar que la ganancia de una diana, en el peor de los casos requerirá la transformación de cuatro o seis, dependiendo del tipo de enzima, nucleótidos contiguos. La parsimonia de Camin-Sokal asume la ordenación de los caracteres, el conocimiento sobre su polaridad, es decir, el establecimiento de los estados apomórficos, y la irreversibilidad de los cambios. Esta última premisa se incorpora a la matriz prohibiendo la transformación en uno de los sentidos (Fig. 9). Este tipo de parsimonia es raramente utilizada, debido a la dificultad de justificar la irreversibilidad, tanto en datos morfológicos como moleculares (Kitching, 1992).

Las diferentes asunciones enumeradas constituyen de hecho casos particulares de lo que se ha denominado parsimonia generalizada (Swofford et al., 1996a). El desarrollo de algoritmos capaces de calcular exactamente el número de pasos total de una matriz de costos (Sankoff y Cedergren, 1983), permite incorporar al análisis una amplísimo abanico de modelos evolutivos mediante la asignación de un valor de coste a un tipo particular de transformación. En cualquier caso, la adopción de un determinado modelo de evolución, restricción o ponderación sobre las transformaciones de los estados, deberá ser debidamente justificado, a la vez que formulada de manera explícita y clara (Wheeler, 1986).

A diferencia de los caracteres morfológicos, los caracteres derivados de secuencias nucleotídicas tienen un número limitado de estados posibles, el cual es potencialmente compartido por todas las posiciones. En este tipo de datos la principal fuente de homoplasia es la existencia de sustituciones múltiples en una misma posición nucleotídica y, a su vez, la probabilidad que éstas hayan tenido lugar está directamente relacionada con la frecuencia de cambio de la posición. Asimismo, cuanto más frecuente sea un tipo concreto de transformación (=sustitución), más probable será que ésta esconda la existencia de cambios múltiples y, también, que haya tenido lugar en distintos linajes independientemente. Por tanto, la resolución filogenética de este tipo de datos puede

potenciarse si se da un mayor coste a las transformaciones menos frecuentes (Williams y Fitch, 1989). Una forma de poner en evidencia la posible existencia de sustituciones múltiples es mediante las denominadas curvas de saturación (Fig. 10), las cuales no son más que gráficos donde se representa en abscisas el número de cambios de una determinada clase de transformación presente entre todas las comparaciones de pares de taxones, y en coordenadas una determinada medida del nivel de divergencia entre los mismos. En ausencia de sustituciones múltiples, el número de cambios aumentará a medida que aumente la divergencia de los taxones, con lo que los puntos se ordenarán siguiendo poco más o menos una línea diagonal. Sin embargo, en ciertos casos puede observarse como llegado a un determinado nivel de divergencia, el número de cambios deja de aumentar para pasar a permanecer constante, independientemente de que la divergencia siga aumentando. Esta situación es consecuencia de que los nuevos cambios que tienen lugar lo hacen sobre posiciones que ya habían cambiado anteriormente, de forma que no serán contabilizadas en el computo de cambios global. La información sobre la frecuencia relativa de un determinado tipo de transformación puede obtenerse a partir del conocimiento previo sobre como evolucionan las secuencias, o bien, a partir de la observación de las características de las secuencias en el estudio particular (Knight y Mindell, 1993).

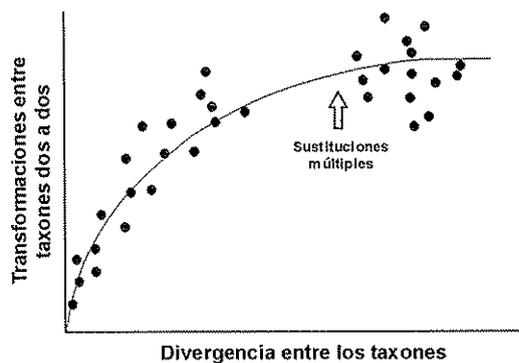


Fig. 10.- Ejemplo de curva de saturación.

Las transformaciones entre estados nucleotídicos suelen dividirse en dos tipos (Fig. 11). Los cambios de base de la misma naturaleza química, es decir, entre pirimidinas (timina y citosina) o entre purinas (adenina y guanina), se denominan transiciones (s), mientras que los cambios entre bases de naturaleza diferente (purina a pirimidina o al contrario) reciben el nombre de transversiones (v). A pesar que potencialmente el número de transversiones posibles es el doble que el de transiciones, éstas últimas suelen darse con más frecuencia que las anteriores, especialmente en el ADN mitocondrial (Brown et al., 1982; Li et al., 1984). Por esta razón, se ha propuesto que las transiciones reciban un coste inferior al de las transversiones (Kocher et al., 1989) (Fig. 11). El problema es decidir que valor exacto de coste es el correcto o, al menos, establecer un criterio objetivo de evaluación; aunque se observa que en la mayoría de estudios la elección de dicho valor es muy arbitraria. Una aproximación habitual es la de aplicar diferentes esquemas de coste y comprobar su efecto sobre el análisis p. ej.: 2:1, 3:1, 5:1, 10:1). Un método más objetivo consiste en utilizar la tasa de acumulación de las transiciones respecto a las transversiones en comparaciones entre pares de taxones no afectadas por las sustituciones múltiples, es decir, en la incluidas en la zona de crecimiento lineal. El valor de coste relativo se obtiene calculando la pendiente de la zona de crecimiento conjunta a partir de un modelo II de regresión lineal (Sturmbauer y Meyer, 1992). La situación extrema de rebajar las transiciones respecto las transversiones es la denominada parsimonia de transversión (Fig. 11), en donde las primeras son completamente eliminadas del análisis asignándoles un coste de 0. Esta aproximación comporta generalmente una pérdida importante de resolución en las zonas terminales del árbol, ya que en muchos casos las transiciones son los únicos cambios que se producen entre taxones evolutivamente muy cercanos.

Conviene destacar que, en algunas ocasiones, la asunción de que la tasa de acumulación de transiciones es mayor que la de transversiones puede resultar falsa. Así, por ejemplo, se ha demostrado que en algunos genes ribosomales la proporción de los dos tipos de sustituciones es similar (Vawter y Brown, 1993) o, como pasa en genes ribosomales mitocondriales, la mayoría de cambios corresponden a transversiones (DeSalle, 1992). Además, en genomas con un sesgo importante en la composición de bases en favor de A y T, el número de transversiones AT puede superar a cualquier otro tipo de transformación (Dowton y Austin, 1997). Por otra parte, la separación de las sustituciones entre bases en solamente dos clases de cambio, puede estar escondiendo relaciones muy diferentes entre las diferentes transformaciones. Una manera de evitar el posible sesgo derivado de la agrupación de los cambios en tan solo dos clases, es simplemente aumentando el número de clases consideradas, ya sean los 6 tipos distintos de transformación (Fig. 11) (parsimonia de seis parámetros, Cunningham, 1997b) o añadiendo la asimetría potencial de los cambios (p.ej.  $A \rightarrow T$  distinto de  $T \rightarrow A$ ), estableciendo hasta 12 clases posibles (Fig. 11) (parsimonia de doce parámetros, siguiendo la nomenclatura anterior).

Se han propuesto un conjunto de metodologías para derivar de una manera más o menos objetiva el coste de los distintos tipos de transformación, a partir de la información de los datos originales, con lo que se evita tener que asumir un modelo general de evolución para todas las moléculas y los taxones (Wheeler, 1990a). Entre estos se destacan el de los costes combinatoriales (Wheeler, 1990a) y el de los costes EOR (Knight et al., 1993). El primero está basado en la frecuencia con que determinados nucleótidos aparecen conjuntamente en las posiciones variables, de forma que,

cuanto mayor sea ésta más probable será la de que dos se hayan intercambiado. Operacionalmente, el método consiste en asignar un valor ( $a_{ijk}$ ) a cada coincidencia de dos nucleótidos ( $i, j$ ) en una posición derivado de la fórmula (de ahí el nombre de costes combinatoriales):

$$a_{ijk} = (n_k - 1) / C_n^2$$

donde  $n_k$  = número de nucleótidos distintos en la posición  $k$ , y  $n$  = número total de nucleótidos. Sumando los valores de todas las posiciones donde aparecen conjuntamente dos nucleótidos determinados, obtendremos una matriz simétrica de las asociaciones entre los cuatro nucleótidos. El siguiente paso es normalizar la matriz, dividiendo cada valor por la suma de los valores de la columna a la que pertenece. La finalidad de esta normalización es representar las frecuencias de transformación de un nucleótido concreto a cada una de las alternativas. Asimismo, esta normalización introduce asimetría en la matriz, el valor de los cambios entre los nucleótidos es diferente en un sentido que en el contrario, de forma que se consideran las 12 clases diferentes de sustituciones. Finalmente, para pasar de la matriz de asociaciones (T) a la matriz de costes (W), se aplica un transformación logarítmica:

$$W_{ij} = |\ln T_{ij}|$$

Cabe destacar que, la normalización originalmente propuesta ha sido criticada argumentando que ésta se corresponde en la práctica a una probabilidad condicionada. Para obtener una auténtica probabilidad total, el denominador no debe ser la suma sólo de las posiciones variables, sino la de todas las posiciones, variables e invariables (Rodrigo, 1992).

El método EOR (*expected/observed ratio*) propone una aproximación diferente. La corrección no tiene en cuenta tan solo la frecuencia de aparición de dos bases concretas, sino también la composición nucleotídica de la secuencia. Para cada pareja de taxones se calcula el "número esperado" de cada tipo de cambio, multiplicando el número total de transiciones o transversiones (según el tipo de transformación considerada) que presentan por la suma de las medias de las proporciones de cada una de las bases implicadas. El coste se deriva directamente del cociente entre el valor obtenido de la manera indicada y el número de cambios "observado". Así, las sustituciones que se den con una frecuencia superior a la esperada recibirán un coste menor que el de aquellas que se den más raramente. Los puntos débiles de esta aproximación son: (1) la ausencia de un modelo explícito del proceso aleatorio de donde derivar el valor "esperado" y (2) que los costes calculados para las diferentes bases tienen diferentes escalas y, por tanto, su combinación en una misma matriz es incorrecta (Collins et al., 1994). Una crítica adicional, la cual puede hacerse extensiva a otros métodos basados en comparaciones entre pares de taxones, es que los valores así obtenidos no son evolutivamente independientes, de forma que los cálculos derivados estarán sesgados por la filogenia subyacente (Collins et al., 1994; Swofford et al., 1996a). Una manera de evitar este sesgo, es reconstruir los estados ancestrales a partir de un árbol previo. La denominada ponderación dinámica (*dynamic weighting*) (Fitch y Ye, 1991; Williams y Fitch, 1989; Williams y Fitch, 1990), deriva la matriz de costes de la frecuencia de las diferentes sustituciones calculadas a partir de un árbol inicial. A pesar de no ser un requerimiento del método, este árbol inicial es generalmente el árbol más parsimonioso bajo ponderación uniforme. Se han propuesto dos tipos de función para transformar la matriz de frecuencias en una matriz de costes: lineal, es decir, utilizando el inverso de la frecuencia, o cuadrático, el inverso del cuadrado de la misma y, por tanto, mucho más extremo. A diferencia de otros métodos, éste permite incorporar también pesos diferencia-

les a los caracteres. El razonamiento es exactamente el mismo que el utilizado para los costes, con la única diferencia de que, en lugar de a través de una matriz, los pesos se incorporan al análisis en forma de vector. Esta aproximación es iterativa, de manera que tras calcular la matriz de costes y el vector de pesos, se realiza un nuevo análisis, del cual se derivaran costes y pesos, y así consecutivamente hasta que el árbol obtenido sea igual al árbol precedente o hasta que se hayan realizado un número predeterminado de repeticiones (generalmente 20), sin que se haya conseguido una estabilización de los resultados. La principal crítica a este método y, en general, a todos aquellos que derivan las frecuencias de cambio a partir de un árbol previo, es que los resultados son dependientes del árbol seleccionado. Un problema adicional de la ponderación dinámica es su elevado coste computacional, aunque pueden implementarse versiones simplificadas del mismo (Maddison y Maddison, 1992).

Finalmente, debe destacarse que ciertos valores de una determinada matriz de costes pueden conducir a la violación de la conocida como desigualdad triangular. Ésta se define como:

$$d_{ij} = d_{ik} + d_{kj}$$

donde  $i, j, k$  representan los estados de un carácter y  $d$  la distancia o coste de su transformación. El no cumplimiento de esta desigualdad puede resultar en inconsistencias lógicas. Por ejemplo, suponiendo que los estados sean A, C, G con costes: AC=3, AG=1, CG=1, se comprueba que  $3 > 1+1$  y que, por tanto, violan la desigualdad triangular, lo que implica que si en una cierta posición solo hay A y C, la reconstrucción más parsimoniosa es aceptar que el estado ancestral era una G, cuando resulta ilógico asignar a un antepasado un carácter no observado (Wheeler, 1993). La violación de la desigualdad no solo puede darse en los caracteres moleculares, pero como es este el contexto en que las matrices de costes son mayormente utilizadas, resulta más fácil que se de dicho problema. Existen tres situaciones principales que pueden conducir al incumplimiento de la desigualdad: si lo costes de las transversiones son superiores al doble del de las transiciones, si el coste de los *gaps* es muy bajo (p.ej. si se tratan como *missing data*) o si la matriz es asimétrica. Estas situaciones, si bien muchas veces evitables, reducen el número de valores posibles asignables y, por tanto, seleccionan entre aquellos esquemas que resultan filogenéticamente más razonables. Sin embargo, hay autores que han argumentado que esta inconsistencia puede tener, en algunos casos, una explicación biológica que la legitime (Fitch, 1993; Maddison y Maddison, 1992, pero ver Allard y Carpenter, 1996).

### 3.5.2. Peso de los caracteres

Paralelamente, y de forma similar a la asignación de costes diferenciales de las transformaciones de estados, los caracteres pueden ser ponderados. Como se ha comentado anteriormente para el caso de los estados, el peso relativo de un carácter frente a otro puede derivarse a partir de consideraciones externas o precedentes al análisis (*a priori*) o basarse en los resultados de un análisis preliminar (*a posteriori*). En el caso concreto de los datos morfológicos, los esquemas de ponderación *a priori* son en general de difícil justificación (Wheeler, 1986) pero ver (Maddison, 1993; Neff, 1986). Sin embargo, se han propuesto métodos de ponderación que utilizan el criterio de compatibilidad entre los caracteres como base para el cálculo de sus pesos (Sharkey, 1989). Entre aquellos métodos que derivan pesos a partir de árboles preliminares destaca la denominada ponderación sucesiva (*successive weighting*) (Carpenter, 1988; Carpenter, 1994; Farris, 1969). Esta técnica

se basa en el concepto de fiabilidad cladística (Farris, 1969), es decir, en utilizar el grado de ajuste de un carácter a una jerarquía como criterio para su valoración. El peso de un determinado carácter será función de su ajuste sobre el árbol o árboles resultantes de un análisis previo, generalmente utilizando ponderación uniforme. Los caracteres de la matriz se juzgan a sí mismos en base a su fiabilidad (Carpenter, 1994). Este método fue originalmente propuesto como un criterio adicional para la elección de un árbol en situaciones en donde el análisis producía diversos cladogramas igualmente parsimoniosos (Carpenter, 1988). Sin embargo, se ha constatado que la aplicación de este tipo de ponderación puede dar como resultado un árbol distinto tanto en longitud como en topología al original (Brothers y Carpenter, 1993; Platnick et al., 1991a). Muchos autores proponen que los árboles obtenidos mediante una ponderación "adecuada" de los caracteres deben ser considerados como la mejor hipótesis de la filogenia, independientemente de que éstos sean más o menos parsimoniosos bajo ponderación uniforme (Farris, 1969; Goloboff, 1993; Kluge y Farris, 1969; Platnick et al., 1991a) pero ver (Scharff y Coddington, 1997). La ponderación sucesiva es un método iterativo, de manera que a la obtención de un árbol, o árboles, mediante la aplicación de un determinado esquema de ponderación, le sigue una nueva asignación de pesos y un nuevo análisis bajo los mismos. El proceso se repite hasta que en dos análisis sucesivos se obtienen los mismos árboles, tanto en longitud como en valores de sus índices de homoplasia. A este criterio se le denomina autoconsistencia (Farris, 1969). Un árbol es autoconsistente si es el más corto bajo el esquema de ponderación de los caracteres que él mismo implica, y es el árbol que resuelve los conflictos entre los caracteres a favor de aquellos que son menos homoplásicos. La aplicación práctica de esta técnica plantea algunos problemas. La primera cuestión es la elección del índice a utilizar para medir el ajuste de los caracteres. El más utilizado actualmente es el *CR*, que pasa por ser el menos sensible a la existencia de diferentes tipos de caracteres (diferente número de estados, ordenados o no, etc.), a la vez que, a diferencia del *IC*, puede alcanzar valores de 0 y eliminar, por tanto, completamente un carácter del análisis. No obstante, algunos autores han propuesto utilizar el *IC*, ya que el *CR* (al igual que el *IR*) no solamente asigna un peso menor a aquellos caracteres más homoplásico, sino también a aquellos con una mayor variabilidad informativa ( $g - m$ ) (Goloboff, 1991, 1993). Un problema adicional se plantea cuando existen dos o más árboles igualmente parsimoniosos para los datos uniformemente ponderados. En estos casos puede escogerse entre el mayor valor posible del índice del carácter, la media de los valores del índice en cada árbol o el valor menor. Otro factor importante tiene que ver con la implementación computacional de la técnica y, en especial, con la escala de los valores de los pesos. En algunos programas, como por ejemplo Hennig86 (Farris, 1988), los pesos asignados solo pueden tomar valores entre 0 y 10, quedando los valores decimales truncados, aceptándose únicamente su parte entera, mientras que en otros, como PAUP/PAUP\* (Swofford, 1993), permiten escoger tanto los tipos de índice (*IC*, *IR*, *CR*), como cualquiera de los valores de los mismos dado un conjunto de árboles. A su vez, la escala utilizada permite asignar valores de 0 a 1000, y ofrece la posibilidad de redondear los valores en lugar de truncarlos. Las ventajas de una implementación respecto a las demás, así como el efecto que éstas pueden tener sobre la elección del árbol resultante, no han sido explícitamente investigadas. Si bien una escala de 0 a 1000 permite obviamente una mejor discriminación, no queda claro si la misma es realmente deseable debido a la naturaleza aproximativa de este tipo de asignación de pesos. Por otra parte, algunos autores han

alertado de la posibilidad que, en ponderación sucesiva, caracteres complejos perfectamente definibles y objetivos reciban pesos bajos, mientras que caracteres con interrogantes o más pobremente definidos se vean potenciados (Scharff y Coddington, 1997). Dichos autores abogan, en general, por evaluar críticamente las implicaciones para la evolución de los caracteres derivadas del árbol resultante de la ponderación diferencial sucesiva.

Goloboff (1993, 1995), a pesar de estar de acuerdo con la asignación de pesos utilizada en la ponderación sucesiva, argumenta en contra de su implementación iterativa. Así, considera que: (1) la solución final depende del conjunto inicial de pesos, (2) puede darse el caso que alguno de los árboles iniciales no sea autoconsistente y (3), dado que la evaluación de la autoconsistencia se realiza sobre una submuestra de los árboles posibles, este método no permitirá encontrar todos los árboles autoconsistentes. Para evitar éstas y otras limitaciones de la ponderación sucesiva, se propone una nueva aproximación basada en el cálculo del peso de los caracteres simultáneamente a la búsqueda de los árboles (Goloboff, 1993, 1995). Este objetivo se consigue sustituyendo el criterio de optimización de la parsimonia clásica, es decir, la minimización del número de pasos del cladograma, por uno nuevo: la maximización de una función conjunta, cóncava y decreciente de la homoplasia ( $F$ ):

$$F = \sum f_i \text{ donde } f_i = k / (k + es_i + es_0)$$

El ajuste (*fit*) de un carácter ( $f_i$ ) es función cóncava del número de pasos adicionales ( $es_i = s - m$ ) que muestra en un árbol determinado y del número de pasos adicionales ( $es_0$ ) debidos a polimorfismos en los taxones terminales. La constante de concavidad ( $k$ ) determina el grado en el que los caracteres homoplásicos son rebajados. Valores bajos de  $k$  provocan el aumento de la concavidad de la función, lo que se traduce en que los caracteres homoplásicos tengan poca influencia. Valores altos de  $k$  linealizan la función (la relación entre el peso y la homoplasia en la parsimonia clásica es lineal), aproximando la influencia de los caracteres homoplásicos a la del resto. No obstante, la selección del valor de concavidad idóneo para un determinado análisis no ha sido motivo de estudio y es, por el momento, totalmente subjetivo. La asunción principal de este método es que los árboles que maximizan esta función de la homoplasia son los que maximizan la fiabilidad de los caracteres. Esta aproximación ha sido implementada en el programa de ordenador PEE-WEE (*Parsimony and Implied Weights*) v 2.50 (Goloboff, 1996b). Además de la elección del valor de  $k$ , la utilización de este criterio de ponderación esta sujeto a otras limitaciones. Por ejemplo, y dado que dicho criterio se basa en una función continua, uno puede preguntarse sobre el sentido, desde un punto de vista filogenético, de una diferencia de ajuste de décimas entre dos árboles (Coddington, com. personal). En este contexto, conviene destacar que el propio autor ha dotado al programa de la capacidad de retener árboles subóptimos, aunque la cuestión es ciertamente más compleja, no reduciéndose tan solo a examinar la reducción en el valor de ajuste de los caracteres, sino también a las causas de dicha reducción, es decir, los tipos de caracteres implicados en la misma (Goloboff, 1995). Por otra parte, cuando se utilizan datos moleculares, la diferencia entre el número de pasos mínimo y máximo de un carácter puede ser muy grande, y provocar una reducción demasiado drástica del ajuste del mismo. Con el objeto de evitar este efecto, se ha propuesto corregir la función de ajuste dividiendo su denominador por el número de pasos mínimo del carácter ( $m$ ) (Gladstein y Wheeler, 1997).

Tal y como se ha comentado al hablar del coste de las transformaciones, las peculiaridades de los datos moleculares, como por ejemplo el número fijo de estados compartido por todos los caracteres, así como el conocimiento previo de un conjunto de procesos que afectan la evolución a nivel molecular, ofrecen criterios adicionales para la ponderación, en este caso, de los caracteres. La asunción más generalizada es que las posiciones con una tasa elevada de cambio han sufrido, muy probablemente, sustituciones múltiples (=homoplasia) y que, para evitar que tengan una influencia negativa sobre los análisis, su peso ha de ser rebajado (Hillis et al., 1993). En el caso de genes codificadores de proteínas y debido a la degeneración del código genético, la frecuencia de cambios en las terceras bases de los codones acostumbra a superar en mucho las de las primeras y segundas bases nucleotídicas, es pues una práctica muy generalizada la de dar pesos superiores a primeras y segundas posiciones. El valor concreto del peso se suele derivar del inverso de la frecuencia relativa de cambios, ya sea para cada posición o uniendo primeras y segundas posiciones en una misma clase. A su vez, la cantidad de cambios de cada clase de posiciones puede determinarse mediante la comparación de pares de taxones o, para evitar la dependencia de los caracteres, inferirla de la reconstrucción de los estados ancestrales a partir de un árbol preliminar (Maddison y Maddison, 1992; Williams y Fitch, 1989). Sin embargo, esta aproximación supone una excesiva simplificación, ya que en algunos casos un cambio en una tercera posición puede suponer también un cambio aminoacídico y, al contrario, hay situaciones en las que un cambio en una primera posición no conlleva una sustitución aminoacídica. Una asignación más "realista" de los pesos podría resultar excesivamente compleja (Simon et al., 1994). Cuando la divergencia entre los taxones es suficientemente grande, existe la posibilidad de traducir la secuencia nucleotídica a aminoácidos y realizar el análisis utilizando los mismos como caracteres. Recientemente, ha sido propuesta una opción intermedia, consistente en combinar los caracteres nucleotídicos y su traducción aminoacídica en una misma matriz, consiguiendo de esta manera la potenciación de los patrones derivados de las bases nucleotídicas que sean más congruentes con los cambios aminoacídicos, en principio más conservados (Agosti et al., 1996). La principal limitación de esta técnica es que los dos tipos de caracteres (nucleótidos y aminoácidos) no son evidentemente independientes, ya que son dos formas de codificación de una misma información subyacente, violando, por tanto, uno de las asunciones básicas de la parsimonia. A pesar de esto, los defensores de dicho método argumentan que, en la práctica, estos caracteres se comportan como si fuesen independientes y, por tanto, la aplicación del criterio de parsimonia queda garantizado.

A diferencia de los genes codificantes, los genes ribosomales son fundamentalmente estructurales, sufriendo el ARN transcrito un proceso de plegamiento que determinará su estructura funcional. Ésta será, básicamente, la consecuencia de la existencia de regiones de la secuencia que son complementarias de otras, a las que se unen. El resultado es una estructura espacial con regiones de cadena doble, los *stems*, alternadas con regiones de cadena simple, los *loops*. La distinción de las posiciones ribosomales en estas dos clases es la base de algunos esquemas de ponderación aplicados a estos genes. Por una parte, se ha sugerido que debido a su papel en el mantenimiento de la estructura secundaria, los nucleótidos que forman parte de *stems* son más conservativos que los de los *loops* y, por tanto, debería aumentarse el peso de los mismos en el análisis. Sin embargo, la frecuencia de cambio de diferentes regiones de los genes ribosomales no parece apoyar esta afirmación (Simon et al., 1994; Vawter y Brown, 1993). Contrariamente, se ha

propuesto que el peso de las posiciones en *stems* debería ser rebajado en el análisis ya que una mutación en una zona de *stem* requerirá una substitución compensatoria en la cadena complementaria para mantener la estructura secundaria (Wheeler y Honeycutt, 1988). Así pues, y debido a la no independencia de los cambios en estas posiciones, los *stems* deberían ser rebajados en un factor de 1/2 respecto a los *loops*. Algunos autores consideran que este valor es el resultado de una excesiva simplificación del proceso, ya que dado que existen emparejamientos no clásicos entre las bases (p. ej. G-T o T-C), las compensaciones pueden no ser perfectas (Simon et al., 1994). Los datos empíricos parecen sugerir que los *stems* no deberían ser rebajados más de un 80% respecto a los *loops* (Dixon y Hillis, 1993).

La gran mayoría de aproximaciones para la asignación de pesos o costes en datos moleculares, adolecen de la misma limitación: tratar los estados y las posiciones como clases (Carpenter, 1993). Esta uniformización de los estados y las posiciones conlleva el inconveniente de esconder las diferencias que puedan existir entre sus componentes. Así, por ejemplo, cuando se dividen las substituciones en transiciones y transversiones y se asigna un único coste a cada una de ellas, se está escondiendo el hecho conocido de que la proporción de transiciones respecto al de transversiones tiene un rango de variabilidad muy amplio, más o menos relacionado con el tiempo de divergencia.

Una desventaja adicional de la ponderación diferencial es que los estadísticos (=índices) de los árboles obtenidos bajo un determinado esquema de ponderación, ya sea en datos moleculares, morfológicos o su combinación, no pueden ser comparados entre esquemas diferentes (Simon et al., 1994).

### 3.6. Técnicas para la construcción del árbol y optimización de los caracteres

Una vez se ha construido la matriz de caracteres y las asunciones sobre el coste de las transformaciones y/o el peso relativo de los diferentes caracteres han sido explícitamente incorporados, el siguiente paso es encontrar el árbol o árboles más parsimoniosos. La parsimonia, al igual que otros métodos de reconstrucción filogenética como la máxima verosimilitud, incorpora un criterio de optimalidad, es decir, un criterio para, dado un conjunto de árboles decidir cual es el mejor. En el caso de la parsimonia el criterio es: "el mejor árbol es aquel con el menor número de pasos". El número de pasos de un carácter es la suma del número total de transformaciones entre sus estados, cada una de ellas multiplicada por el coste de dicha transformación. Cuando se aplica un método que incorpora un criterio de optimalidad, la problemática de hallar el mejor árbol puede dividirse en dos fases: (1) construir los árboles, y (2) evaluarlos (Swofford et al., 1996a). Evaluar un árbol consiste en determinar su longitud y compararla con las del resto de árboles. La longitud de un árbol es la suma del número de pasos de cada uno de sus caracteres multiplicado por su peso correspondiente. Así, dados dos caracteres, el primero con 2 pasos y un peso de 3, y el segundo con 4 pasos y un peso de 1, la longitud total será:  $(2 \times 3) + (4 \times 1) = 10$ . El cálculo del número de pasos de un carácter en un árbol concreto, es el resultado de optimizar el carácter en éste. Se denomina optimización de un carácter a la asignación de un estado concreto del carácter a cada uno de los nodos internos (antepasados hipotéticos) de un determinado árbol, de forma que el número de transformaciones necesarias para explicar la distribución de estados en el árbol sea el mínimo posible. En ciertos casos la optimización puede ser ambigua, es decir, la asignación de uno u otro estado a un determinado nodo sea

igualmente parsimoniosa. En estos casos, para escoger uno de los estados pueden utilizarse criterios adicionales. Así, por ejemplo, puede preferirse que la homoplasia del cladograma sea principalmente el resultado de reversiones, lo cual se consigue favoreciendo que los cambios de estado se produzcan hacia la raíz del árbol. A este tipo de optimización se la denomina ACCTTRAN (Swofford y Maddison, 1987; Swofford y Olsen, 1990). Al contrario, podría preferirse que las convergencias sean la causa fundamental de homoplasia, favoreciendo la optimización que concentra los cambios tan lejos de la raíz como sea posible. Esta optimización recibe el nombre de DELTRAN (Swofford y Maddison, 1987; Swofford y Olsen, 1990). A su vez, la optimización depende de las asunciones sobre las relaciones entre los estados que se hayan incorporado, la optimización de un carácter será distinta dependiendo de si dicho carácter es ordenado o desordenado. Como se ha comentado en el apartado de costes de las transformaciones, existen diferentes algoritmos computacionales que permiten calcular el número de pasos de un carácter considerando ciertas asunciones sobre las relaciones entre sus estados. Por otra parte, debe tenerse en cuenta que la longitud de un árbol es la suma ponderada de los pasos de cada carácter, dos árboles pueden tener la misma longitud y ser topológicamente diferentes. Esto se explica porque una topología puede ahorrar pasos en ciertos caracteres y aumentarlos en otros, de forma que la suma final sea la misma que la de la otra topología, la cual puede haber ahorrado o ganado pasos en caracteres distintos. La consecuencia de esta situación será la obtención de más de un árbol máximo parsimonioso.

#### 3.6.1. Búsquedas exhaustivas

A primera vista, la búsqueda del árbol más parsimonioso puede parecer trivial. Solamente se necesita implementar un algoritmo computacional que construya todos los árboles posibles para los taxones estudiados, calcular la longitud de cada uno de ellos y seleccionar los más cortos. A este tipo de aproximación se la denomina búsqueda exhaustiva. Sin embargo, el problema no es, ni mucho menos, tan simple. La principal limitación de las búsquedas exhaustivas es que el número de árboles posibles aumenta de forma explosiva con el número de nodos terminales (taxones). Dado un número  $n$  de taxones, el número de árboles sin raíz posibles viene dado por la fórmula:

$$\prod (2i - 1), \quad i = 1, 2, \dots, n-2$$

Así, para 7 taxones existen 945 árboles dicotómicos sin raíz posibles, pero si el número de taxones aumenta a 20, el número de árboles posibles será de  $2,2 \times 10^{20}$ . Una búsqueda exhaustiva será solo factible cuando la matriz contenga un máximo de 11 taxones (34.459.425 árboles) (Swofford et al., 1996a). Afortunadamente, si el valor óptimo de un método no disminuye nunca al añadir un nuevo elemento al cálculo, como ocurre en parsimonia, existen algoritmos que permiten encontrar el mejor árbol o árboles sin tener que evaluar todos los posibles. Estos algoritmos se basan en el método denominado *branch-and-bound* (Hendy y Penny, 1982). Este método permite búsquedas con un número mayor de taxones y en menor tiempo computacional. A pesar de todo, en situaciones donde sea factible, es preferible optar por utilizar una búsqueda exhaustiva, ya que de esta manera, se obtendrá información adicional, como por ejemplo la distribución de árboles para cada número extra de pasos o la evaluación de árboles subóptimos (Kitching, 1992). Desgraciadamente, la utilización del *branch and bound* está también limitada por el número de taxones, en este caso situándose entre 20 y 25, dependiendo del número y nivel de homoplasia de los caracteres.

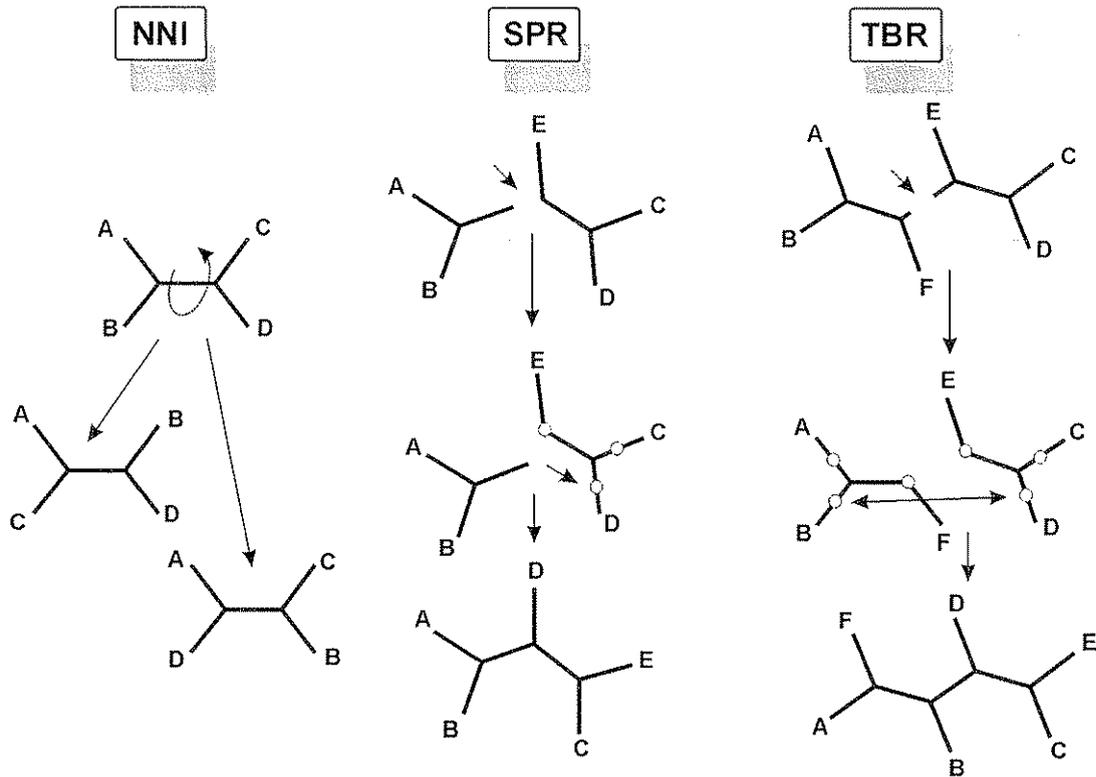


Fig. 12.- Representación esquemática de los diferentes métodos de reordenamiento de ramas.

### 3.6.2. Búsquedas heurísticas

Desde un punto de vista matemático, la búsqueda del árbol más corto dado un cierto número de taxones y caracteres es un problema NP-completo (NP= polinomial no determinístico). Esto quiere decir que no es probable que exista un algoritmo general que garantice solucionar todas las vertientes del problema, de forma que el tiempo computacional venga dado por una función polinómica del tamaño del problema (Penny et al., 1992). Por tanto, en ausencia de este tipo de algoritmos, conocidos como eficientes, el tiempo invertido por los algoritmos exactos para solucionar problemas NP-completos aumentan extremadamente rápido a medida que aumenta la complejidad de los datos. La única aproximación posible en estos casos es la utilización de métodos de “prueba y error”, los cuales consisten de dos pasos básicos: (1) construcción de un árbol inicial y cálculo de su número de pasos y (2) modificación de este árbol reordenando sus ramas, es decir, cambiando la posición de los taxones con el objetivo de encontrar topologías más cortas (Maddison, 1991). A este tipo de búsquedas se las denomina heurísticas. En general, la obtención del árbol inicial se realiza mediante un protocolo de incorporación secuencial de los taxones (*stepwise addition*), así, se construye un árbol preliminar seleccionando tres taxones y, posteriormente, se van añadiendo los taxones restantes uno a uno. El criterio utilizado para seleccionar los taxones iniciales y el orden de adición de los siguientes distingue diversas variantes del protocolo:

1. *As is*. Los tres taxones iniciales son los tres primeros de la matriz, y la incorporación de los taxones siguientes sigue estrictamente el orden de la misma.
2. *Random*. Los taxones iniciales y el orden de incorporación de los siguientes se establece al azar.

3. *Simple* (Farris, 1970). Se selecciona un taxón, generalmente el definido como grupo externo, y se calcula la distancia de *Manhattan* de todos los taxones restantes respecto al seleccionado. Los dos con la distancia más corta formarán con el primero el árbol preliminar. El orden de incorporación de los taxones restantes se realiza en función de su distancia al primero, de menor a mayor. En caso de que dos taxones tengan el mismo valor, se selecciona uno de ellos al azar.
4. *Closest*. Se examinan todos los árboles posibles de tres taxones y se selecciona el más corto. Seguidamente, se añade cada uno de los taxones restantes en todas las ramas posibles del árbol y se retiene el taxón y la disposición que hacen el árbol más corto. Se repite este procedimiento hasta que se unen todos los taxones. Es, evidentemente, el más costoso computacionalmente.

Una vez construido el árbol o árboles iniciales, el siguiente paso consiste en intercambiar el orden de sus ramas (*branch swapping*). Las reordenaciones más utilizadas por los programas de ordenador de inferencia filogenética son (Swofford et al., 1996a) (Fig. 12):

1. *Nearest neighbour interchange (NNI)*. Se selecciona una rama interna y se intercambian entre ellos los dos grupos que quedan en cada uno de sus extremos. Para cada rama existen dos cambios posibles. Esto se realiza para todas las ramas internas del árbol.
2. *Subtree pruning and regrafting (SPR)*. Se rompe una rama interna de forma que se obtengan dos subárboles, uno con una rama libre y el otro sin ella. El subárbol con la rama libre se une, a través de ésta, a las diferentes ramas del otro subárbol. Se evalúan todas las posibles combinaciones de romper ramas internas y reconectar los subárboles.

3. *Tree bisection and reconnection (TBR)*. Similar a la anterior pero, en este caso, ninguno de los dos subárboles conserva una rama libre. Tras separar los subárboles, cada uno de ellos se conecta por cada una de sus ramas internas a todas las ramas internas del subárbol complementario. Se evalúan todas las posibles combinaciones posibles de romper ramas internas y reconectar los subárboles.

La relación entre el conjunto de árboles que puede generar cada tipo de reordenación es  $NNI < SPR < TBR$  (Maddison, 1991). Se puede comprobar que las reordenaciones TBR son las más exhaustivas, a pesar de que son, también, las computacionalmente más costosas.

Debe tenerse en cuenta que las búsquedas heurísticas no garantizan obtener el árbol más corto, sino que dados unos árboles iniciales, pueden encontrar los árboles más cortos derivados de los mismos, lo que se conoce con el nombre de mínimo local. Si alguno de los mínimos locales hallados coincide con el mínimo global, entonces se habrá hallado al menos uno de los árboles más parsimoniosos, aunque nunca se podrá tener la certeza de que esto sea así. Un problema añadido es la posibilidad de que los árboles más parsimoniosos pertenezcan a islas diferentes. Una isla se definen como una colección de árboles parsimoniosos conectados y separados de otros grupos de árboles igualmente parsimoniosos por árboles intermedios de mayor longitud (Maddison, 1991). La conexión hace referencia al hecho de que estos árboles están separados los unos de los otros por una sola reordenación de ramas. Evidentemente, el número y la topología de los árboles de una isla estará determinado por el tipo de reordenación (NNI, SPR o TBR). Si la búsqueda utiliza solamente un árbol preliminar no será posible encontrar árboles pertenecientes a islas distintas, ya que estos estarán separados por árboles más largos y en la evaluación de los reordenamientos tan solo se retienen árboles cada vez más cortos. Se ha demostrado empíricamente que si los árboles obtenidos en un análisis exhiben un índice de consistencia inferior a 0,67 y el número de taxones es mayor que 20, muy probablemente existe más de una isla de árboles parsimoniosos. Una manera de encontrar diferentes islas de árboles es utilizando diversos árboles preliminares. Una de las búsquedas más eficientes consiste en generar muchos árboles iniciales utilizando el protocolo de incorporación de taxones *random* (p. ej. 100 árboles iniciales) y aplicar posteriormente a cada uno de ellos reordenaciones tipo TBR. Disponer de diversos puntos iniciales donde empezar a buscar, permitirá una mejor exploración del espacio de árboles. Dado que, en principio, cada uno de los árboles será distinto a los demás, la convergencia de los diferentes resultados en un conjunto de árboles comunes confiere una mayor confianza en el análisis. Otra opción, compatible con la anterior, consiste en utilizar reordenaciones de ramas más complejas, con el objetivo similar de aumentar el tamaño de la muestra de árboles investigados. Así el programa PEE-WEE y afines (Goloboff, 1996a, 1996b, 1996c, 1996d), incorporan algoritmos de reordenación entre ramas múltiples. Sin embargo, dicha aproximación todavía se halla en periodo experimental y no existe ninguna referencia sobre su eficacia.

### 3.7. Fiabilidad de los árboles y clados obtenidos

Dado un cierto método de inferencia filogenética con un criterio explícito de optimización, la hipótesis escogida de relaciones filogenéticas entre los taxones estudiados será aquella que maximiza el criterio escogido. Sin embargo, existen situaciones que pueden hacer poner en duda que el árbol escogido, ya sea en su conjunto o en alguno de sus clados, refleje fielmente las relaciones filogenéticas entre los organismos o, al menos, que lo

haga sensiblemente mejor que otro árbol desechado. En este contexto, surge el interrogante sobre cual es el grado de confianza o fiabilidad que merece la interpretación del árbol resultante como hipótesis filogenética. En realidad, este grado de confianza o fiabilidad puede separarse en dos aspectos distintos: (1) tiene el árbol en su conjunto una señal o estructura filogenética significativas y (2) cuál es el grado de apoyo o la estabilidad de los diferentes clados contenidos en el mismo. Antes de continuar, conviene recordar que la aplicación de ciertos conceptos estadísticos, como intervalos de confianza o significación de un determinado test, se basan en la existencia de una distribución de probabilidad conocida, o al menos asumida, de los datos. En inferencia filogenética, dicha distribución de probabilidad sólo puede ser conocida en el caso que se acepte un determinado modelo evolutivo de generación de los caracteres. La parsimonia se caracteriza frente a otros métodos de reconstrucción filogenética por la no incorporación en el análisis de un modelo evolutivo. No obstante, existen una serie de métodos matemáticos que permiten el cálculo de ciertos estadísticos aún en ausencia, o desconocimiento, de la distribución de probabilidad. Estas metodologías se conocen con el nombre de técnicas de remuestreo, ya que mediante el muestreo iterativo de los datos originales se obtiene un conjunto de réplicas, denominadas pseudoréplicas, a partir de las cuales se elabora la distribución de interés. Dichas pseudoréplicas se generan mediante la aleatorización de los datos originales. En el caso particular de las matrices de caracteres, este remuestreo puede realizarse tanto a nivel de columnas (caracteres) como de filas (taxones). Las técnicas de aleatorización computacional intensiva pueden dividirse en tres tipos básicos: *jackknife*, *bootstrap* (un subtipo de los denominados métodos Monte Carlo) y permutaciones (Siddall, 1995), cada una de ellas caracterizada por la forma en que los datos originales son remuestreados (Fig. 13). Así, en *jackknife* las pseudoréplicas se obtienen por eliminación de uno (primer orden) o de un número  $n$  (orden  $n$ ) de los datos iniciales. En *bootstrap* las pseudoréplicas se obtienen seleccionando al azar y con reemplazamiento los datos iniciales, hasta obtener el mismo número de observaciones que en la muestra original. En la pseudoréplica algunos de los datos originales estarán repetidos  $n$  veces, mientras que otros estarán ausentes. En las permutaciones, los datos son simplemente reordenados, rompiendo de esta forma cualquier tipo de covariación que pudiese existir entre ellos.

#### 3.7.1. Señal, estructura o información filogenética de un árbol

Dado que todos los métodos de reconstrucción filogenética asumen implícitamente que existe una relación entre los taxones susceptible de ser representada en forma de árbol, incluso en los casos en que los estados de los caracteres hubiesen sido asignados utilizando un dado, se generaría uno o más árboles. La parsimonia optimiza la congruencia entre los caracteres, pero no discrimina entre las posibles fuentes de congruencia. Así en ausencia de un proceso unificador (=evolución), esta congruencia podría darse por simple azar. Por tanto, la cuestión es si la matriz muestra una congruencia entre los caracteres mayor a la esperada por azar. Existen dos técnicas principales para abordar esta cuestión:

1. *Skewness*. La distribución del número de árboles posibles para cada longitud para una determinada matriz está fuertemente sesgada hacia los árboles con más pasos. El estadístico  $g$ , (Sokal y Rohlf, 1981) mide el grado de sesgo (*skewness*) de una distribución, siendo negativo cuando la distribución está sesgada hacia la izquierda (=menos árboles cortos que largos). Cuanto mayor sea la congruencia entre los caracte-

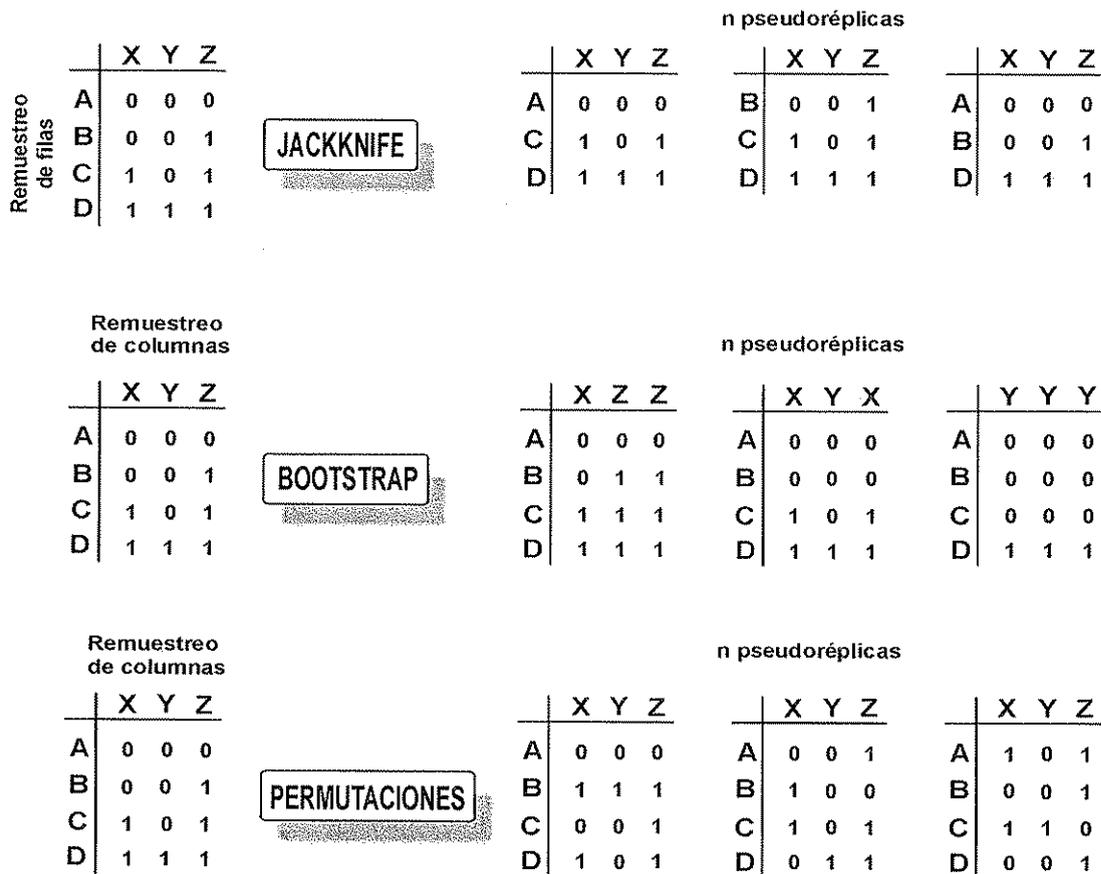


Fig. 13.- Tipos de aleatorizaciones aplicables a matrices de caracteres.

res, cualquiera que sea la razón, más sesgada será la distribución, ya que solamente un pequeño grupo de árboles será mucho más corto que el resto. Se ha propuesto la utilización del índice  $g$ , como medida del grado de estructura filogenética de una matriz, cuanto más negativo sea este valor mayor será la señal filogenética del mismo (Hillis, 1991; Huelsenbeck, 1991). Además, a partir de los valores de  $g$ , obtenidos en matrices generadas al azar, se puede construir un test de significación que permita ver si el valor de la matriz original se separa significativamente del de matrices sin señal filogenética, aquellas en las que la congruencia entre los caracteres se debe solamente al azar (Hillis y Huelsenbeck, 1992). Sin embargo, se ha demostrado que el índice  $g$ , puede estar más influenciado por la frecuencia de los estados dentro de los caracteres que por la congruencia entre éstos, además de ser sensible al número de caracteres (Kallersjö et al., 1992). Otras críticas son que no existe un conocimiento suficiente de los valores de  $g$ , como para determinar exactamente los umbrales de significación del estadístico  $y$ , si realmente una diferencia significativa respecto a un valor generado al azar implica necesariamente la existencia de una estructura jerárquica significativa (Trueman, 1993).

2. *Test PTP (permutation tail probability)*. Siguiendo un razonamiento similar al anterior, la congruencia entre los caracteres en una matriz con señal filogenética deberá ser mayor que la simplemente producida por el azar. El denominado PTP (Archie, 1989; Faith y Cranston, 1991) es un test de significación basado en la permutación de los estados de cada uno de los caracteres de la matriz original, generando matrices con las mismas características (igual número de taxones, de caracteres, de estados y de *missing data* por

carácter), pero con congruencia entre los caracteres debida solamente al azar. Dado que el grado de congruencia entre los caracteres es el responsable último de que unos árboles sean más cortos que otros, se toma la longitud del árbol más parsimonioso como medida de referencia de la congruencia. La distribución de probabilidades, o mejor frecuencias, se establece a partir del cálculo del árbol más parsimonioso para un número  $N$  de matrices permutadas (=pseudoréplicas) (Fig. 13). La matriz original será significativamente diferente de las permutadas, es decir, tendrá señal filogenética si la probabilidad de que la longitud del árbol más parsimonioso para los datos originales se sitúa dentro de la cola inferior de la distribución de dichas probabilidades. Formalmente:

$$\alpha' = 1 - E / (N + 1), \text{ donde}$$

$$\alpha' < \text{nivel de significación escogido (p. ej. 0,05)}$$

Donde,  $\alpha'$  es la tasa de error y  $E$  es el número de casos (observados y permutados) donde las longitudes son más cortas que la observada. En términos de intensidad computacional, se ha comprobado que pueden alcanzarse niveles de significación del 0,05, en tan solo 20 permutaciones (Trueman, 1993). Es importante recordar que este test sólo permite responder a la pregunta si la congruencia en los caracteres de la matriz es mayor que la que se esperaría al simple azar, pero en ningún caso si esta congruencia es debida a la evolución o a algún otro proceso subyacente, como pudiera ser la adaptación ecológica. Si bien este test ofrece ciertas ventajas sobre el *skewness*, como la no sensibilidad a las frecuencias de cada estado o al número de caracteres, se ha observado que, en ciertos casos, puede

obtenerse una alta significación en matrices con una estructura filogenética ambigua (Källersjö et al., 1992) o, incluso, ser significativo cuando ésta es simplemente inexistente (Alroy, 1994). Con el objetivo de evitar este efecto del todo indeseable, se ha propuesto una modificación de este test, consistente en utilizar como estadístico el denominado apoyo total (TS), el cual resulta de sumar los valores de apoyo de Bremer (ver siguiente apartado) de todos los cladogramas del árbol (Källersjö et al., 1992). Otra forma de mejorar el test original es mediante la no inclusión en dicho test de los grupos externos, ya que de esta manera se evita un sesgo hacia el rechazo de la hipótesis nula (no hay diferencias significativas) la cual puede aparecer en caso contrario (Trueman, 1996).

### 3.7.2. Grado de apoyo de los cladogramas de los árboles obtenidos

No todos los cladogramas que aparecen en un clado reciben el mismo apoyo por parte de los datos. Dicho apoyo puede interpretarse como una medida de la estabilidad de un determinado clado frente a perturbaciones en los datos, tales como revisión y recodificación de los caracteres, o la inclusión de nuevos datos, ya sea en forma de nuevos taxones o nuevos caracteres. Una forma simple de medir este apoyo es mediante el número de sinapomorfias que definen un determinado clado. Un criterio adicional es evaluar el tipo de caracteres implicados en dicho apoyo, p.ej. caracteres morfológicos complejos frente a caracteres simples, o transiciones frente a transversiones. Un refinamiento de la simple longitud (número de cambios) de las ramas, es considerar tan solo aquellas sinapomorfias únicas, es decir, que no se dan en ningún otro clado del árbol (Kluge, 1989). No obstante, se ha sugerido que este tipo de medidas puede conducir a error, ya que la homoplasia no tiene porque estar uniformemente distribuida a lo largo del árbol, concentrándose en ciertas partes del mismo (Sanderson, 1995). En la actualidad los índices más utilizados para medir la estabilidad o grado de apoyo de los cladogramas son:

1. *Bootstrap*. Este test estadístico ha sido propuesto como una metodología para estimar los límites de confianza de las ramas internas de un árbol filogenético (Felsenstein, 1985). Desde un punto de vista puramente estadístico, el *bootstrap*, como técnica de remuestreo, es un método para estimar varianzas, intervalos de confianza y otras características de ciertos estadísticos cuando no se conoce, o es difícil de establecer, la distribución de las muestras (Efron, 1979, 1982, 1987). Con este fin, se asume que la muestra inicial representa fielmente las características del universo al que representa, por lo que, en lugar de obtener más muestras de dicho universo, pueden generarse nuevas muestras a partir de la original por remuestreo de la misma. En el caso del *bootstrap*, cada pseudoréplica se obtiene mediante remuestreo con reemplazamiento de los datos originales, poseyendo éstas el mismo tamaño que la original. El nivel de confianza de una hipótesis puede estimarse como el porcentaje de veces que la hipótesis es apoyada por las pseudoréplicas construidas mediante *bootstrap*. En inferencia filogenética, la matriz de caracteres representa la muestra inicial. Las pseudoréplicas son matrices con el mismo número de taxones y caracteres pero en las cuales los caracteres han sido remuestreados con reemplazamiento, de manera que algunos estarán repetidos una o más veces y otros estarán ausentes (Fig. 13). A continuación, se obtiene el árbol más parsimonioso para cada una de las pseudoréplicas. El apoyo de una rama

determinada será el porcentaje de veces que el clado que define aparece en el conjunto de árboles de las pseudoréplicas, lo que se conoce como proporción de *bootstraps* (BP). El primer problema que plantea la utilización del *bootstrap* como medida de apoyo, es saber que es exactamente lo que se está midiendo. Según su formulación original (Felsenstein, 1985) el *bootstrap* sería la probabilidad de que una rama interna fuese recuperada en un análisis posterior, el cual incluyera nuevos datos independientes. Posteriormente, se ha sugerido que la proporción de *bootstraps* se interpretaría mejor como la probabilidad de que una determinada rama esté representada en la verdadera filogenia del grupo (Felsenstein y Kishino, 1993). Estas dos interpretaciones se refieren a propiedades radicalmente distintas, la primera es la repetibilidad y la segunda la exactitud. Estas dos interpretaciones de los valores de BP han sido testadas mediante simulaciones por ordenador y filogenias de laboratorio (Hillis y Bull, 1993), y las conclusiones a las que se han llegado sugieren que las BP son: (1) una medida muy poco fiable del grado de repetitividad de la rama o clado considerado y (2) una medida muy sesgada de la exactitud de la rama o clado, ya que, generalmente, subestiman los valores altos y sobrestiman los bajos. Además este sesgo es dependiente de variables tales como: el número de taxones, el de caracteres o la propia topología del árbol; y, por tanto, los resultados de estudios distintos no pueden ser directamente comparados. Por ejemplo, los valores de *bootstrap* son sensibles a la presencia de autapomorfias, de forma que cuando éstas están presentes las BP tienden a ser más bajas (Carpenter, 1996). A pesar de todo, estos autores consideran el *bootstrap* como una medida útil de la confianza relativa en los diferentes cladogramas de un árbol. Por otra parte, se han propuesto diversas correcciones del cálculo del *bootstrap* para corregir dicho sesgo (Bull et al., 1993; Rodrigo, 1993; Zharkikh y Li, 1995), aunque su implementación computacional está muy limitada o es muy costosa. Un tercer concepto que puede prestarse a confusión es el de la precisión, el cual simplemente es simplemente el grado con que un número limitado de pseudoréplicas se aproximan al valor de BP obtenido mediante un número infinito de las mismas. La precisión es una función simple del número de pseudoréplicas, las cuales según datos empíricos deben ser de entre 400 y 2000 para obtener unos valores de BP similares a los reales (Hedges, 1992). Existe un largo listado de críticas a la utilización de los *bootstraps* en reconstrucción filogenética, la mayoría de ellas basadas en el incumplimiento de muchas de sus asunciones. Quizás la más fundamental sea que, para que el *bootstrap* sea válido, los caracteres sujetos a remuestreo deben ser independientes los unos de los otros y, además, han de estar idénticamente distribuidos. En el contexto de caracteres biológicos, eso es tanto como suponer que la evolución es estocástica, cosa que contradice totalmente el concepto de historicidad y, por tanto, ciertamente falsa (Kluge y Wolf, 1993). Por otra parte, el *bootstrap* asume que el estadístico objeto de análisis es una función "suave" de la distribución de la variable remuestreada, lo cual es cierto para estadísticos que tomen valores reales. Sin embargo, la presencia o ausencia de un determinado clado es un valor binario, y la teoría del *bootstrap* todavía no ha sido desarrollada para estos supuestos (Brown, 1994). Otra asunción del *bootstrap* es que el muestreo, elección y codificación de los caracteres se realiza al azar, es decir, sin sesgo por parte del investigador. No obstante, los sistemáticos tienden a recopilar los caracteres evolutiva-

mente más conservativos y más indicados para el nivel jerárquico de interés (Jones, Kluge y Wolf, 1993). Para conseguir que la muestra original de caracteres recupere la mayor parte de características del universo representado, su tamaño debe ser muy grande. Desgraciadamente, la mayoría de investigadores se ven obligados a trabajar con un número de datos muy limitado, si se tiene en cuenta la enorme cantidad de datos potencialmente disponibles en biología comparativa (Jones et al., 1993; Kluge y Wolf, 1993). Otro problema es que, al calcular las proporciones de *bootstrap*, se asume que se han hallado todos los árboles más parsimoniosos para cada réplica, es decir, la mejor estima de la filogenia que éstas apoyan. Sin embargo, cuando el número de datos es grande, o lo es el de las pseudoréplicas que se pretenden analizar, se suelen realizar búsquedas heurísticas que, como se ha comentado anteriormente, no garantizan que se encuentre el árbol más parsimonioso o todos los árboles más parsimoniosos (Sanderson, 1995). Por otra parte, dado que en cada pseudoréplica se eliminan unos caracteres y se multiplican otros, el *bootstrap* debería mejor ser interpretado como una forma de ponderación al azar de los caracteres, de manera que las zonas del árbol menos sensibles al peso asignado aparecerán más frecuentemente en las pseudoréplicas (Trueman, 1993). Cabe pues preguntarse si esto es en realidad de algún interés (Bremer, 1994; Trueman, 1993). Tal y como fue originalmente propuesto (Felsenstein, 1985), y como todavía puede verse en muchos estudios, los resultados finales del *bootstrap* suelen presentarse en forma de árboles de consenso (ver apartado 3.6), donde aparecen los clados con valores de BP superiores a un cierto mínimo preestablecido. En algunos casos, dicho consenso puede presentar clados que no aparecen en el árbol más parsimonioso o, al contrario, clados que estaban presentes en dicho árbol desaparezan. Esto supone que la hipótesis filogenética escogida resulte no ser la más parsimoniosa, lo cual es epistemológicamente injustificable (Miyamoto, 1985). A pesar de este largo listado de críticas y que en palabras de uno de los propios proponentes de la técnica "...la utilización del *bootstrap* en inferencia filogenética está desprovista de cualquier tipo de justificación rigurosa (Brown, 1994)", este índice sigue siendo el más intensamente utilizado en los estudios filogenéticos, principalmente en aquellos que utilizan datos moleculares.

2. *Apoyo de Bremer (=índice de deterioro, índice de apoyo, apoyo de las ramas)*. Una aproximación radicalmente diferente al problema de la evaluación de la estabilidad de los clados del árbol más parsimonioso, es considerar el grado de apoyo como una función del número de pasos extras necesarios para colapsar un clado, es decir, el apoyo de un clado es la diferencia entre el número de pasos del árbol más parsimonioso encontrado, y el árbol más parsimonioso bajo la restricción de que dicho clado no aparezca (Bremer, 1988, 1994). En principio, este valor es comparable entre matrices que contengan caracteres del mismo tipo, ya sean morfológicos o moleculares (Bremer, 1994). En el caso que los caracteres hayan sido anteriormente ponderados, es necesario corregir el valor de apoyo respecto la escala de ponderación aplicada. El valor de apoyo corregido se obtienen de dividir el apoyo ponderado por el cociente entre el número de pasos del cladograma ponderado y el número de pasos del mismo cladograma con ponderación uniforme e igual a 1 (Gustafsson y Bremer, 1995). Quizás una de las ventajas más importantes

de este método sea la de establecer el apoyo de los distintos clados basándose en los datos reales, no en perturbaciones de los mismos (Bremer, 1994). Como se ha visto anteriormente (apartado 3.7.1), a partir del apoyo de Bremer se ha derivado el Apoyo Total (TS), resultado de sumar los apoyos de cada uno de los clados del árbol, como medida de la estructura filogenética de la matriz (Källersjö et al., 1992). Dicha medida no debe confundirse con el denominado Índice de Apoyo Total (*t*) (Bremer, 1994), el cual mide la estabilidad global del árbol. Éste se define como el cociente entre el apoyo total (TS) y el número de pasos del cladograma más parsimonioso bajo ponderación uniforme y con peso igual a 1 (*p*). Dado que *p* es el valor máximo que puede tomar el TS, el valor de *t* puede variar entre 0, caso que no exista ningún apoyo para ningún clado, es decir, una politomía total, y 1, cuando *t* y *p* sean iguales, situación que solo se dará en ausencia total de homoplasia. La principal limitación del apoyo de Bremer es que se trata de un índice cualitativo, y por tanto carente de valores umbral a partir de los cuales tomar decisiones sobre la credibilidad del clado en cuestión. Se ha desarrollado un método de permutación que permite la evaluación estadística del grado de apoyo de la monofilia de un determinado grupo. El denominado *T-PTP* (*topology-dependent permutation tail probability*) (Faith, 1991) es un test de significación diseñado para comprobar si un determinado clado está significativamente apoyado por los datos. Sin embargo, simulaciones por ordenador de su comportamiento han demostrado que dicho test no es en realidad una medida de monofilia (Huelsenbeck, Hillis, y Jones, 1995; Swofford, Thorne, Felsenstein y Wiegmann, 1996b).

3. *Índice de monofilia Jackknife (JMI)*. Otra vertiente de la estabilidad de los clados tiene que ver con su sensibilidad a la inclusión de nuevos taxones en el análisis. Con el objetivo de medir dicho efecto, se construyen pseudoréplicas de la matriz original utilizando *jackknife* de primer orden (Fig. 13). A diferencia de los otros métodos de aleatorización comentados, en este caso se remuestran los taxones, de forma que cada nueva pseudoréplica es el resultado de la eliminación de uno de los taxones originales. Se busca el árbol más parsimonioso para cada pseudoréplica y se calcula la frecuencia con que cada uno de los clados del árbol original aparece en los árboles del remuestreo (Siddall, 1995). Un clado se considera que está presente en una pseudoréplica si aparece exactamente igual que en el árbol original o tan solo le falta el taxón eliminado en la réplica considerada. Se recomienda que el grupo externo sea mantenido al margen de las perturbaciones, ya que la existencia de una raíz es indispensable para la definición de monofilia. A diferencia del *bootstrap*, este índice no parece ser sensible al número de sinapomorfias en que se apoyan los grupos ni a la presencia de caracteres no informativos. Se ha aconsejado explícitamente no utilizar los valores del índice como una medida estadística, sino simplemente una aproximación cualitativa a la estabilidad de los clados. Se ha observado que, la eliminación de ciertos taxones comporta un aumento destacado del número de árboles más parsimoniosos encontrados y que, por tanto, su presencia en el análisis contribuye a estabilizar los resultados. Estos taxones se denominan críticos. Al contrario, la eliminación de otros taxones tienen un efecto opuesto, reduciendo el número inicial de árboles hallados, dichos taxones se conocen como problemáticos.

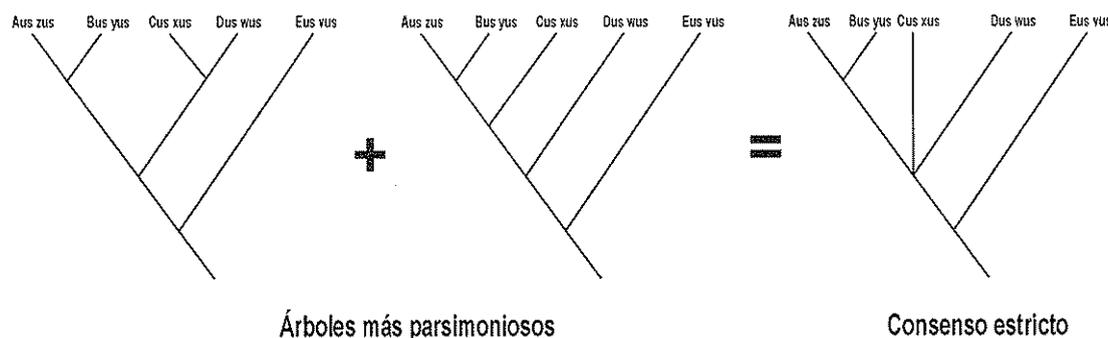


Fig. 14.- Ejemplo de construcción de un árbol de consenso estricto.

### 3.8. Árboles de consenso

Los árboles resultantes del análisis filogenético se denominan árboles fundamentales, ya que son una representación de la información contenida en los datos. Existe otro tipo de árboles donde lo que se representa es la información contenida en otros árboles, los árboles de consenso. El objetivo principal de los métodos de consenso es representar gráficamente los puntos de acuerdo entre los diferentes árboles fundamentales. Existen diferentes protocolos para derivar consensos (para descripción y resultados de su aplicación ver (Forey et al., 1992; Quicke, 1993; Wilkinson, 1994). Sin embargo, la gran mayoría de estos métodos producen árboles con clados que solo se encuentran en algunos de los árboles fundamentales o, en el peor de los casos, clados que no se encuentran en ninguno de ellos. Por tanto, su uso como sumario del conjunto de árboles más parsimoniosos provenientes de una matriz es, como mínimo, poco recomendable (Nixon y Carpenter, 1996). Aún así, su aplicación en el contexto de sintetizar la información de árboles provenientes de distintas matrices de datos, p. ej. biogeográficas, puede ser justificable. El método de consenso más ampliamente utilizado es el denominado consenso estricto, en el cual se produce un árbol donde únicamente se representan los clados que están presentes en todos los árboles fundamentales (Fig. 14). Un aspecto negativo del consenso estricto es que debido a su naturaleza tan conservativa, en muchos casos el árbol resultante muestra un elevado número de politomías, las cuales comprometen seriamente su nivel de resolución. No obstante, esta pérdida de resolución queda justificada por la inclusión en el consenso de aquellos componentes que no son ambiguos, es decir, para los cuales los datos son absolutamente claros. Desgraciadamente, en ciertas ocasiones esto puede no ser absolutamente cierto, debido a la existencia en los árboles fundamentales de ramas de longitud cero y/o de politomías. Es bien conocido el hecho de que en ciertas situaciones algunos programas de ordenador de análisis filogenético pueden generar cladogramas cuyos nodos no estén todos simultáneamente apoyados por cambios de carácter. Este fenómeno es el resultado de la existencia de ambigüedad en la optimización de algunos caracteres. Así, un determinado clado puede estar apoyado por un cambio bajo una cierta optimización, mientras otro clado puede estar apoyado por una optimización distinta del mismo carácter. Estos dos clados son mutuamente excluyentes, ya que dependen de optimizaciones alternativas. A pesar de ello, algunos programas ofrecen el árbol que presenta los dos clados al mismo tiempo entre sus soluciones (Coddington y Scharff, 1994). Por otra parte, no existe acuerdo sobre cual es la mejor manera de tratar los árboles con ramas de longitud cero. Algunos autores proponen colapsar todos aquellos clados que

no estén apoyados en cada una de las posibles optimizaciones de los caracteres, es decir, que tengan una longitud mínima de cero (Goloboff, 1996b; Nixon y Carpenter, 1996). Esta aproximación, conocida como apoyo estricto, comporta la eliminación de los árboles que difieren de los restantes por la existencia de un clado apoyado por una optimización particular de un carácter. Según dichos autores, estos árboles no representan hipótesis filogenéticas dignas de ser consideradas. Al contrario, otros autores consideran dichos árboles como legítimos, y defienden su inclusión dentro de las soluciones halladas (Coddington y Scharff, 1994). Esta aproximación se denomina apoyo semiestricto y consiste en el mantenimiento de aquellos árboles en los que sus clados estén apoyados al menos bajo una optimización. En ciertas ocasiones algunos de los árboles más parsimoniosos pueden mostrar una o varias politomías debido, en este caso, a la ausencia de cambios de estado que apoyen alguna de las resoluciones posibles dentro de dicha politomía. A su vez, estos árboles pueden ser compatibles o incompatibles con el resto de árboles más parsimoniosos. A pesar que, en general, estas politomías fundamentales son interpretadas como "suaves", los programas informáticos los interpretan como "duras". Se ha argumentado que bajo la asunción de que las politomías son "suaves", cualquier árbol dentro del conjunto de los árboles más parsimoniosos que presente alguna politomía, compatible o no con el resto, debe ser eliminada de cualquier análisis posterior (Coddington y Scharff, 1996). La justificación de dicha acción es que los árboles más resolutivos maximizan el número de clados, el contenido de información y el grado de falsibilidad de las hipótesis filogenéticas. Una aproximación diferente a esta situación es considerar los árboles de consenso estricto como una fuente de información sobre el grado de conflicto entre los caracteres de los árboles fundamentales (Nixon y Carpenter, 1996). Este razonamiento se basa en la constatación de que el consenso representa el límite superior del número de pasos de los caracteres en los árboles fundamentales. A partir de esta observación, se sugiere la utilización de un nuevo índice, la concordancia entre clados ( $CC$ ), definido como:

$$CC = 1 - [(\sum g_n - S_p) / (S_c - S_p)]$$

donde,  $g_n$  es el número máximo de pasos de un carácter en los árboles fundamentales,  $S_c$  es la longitud del árbol de consenso y  $S_p$  es la longitud de los árboles fundamentales más parsimoniosos. El numerador  $(\sum g_n - S_p)$  representa el grado de conflicto de los caracteres entre los cladogramas o, también, la homoplasia entre cladogramas. Este índice puede variar entre 0 y 1. El índice es 0 cuando el número de pasos de todos los

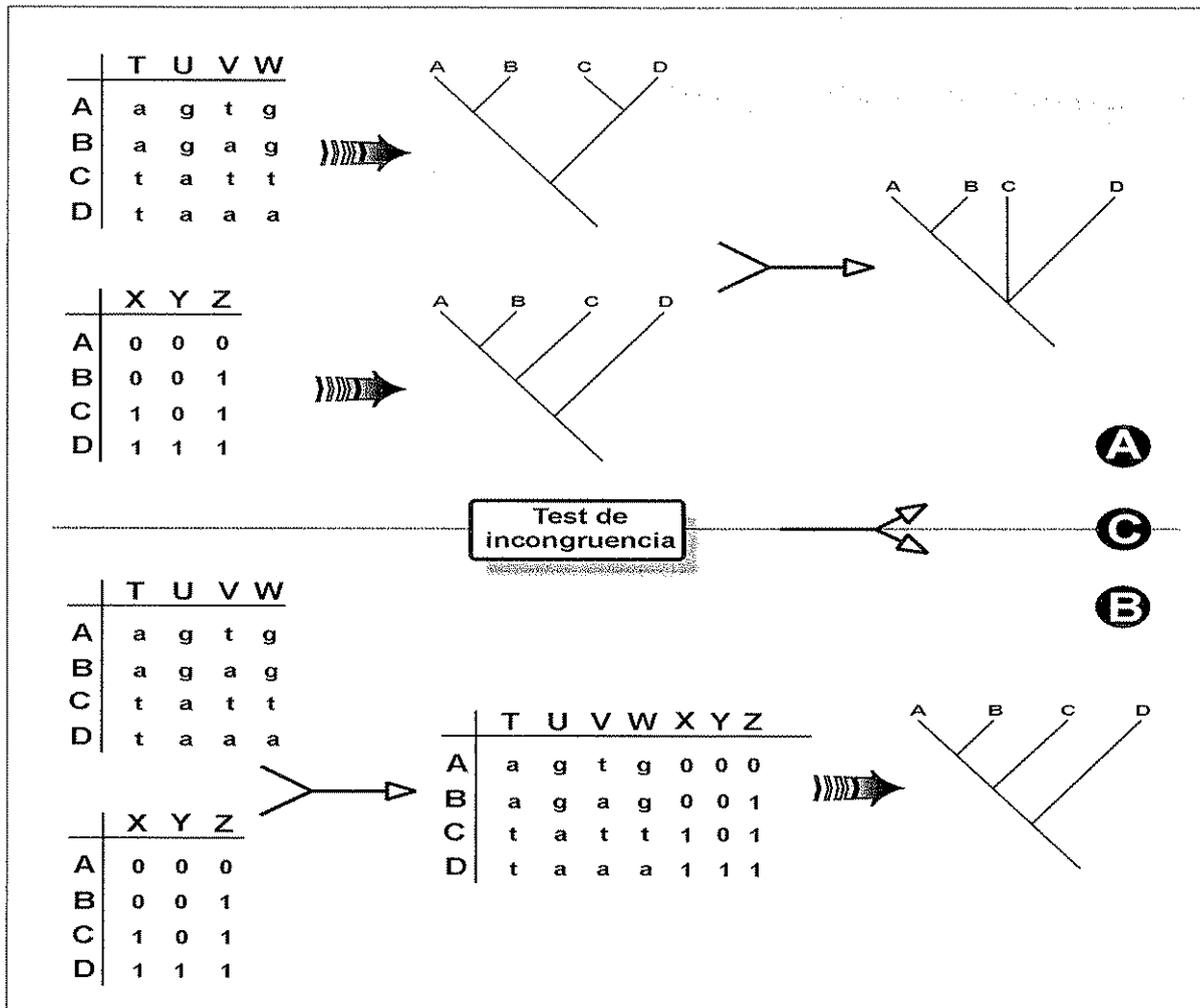


Fig. 15.- Representación esquemática de los diferentes tipos de análisis de caracteres de diferente naturaleza: A. Análisis separado, B. Análisis conjunto, C. Análisis conjunto condicionado.

caracteres alcanza su valor máximo al menos en uno de los árboles fundamentales que es, a su vez, el número de pasos del carácter en el consenso. El índice tiende a 1 a medida que la diferencia entre la longitud del consenso y la longitud máxima de los caracteres entre los árboles fundamentales aumenta. En la práctica, valores bajos del índice indican que las incongruencias (=politomías) del consenso son principalmente debidas a conflictos entre los caracteres de los árboles fundamentales. Al contrario, valores altos sugieren que no hay conflicto entre los caracteres, lo que implica que todavía hay información para agrupar algunos de los taxones incluidos en las politomías. El índice será 1 cuando, o bien hay un único cladograma máximo parsimonioso, o bien los grupos conflictivos solamente están apoyados por optimizaciones ambiguas.

### 3.9. Combinación de datos de diferente naturaleza: ¿análisis sobre particiones o análisis simultáneo?

El gran desarrollo y la accesibilidad de muchas técnicas de biología molecular, han permitido que en la actualidad resulte relativamente sencillo para los sistemáticos obtener diferentes tipos de datos moleculares de los taxones objeto de estudio. Paralelamente, en los últimos tiempos se ha demostrado la valía como indicadores de relaciones filogenéticas de caracteres antaño considerados como extremadamente variables desde un punto de vista histórico, como puedan ser los caracteres comportamentales o ecológicos. La posibilidad de incorporar conjuntos

de datos tan heterogéneos en el análisis filogenético, plantea el dilema de decidir cual es la mejor manera de analizarlos. Se han propuesto tres aproximaciones distintas.

#### 3.9.1. Análisis separado de las diferentes particiones de datos (= congruencia taxonómica o análisis partido)

Esta aproximación considera que la mejor manera de tratar las diferentes clases de caracteres (=particiones) es independientemente, es decir, obtener árboles para cada una de las matrices y utilizar un método de consenso para representar gráficamente la información común (Fig. 15). Las supuestas ventajas de esta opción son (Miyamoto y Fitch, 1995): (1) que se trata de una estima conservativa de la filogenia, (2) que los caracteres provienen de fuentes independientes y, por tanto, es menos probable que apoyen la misma filogenia si esta es errónea y (3) hay situaciones en que, por la propia naturaleza de los datos, es imposible aplicar el mismo método de reconstrucción a todas las particiones simultáneamente (p. ej. caracteres morfológicos con datos de hibridación ADN-ADN). No obstante, la mayoría de estos puntos han sido rebatidos. Así, la utilización del consenso no permite evaluar el apoyo relativo de los grupos conflictivos en los árboles originales (Nixon, 1996). Por otra parte, el análisis separado seguido del consenso comporta la ponderación diferencial de los caracteres de las distintas particiones (Kluge y Wolf, 1993), ya que si, como se ha comentado, se asume que hay más probabilidad de que los caracteres sean independientes entre clases distintas de datos, al analizar cada clase por separa-

do la dependencia de los caracteres será mayor. La presencia de caracteres dependientes tiene el mismo efecto que el aumento de peso de una carácter que apoye las mismas agrupaciones que éstos. La única manera de evitar dicho efecto es combinar las matrices para aumentar el número de caracteres independientes (Nixon, 1996). Finalmente, todos los árboles a partir de los cuales se deriva el consenso contribuyen de la misma forma, a pesar de que algunos de ellos se hayan obtenido en particiones con una estructura filogenética mayor que otras.

### 3.9.2. Análisis conjunto de los datos (= congruencia de caracteres, evidencia total o análisis simultáneo)

Esta aproximación propone combinar los datos en una sola matriz y analizarlos conjuntamente (fig. 15). Las ventajas de este tratamiento son (Eernisse y Kluge, 1993; Kluge y Wolf, 1993; Nixon, 1996): (1) las hipótesis obtenidas están basadas en la máxima evidencia disponible, (2) conjuntos diferentes de datos pueden ofrecer información a distintos niveles filogenéticos, (3) un conjunto de datos con señal filogenética débil debido a la presencia de mucha homoplasia, puede ver incrementada su señal por congruencia con otros caracteres de conjuntos distintos y (4) los caracteres combinados en una sola matriz pueden ser reestructurados durante el análisis y apoyar cladogramas que no se encontrarían en análisis separados. Un problema frecuente al analizar conjuntamente los datos es la existencia de taxones para los cuales se carece de información para alguna de las particiones. En estos casos las únicas opciones posibles son, o bien, sustituir dichos caracteres con interrogantes en el taxón incompleto, o bien, eliminar dichos taxones. Algunos autores han argumentado en contra de la combinación de los datos en caso que pueda demostrarse que los datos por separado apoyen árboles irreconciliables (Bull et al., 1993; De Queiroz, 1993; Huelsenbeck et al., 1996). El razonamiento es que dicha situación es consecuencia de que las particiones incongruentes están sometidas a diferentes procesos evolutivos que no pueden ser englobados bajo la aplicación de una sola metodología de inferencia filogenética.

### 3.9.3. Análisis conjunto condicional

Esta aproximación propone fundamentar la elección de uno de los dos tipos de análisis anteriormente expuestos en la demos-

tración previa de la congruencia o homogeneidad de los datos (Fig. 15). Evidentemente, la aplicación de dicha estrategia requiere la definición de un test estadístico que permita investigar la posible existencia de heterogeneidad entre las particiones. Aunque se han propuesto diversos índices para investigar la congruencia entre particiones utilizando parsimonia (Rodrigo et al., 1993; Templeton, 1983), el denominado test *ILD* (*Incongruence Length Difference*) (Farris, Källersjö, Kluge y Bult, 1994) ha sido considerado como el más útil (Cunningham, 1997a). Este test se basa en un índice anterior, el *ILD* (Mickelthwait y Farris, 1981), originalmente concebido para distinguir la homoplasia resultante de la combinación de matrices diferentes, de la debida a la incongruencia entre caracteres dentro de cada matriz. Este índice se define como:

$$ILD = S_{a+b} - (S_a + S_b)$$

donde,  $S_{a+b}$  es el número de pasos del cladograma resultante de la matriz combinada y  $S_a$ ,  $S_b$  los pasos de los árboles para cada matriz por separado. La distribución del estadístico *ILD* se construye calculando su valor original y los valores resultantes de la aleatorización de las particiones. Estas particiones aleatorizadas tienen el mismo tamaño que las originales, pero están formadas por caracteres seleccionados al azar entre las matrices combinadas, es decir, mezclando los caracteres de los dos particiones originales. La hipótesis nula es que existe congruencia entre las particiones, y es desestimada cuando la proporción de veces en que los valores de *ILD* aleatorizados son mayores o iguales al original, es menor que un cierto nivel de significación puede ser aplicado en presencia de ponderación diferencial de los datos y, también, simultáneamente a más de dos particiones de los datos. Por otra parte, se recomienda eliminar los caracteres invariables en el caso que las particiones difieran sensiblemente en el número de los mismos (Cunningham, 1997a). Finalmente, cabe destacar que algunos autores consideran que, si bien el índice *ILD* es útil para detectar la existencia de caracteres discordantes y el patrón de congruencia entre las diferentes particiones, las mismas deberían ser igualmente combinadas y analizadas simultáneamente incluso en el caso de que dicho test demostrara una heterogeneidad significativa (Nixon, 1996).

## BIBLIOGRAFÍA

- AGOSTI, D., JACOBS, D. & DESALLE, R., 1996. On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. *Cladistics*, **12**(1): 65-82.
- ALLARD, M. W. & CARPENTER, J. M., 1996. On weighting and congruence. *Cladistics*, **12**(3): 183-198.
- ALROY, J., 1994. Permutation test for the presence of phylogenetic structure. *Syst. Biol.*, **43**: 430-437.
- ARCHIE, J. W., 1989. A randomization test for phylogenetic information in systematic data. *Syst. Zool.*, **38**: 239-252.
- BREMER, K., 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, **42**: 795-803.
- BREMER, K., 1994. Branch support and tree stability. *Cladistics*, **10**: 295-304.
- BROTHERS, D. J. & CARPENTER, J. M., 1993. Phylogeny of Aculeata: Chrysioidea and Vespoidea (Hymenoptera). *J. Hym. Res.*, **2**: 227-304.
- BROWER, A. V. Z. & DESALLE, R., 1994. Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. *Ann. Entomol. Soc. Am.*, **87**: 702-716.
- BROWER, A. V. Z. & SCHAWARROCH, V., 1996. Three steps of homology assessment. *Cladistics*, **12**: 265-272.
- BROWN, J. K., 1994. Bootstrap hypothesis tests for evolutionary trees and other dendrograms. *Proc. Natl. Acad. Sci. U.S.A.*, **91**(25): 12293-12297.
- BROWN, W. M., PRAGER, E. M., WANG, A. & WILSON, A. C., 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, **18**: 225-239.
- BULL, J. J., HUELSENBECK, J. P., CUNNINGHAM, C. W., SWOFFORD, D. L. & WADDELL, P. J., 1993. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.*, **42**(3): 384-397.
- CAMIN, J. H. & SOKAL, R. R., 1965. A method for deducting branching sequences in phylogeny. *Evolution*, **19**: 311-326.
- CARPENTER, J. M., 1988. Choosing among multiple equally parsimonious cladograms. *Cladistics*, **4**: 291-296.
- CARPENTER, J. M., 1992. Random cladistics. *Cladistics*, **8**: 147-153.
- CARPENTER, J. M., 1994. Successive weighting, reliability and evidence. *Cladistics*, **10**(2): 215-220.
- CARPENTER, J. M., 1996. Uninformative bootstrapping. *Cladistics*, **12**(2): 177-181.
- CARPENTER, K. E., 1993. Optimal cladistic and quantitative evolutionary classifications as illustrated by fusilier fishes (Teleostei: Caesionidae). *Syst. Biol.*, **42**: 142-154.
- CAVALLI-SFORZA, L. L. & EDWARDS, A. W. F., 1967. Phylogenetic analysis: models and estimation procedures. *Evolution*, **32**: 550-570.

- CHAPPILL, J. A., 1989. Quantitative characters in phylogenetic analysis. *Cladistics*, 5: 217-234.
- CHIPPINDALE, P. T. & WIENS, J. J., 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. *Syst. Biol.*, 43: 278-287.
- CODDINGTON, J. A. & SCHARFF, N., 1994. Problems with zero-length branches. *Cladistics*, 10(4): 415-424.
- CODDINGTON, J. A. & SCHARFF, N., 1996. Problems with 'soft' polytomies. *Cladistics*, 12(2): 139-146.
- COLLINS, T. M., WIMBERGER, P. H. & NAYLOR, G. J. P., 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.*, 43: 482-496.
- CUNNINGHAM, C. W., 1997a. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.*, 14(7): 733-740.
- CUNNINGHAM, C. W., 1997b. Is congruence between data partitions a reliable predictor of phylogeny accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. *Syst. Biol.*, 46(3): 464-478.
- DE PINNA, M. C. C., 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics*, 7: 367-394.
- DE QUEIROZ, A., 1993. For consensus (sometimes). *Syst. Biol.*, 42(3): 368-372.
- DESALLE, R., 1992. The origin and possible time of divergence of the Hawaiian Drosophilidae: evidence from DNA sequences. *Mol. Biol. Evol.*, 9(5): 905-916.
- DIXON, M. T. & HILLIS, D. M., 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.*, 10(1): 256-267.
- DOWTON, M. & AUSTIN, A. D., 1997. Evidence for AT transversion bias in wasp (Hymenoptera, Symphita) mitochondrial genes and its implications for the origin of parasitism. *Journal of Molecular Evolution*, 44: 398-405.
- EDWARDS, A. W. F. & CAVALLI-SFORZA, L. L., 1963. The reconstruction of evolution. *Ann. Hum. Genet.*, 27: 105.
- EERNISSE, D. J. & KLUGE, A. G., 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Mol. Biol. Evol.*, 10(6): 1170-1195.
- EFRON, B., 1979. Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7: 1-26.
- EFRON, B., 1982. *The jackknife, the bootstrap and other resampling schemes*. Philadelphia: Society for industrial and applied mathematics.
- EFRON, B., 1987. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.*, 82: 171-185.
- ELDRIDGE, N. & CRACRAFT, J., 1980. *Phylogenetic patterns and the evolutionary process*. New York: Columbia University Press.
- FAITH, D. P., 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Cladistics*, 40(3): 366-375.
- FAITH, D. P. & CRANSTON, P. S., 1991. Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. *Cladistics*, 7: 1-28.
- FARRIS, J. S., 1969. A successive approximations approach to character weighting. *Syst. Zool.*, 18: 374-385.
- FARRIS, J. S., 1970. Methods for computing Wagner trees. *Syst. Zool.*, 19: 83-92.
- FARRIS, J. S., 1974. Formal definitions of paraphyly and polyphyly. *Syst. Zool.*, 23: 548-554.
- FARRIS, J. S., 1977. Phylogenetic analysis under Dollo's Law. *Syst. Zool.*, 26: 77-88.
- FARRIS, J. S., 1982. Outgroups and parsimony. *Syst. Zool.*, 31: 328-334.
- FARRIS, J. S., 1983. The logical basis of phylogenetic systematics. En: N. I. PLATNICK & V. A. FUNK (Eds.) *Advances in cladistics*. (Vol. 2, pp. 7-36). New York: Columbia University Press.
- FARRIS, J. S., 1988. *HENNIG 86* (Version 1.5). Port Jefferson Station, New York.
- FARRIS, J. S., 1989. The retention index and the rescaled consistency index. *Cladistics*, 5: 417-419.
- FARRIS, J. S., 1991. Excess homoplasy ratios. *Cladistics*, 7: 81-91.
- FARRIS, J. S., KALLERSJO, M., KLUGE, A. G. & BULT, C., 1994. Testing significance of incongruence. *Cladistics*, 10(3): 315-320.
- FELSENSTEIN, J., 1978. Cases in which parsimony or compatible methods will be positively misleading. *Syst. Zool.*, 27: 27-33.
- FELSENSTEIN, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39: 783-791.
- FELSENSTEIN, J., 1988a. Phylogenies and quantitative characters. *Ann. Rev. Ecol. Syst.*, 19: 445-471.
- FELSENSTEIN, J., 1988b. Phylogenies from molecular sequences. *Annu. Rev. Genet.*, 22: 521-565.
- FELSENSTEIN, J. & KISHINO, H., 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.*, 42: 193-200.
- FITCH, W. M., 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, 20: 404-416.
- FITCH, W. M., 1993. Commentary on the letter from Ward C. Wheeler. *Mol. Biol. Evol.*, 10(3): 713-714.
- FITCH, W. M. & YE, J., 1991. Weighted parsimony: Does it work? En: M. M. MIYAMOTO & J. CRACRAFT (Eds.) *Phylogenetic Analysis of DNA sequences*. Oxford University Press, New York: 147-154.
- FOREY, P. L., HUMPHRIES, C. J., KITCHING, I. J., SCOTLAND, R. W., SIEBERT, D. J. & WILLIAMS, D. M., 1992. *Cladistics, a practical course in systematics*. Clarendon Press, Oxford.
- GATESY, J., DESALLE, R. & WHEELER, W., 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol. Phylogenet. Evol.*, 2(2): 152-157.
- GINGERICH, P. D., 1979. Stratophenetic approach to phylogeny reconstruction in vertebrate paleontology. En: J. CRACRAFT & N. ELDRIDGE (Eds.) *Phylogenetic analysis and paleontology*. Columbia University Press, New York: 41-77.
- GLADSTEIN, D. S. & WHEELER, W. C., 1997. *POY: The Optimization of Alignment Characters*. American Museum of Natural History, New York.
- GOLOBOFF, P. A., 1991. Homoplasy and the choice among cladograms. *Cladistics*, 7: 215-232.
- GOLOBOFF, P. A., 1993. Estimating character weights during tree search. *Cladistics*, 9: 83-91.
- GOLOBOFF, P. A., 1995. Parsimony and weighting: a reply to Turner and Zandee. *Cladistics*, 11: 91-104.
- GOLOBOFF, P. A., 1996a. *Nona* (Version 1.5.1). American Museum of Natural History, New York.
- GOLOBOFF, P. A., 1996b. *Pee-Wee* (Version 2.5.1). American Museum of Natural History, New York.
- GOLOBOFF, P. A., 1996c. *PHAST* (Version 1.1). American Museum of Natural History, New York.
- GOLOBOFF, P. A., 1996d. *SPA* (Version 1.1). American Museum of Natural History, New York.
- GUSTAFSSON, M. H. G. & BREMER, K., 1995. Morphology and phylogenetic interrelationships of the Asteraceae, Calyceraceae, Campanulaceae, Goodeniaceae, and related families (Asterales). *Am. J. Bot.*, 82: 250-265.
- HEDGES, S. B., 1992. The number of replications needed for accurate bootstrap *P* value in phylogenetic studies. *Mol. Biol. Evol.*, 9: 366-369.
- HENDY, M. D. & PENNY, D., 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Math. biosci.*, 59: 277-290.
- HENNIG, W., 1953. Kritische Bemerkungen zur Phylogenetischen System der Insekten. *Beitr. Entomol.*, 3: 1-83.
- HENNIG, W., 1966. *Phylogenetic Systematics*. University of Illinois Press, Urbana, Chicago, London.
- HILLIS, D. M., 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. En: M. M. MIYAMOTO & J. CRACRAFT (Eds.) *Phylogenetic analysis of DNA sequences*. Oxford University Press, New York: 278-294.
- HILLIS, D. M., ALLARD, M. W. & MIYAMOTO, M. M., 1993. Analysis of DNA sequence data: phylogenetic inference. *Methods Enzymol.*, 224: 456-487.
- HILLIS, D. M. & BULL, J. J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.*, 42: 182-192.
- HILLIS, D. M. & HUELSENBECK, J. P., 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.*, 83(3): 189-195.
- HILLIS, D. M., MORITZ, C. & MABLE, B. K., 1996. *Molecular systematics*. Sinauer associates, Sunderland.
- HUELSENBECK, J. P., 1991. Tree-length distribution skewness: An indicator of phylogenetic information. *Syst. Zool.*, 40: 257-270.
- HUELSENBECK, J. P., BULL, J. J. & CUNNINGHAM, C. W., 1996. Combining data in phylogenetic analysis. *TREE*, 11: 152-158.
- HUELSENBECK, J. P., HILLIS, D. M. & JONES, R., 1995. Parametric bootstrapping in molecular phylogenetics: Applications and performance. En: J. FERRARIS & S. PALUMBI (Eds.) *Molecular Zoology: Strategies and Protocols*. Wiley, New York.
- JANVIER, P., 1984. Cladistics: theory, purpose and evolutionary implications. En: J. W. POLLARD (Ed.) *Evolutionary theory paths into the future*. Wiley, New York: 39-75.

- JONES, T. R., KLUGE, A. G. & WOLF, A. J., 1993. When theories and methodologies clash: a phylogenetic reanalysis of the North American Ambystomatid Salamanders (Caudata: Ambystomidae). *Syst. Biol.*, **42**(1): 92-102.
- KALLERSJÖ, M., FARRIS, J. S., KLUGE, A. G. & BULT, C., 1992. Skewness and permutation. *Cladistics*, **8**: 275-287.
- KITCHING, I. J., 1992. The determination of character polarity. En: P. L. FOREY, C. J. HUMPHRIES, I. J. KITCHING, R. W. SCOTLAND, D. J. SIEBERT & D. M. WILLIAMS (Eds.) *Cladistics. A practical course in Systematics*. Oxford University Press, Oxford: 22-43.
- KJER, K. M., 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. *Mol. Phylogenet. Evol.*, **4**(3): 314-330.
- KLASSEN, G. J., MOOI, R. D. & LOCKE, A., 1991. Consistency indices and random data. *Syst. Zool.*, **40**: 446-457.
- KLUGE, A. G., 1985. Ontogeny and phylogenetic systematics. *Cladistics*, **1**: 13-27.
- KLUGE, A. G., 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Syst. Zool.*, **38**: 7-25.
- KLUGE, A. G., 1997a. Sophisticated falsifications and research cycles: Consequences for differential character weighting in phylogenetic systematics. *Zool. Scr.*, **26**: 349-360.
- KLUGE, A. G., 1997b. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics*, **13**: 81-96.
- KLUGE, A. G. & FARRIS, J. S., 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.*, **18**: 1-32.
- KLUGE, A. G. & WOLF, A. J., 1993. Cladistic: What's in a word? *Cladistics*, **9**: 183-199.
- KNIGHT, A. & MINDELL, D. P., 1993. Substitution bias, weighting of DNA sequence evolution, and the phylogenetic position of Fea's viper. *Syst. Biol.*, **42**: 18-31.
- KNIGHT, A., STYER, D., PELIKAN, S., CAMPBELL, J. A., DENSMORE, L. D., III & MINDELL, D. P., 1993. Choosing among hypotheses of rattlesnake phylogeny: A best-fit rate test for DNA sequence data. *Syst. Biol.*, **42**: 356-367.
- KOCHER, T. D., THOMAS, W. K., MEYER, A., EDWARDS, S. V., PÄÄBO, S., F.X., V. & WILSON, A. C., 1989. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc. Natl. Acad. Sci. U.S.A.*, **86**: 6196-6200.
- LI, W. H., WU, C. I. & LUO, C. C., 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudo genes and its evolutionary implications. *J. Mol. Evol.*, **21**: 58-71.
- MADDISON, D. R., 1990. *Phylogenetic inference of historical pathways and models of evolutionary change*. PhD. thesis, Harvard University.
- MADDISON, D. R., 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Zool.*, **40**: 315-328.
- MADDISON, W. P., 1993. Missing data versus missing characters in phylogenetic analysis. *Syst. Biol.*, **42**: 576-581.
- MADDISON, W. P. & MADDISON, D. R., 1992. *MacClade. Analysis of phylogeny and character evolution. Version 3*. Sinauer Associates, Sunderland.
- MAYR, E. & ASHLOCK, P. D., 1991. *Principles of systematic zoology*. (2 ed.). McGraw Hill, New York.
- MICKEVICH, M. F. & FARRIS, J. S., 1981. The implications of incongruence in *Menidia*. *Sys. Zool.*, **30**: 351-370.
- MIYAMOTO, M. M., 1985. Consensus cladograms and general classifications. *Cladistics*, **1**: 186-189.
- MIYAMOTO, M. M. & FITCH, W. M., 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.*, **44**: 64-76.
- NEFF, N. A., 1986. A rational basis for a priori character weighting. *Syst. Zool.*, **35**(1): 110-123.
- NEI, M., 1996. Phylogenetic analysis in molecular evolutionary genetics. *Ann. Rev. Genet.*, **30**: 371-403.
- NELSON, G. & PLATNICK, N. I., 1981. *Systematics and Biogeography: Cladistics and Vicariance*. Columbia Univ. Press, New York.
- NIXON, K. C., 1996. On simultaneous analysis. *Cladistics*, **12**(3): 221-242.
- NIXON, K. C. & CARPENTER, J. M., 1993. On outgroups. *Cladistics*, **9**: 413-426.
- NIXON, K. C. & CARPENTER, J. M., 1996. On consensus, collapsibility, and clade concordance. *Cladistics*, **12**: 305-321.
- NIXON, K. C. & DAVIS, J. I., 1991. Polymorphic taxa, missing values and cladistic analysis. *Cladistics*, **7**: 233-242.
- NIXON, K. C. & WHEELER, Q. D., 1990. An amplification of the phylogenetic species concept. *Cladistics*, **6**: 211-224.
- PANCHEN, A. L., 1979. The cladistic debate, continued. *Nature*, **280**: 541.
- PANCHEN, A. L., 1992. *Classification, evolution and the nature of biology*. Cambridge University Press, Cambridge.
- PATTERSON, C., 1982. Morphological characters and homology. En: K. A. JOYSEY & A. E. FRIDAY (Eds.) *Problems of phylogenetic reconstruction*. Academic Press, London, New York: 21-74.
- PATTERSON, C., 1988. Homology in classical and molecular biology. *Mol. Biol. Evol.*, **5**: 603-625.
- PATTERSON, C. & JOHNSON, G. D., 1997. The data, the matrix and the message: comments on Begle's "Relationships of the osmeroid fishes". *Syst. Biol.*, **46**(2): 358-365.
- PENNY, D., HENDY, M. D. & STEEL, M. A., 1992. Progress with methods for constructing evolutionary trees. *TREE.*, **7**(3): 73-79.
- PLATNICK, N. I., CODDINGTON, J. A., FORSTER, R. R. & GRISWOLD, C. E., 1991a. Spinneret morphology and the phylogeny of haplogynae spiders (Araneae, Araneomorpha). *A. Mus. Novit.*, **3016**: 1-73.
- PLATNICK, N. I., GRISWOLD, C. E. & CODDINGTON, J. A., 1991b. On missing entries in cladistic analysis. *Cladistics*, **7**: 337-344.
- PLEIJEL, F., 1995. On character coding for phylogeny reconstruction. *Cladistics*, **11**(3): 309-315.
- POPPER, K. R., 1968a. *Conjectures and refutations*. Harper Torchbooks, New York.
- POPPER, K. R., 1968b. *The logic of scientific discovery*. Harper Torchbooks, New York.
- QUICKE, D. L. J., 1993. *Principles and techniques of contemporary taxonomy*. Chapman y Hall, London.
- REMANE, A., 1952. *Die grundlagen des naturlichen systems, der vergleichenden anatomie und der phylogenetik*. Geest und Portig, Leipzig.
- RODRIGO, A. G., 1992. A modification to Wheeler's combinatorial weights calculations. *Cladistics*, **8**: 165-170.
- RODRIGO, A. G., 1993. Calibrating the bootstrap test of monophyly. *Int. J. Parasitol.*, **23**(4): 507-514.
- RODRIGO, A. G., KELLY-BORGES, M., BERGQUIST, P. R. & BERGQUIST, P. L., 1993. A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *N. Zealand J. Bot.*, **31**: 257-268.
- RZHETSKY, A. & NEI, M., 1992. A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.*, **9**: 945-967.
- SAITOU, N. & NEI, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**: 406-425.
- SANDERSON, M. J., 1995. Objections to bootstrapping phylogenies: A critique. *Syst. Biol.*, **44**: 299-320.
- SANDERSON, M. J. & DONOGHUE, M. J., 1989. Patterns of variation in levels of homoplasy. *Evolution*: 1781-1795.
- SANKOFF, D. & CEDERGREN, R. J., 1983. Simultaneous comparison of three or more sequences related by a tree. En: D. SANKOFF & J. B. KRUSKAL (Eds.) *Time warps, string edits and macromolecules. The theory and practice of sequence comparison*. Addison-Wesley, Reading, Massachusetts: 253-263.
- SANKOFF, D., MOREL, C. & CEDERGREN, R. J., 1973. Evolution of 5S RNA and the non-randomness of base replacement. *Nature*, **245**: 232-234.
- SANKOFF, D. & ROUSSEAU, P., 1975. Locating the vertices of a Steiner tree in arbitrary space. *Math. Prog.*, **9**: 240-246.
- SCHARFF, N. & CODDINGTON, J. A., 1997. A phylogenetic analysis of the orb-weaving spider family Araneidae (Arachnida, Araneae). *Zool. J. Linn. Soc.*, **120**: 355-434.
- SCOTLAND, R. W., 1992. Cladistic theory. En: P. L. FOREY, C. J. HUMPHRIES, I. J. KITCHING, R. W. SCOTLAND, D. J. SIEBERT & D. M. WILLIAMS (Eds.) *Cladistics. A practical course in systematics*. Oxford University Press, Oxford: 3-12.
- SHARKEY, M. J., 1989. A hypothesis-independent method of character weighting for cladistic analysis. *Cladistics*, **5**: 63-86.
- SIDALL, M. E., 1998. Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone. *Cladistics*, **14**(3): 209-220.
- SIDDALL, M. E., 1995. Another monophyly index: Revisiting the jackknife. *Cladistics*, **11**(1): 33-56.
- SIEBERT, D. J., 1992. Tree statistics; trees and 'confidence'; consensus trees; alternatives to parsimony; character weighting; character conflict and its resolution. En: P. L. FOREY, C. J. HUMPHRIES, I. J. KITCHING, R. W. SCOTLAND, D. J. SIEBERT & D. M. WILLIAMS (Eds.) *Cladistics. A practical course in systematics*. Oxford University Press, Oxford: 72-88.
- SIMON, C., FRATI, F., BECKENBACH, A., CRESPI, B., LIU, H. & FLOOK, P., 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences. *Ann. Entom. Soc. Am.*, **87**: 651-701.

- SIMPSON, G. G., 1961. *Principles of animal taxonomy*. Columbia University Press, New York.
- SNEATH, P. H. A. & SOKAL, R. R., 1973. *Numerical taxonomy*. Freeman, W.H., San Francisco.
- SOBER, E., 1988. *Reconstructing the past: Parsimony, evolution and inference*. Cambridge: M.I.T. Press.
- SOKAL, R. R. & ROHLF, F. J., 1981. *Biometry*. Freeman, W.H., San Francisco.
- STEVENS, P. F., 1991. Character states, continuous variation and phylogenetic analysis: a review. *Syst. Bot.*, **16**: 553-583.
- STURMBAUER, C. & MEYER, A., 1992. Genetic divergence, speciation and morphological stasis in a lineage of African chichlid fishes. *Nature*, **358**: 578-581.
- SWOFFORD, D. L., 1993. *PAUP: Phylogenetic analysis using parsimony (Version 3.1.1)*. Illinois Natural History Survey, Illinois.
- SWOFFORD, D. L. & MADDISON, W. P., 1987. Reconstructing ancestral character states under Wagner parsimony. *Math. biosci.*, **87**: 199-229.
- SWOFFORD, D. L. & OLSEN, G. J., 1990. Phylogeny reconstruction. En: D. M. HILLIS & C. MORITZ (Eds.) *Molecular systematics*. Sinauer Associates, Sunderland: 411-501.
- SWOFFORD, D. L., OLSEN, G. J., WADDLE, P. J. & HILLIS, D. M., 1996a. Phylogenetic inference. En: D. M. HILLIS, C. MORITZ & B. K. MABLE (Eds.) *Molecular systematics* (2nd ed.). Sinauer Associates, Sunderland: 407-514.
- SWOFFORD, D. L., THORNE, J. L., FELSENSTEIN, J. & WIEGMANN, B. M., 1996b. The topology-dependent permutation test for monophyly does not test for monophyly. *Syst. Biol.*, **45**(4): 575-579.
- TEMPLETON, A. R., 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the humans and apes. *Evolution*, **37**: 221-244.
- TRUEMAN, J. W. H., 1993. Randomization confounded: A response to Carpenter. *Cladistics*, **9**: 101-109.
- TRUEMAN, J. W. H., 1996. Permutation tests and outgroups. *Cladistics*, **12**(3): 253-262.
- VAWTER, L. & BROWN, W. M., 1993. Rates and patterns of base change in the small subunit ribosomal RNA gene. *Genetics*, **134**: 597-608.
- WHEELER, Q. D., 1986. Character weighting and cladistic analysis. *Syst. Zool.*, **35**(1): 102-109.
- WHEELER, W. C., 1990a. Combinatorial weights in phylogenetic analysis: A statistical parsimony procedure. *Cladistics*, **6**: 269-276.
- WHEELER, W. C., 1990b. When is an outgroup not an outgroup and how to root DNA sequence based topologies without an outgroup. *Cladistics*, **6**: 363-367.
- WHEELER, W. C., 1992. Extinction, sampling, and molecular phylogenetics. En: M. J. NOVACEK & Q. D. WHEELER (Eds.) *Extinction and phylogeny*. Columbia University Press, New York: 205-215.
- WHEELER, W. C., 1993. The triangle inequality and character analysis. *Mol. Biol. Evol.*, **10**(3): 707-712.
- WHEELER, W. C., 1996. Optimization alignment: The end of multiple sequence alignment in phylogenetics? *Cladistics*, **12**: 1-9.
- WHEELER, W. C. & GLADSTEIN, D., 1994. MALIGN: a multiple nucleic acid sequence alignment program. *J. Hered.*, **85**: 417.
- WHEELER, W. C. & HONEYCUTT, R. L., 1988. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. *Mol. Biol. Evol.*, **5**(1): 90-96.
- WILEY, E. O., 1981. *Phylogenetics. The theory and practice of phylogenetic systematic*. John Wiley y sons, New York.
- WILEY, E. O., SIEGEL-CAUSEY, D., BROOKS, D. R. & FUNK, V. A., 1991. *The complete cladist*. The University of Kansas, Museum of Natural History, Special Publications, Lawrence.
- WILKINSON, M., 1994. Common cladistic information and its consensus representation: Reduced Adams and reduced cladistic consensus trees and profiles. *Syst. Biol.*, **43**: 343-368.
- WILKINSON, M., 1995a. A comparison of two methods of character construction. *Cladistics*, **11**(3): 297-308.
- WILKINSON, M., 1995b. Coping with abundant missing entries in phylogenetic inference using parsimony. *Syst. Biol.*, **44**: 501-514.
- WILLIAMS, D. M., 1992. DNA analysis: Theory. En: P. L. FOREY, C. J. HUMPHRIES, I. J. KITCHING, R. W. SCOTLAND, D. J. SIEBERT & D. M. WILLIAMS (Eds.) *Cladistics. A practical course in systematics*. Oxford University Press, Oxford: 89-101.
- WILLIAMS, P. L. & FITCH, W. M., 1989. Finding the minimal change in a given tree. En: B. FERNHÖLM, K. BREMER & H. JÖRNVALL (Eds.) *The hierarchy of life. Molecules and morphology in phylogenetic analysis*. Elsevier Science Pub., Amsterdam: 453-470.
- WILLIAMS, P. L. & FITCH, W. M., 1990. Phylogeny determination using a dynamically-weighted parsimony method. En: R. F. DOOLITTLE (Ed.) *Molecular evolution: Computer analysis of protein and nucleic acid sequences*. Academic Press, New York: 615-625.
- YANG, Z., 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.*, **43**: 329-342.
- YANG, Z., 1996. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, **42**: 294-307.
- ZHARKIKH, A. & LI, W. H., 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol.*, **4**(1): 44-63.