Bol. Soc. Bot. México **59**: 115-129 (1996)

CODIFICACIÓN DE LAS INSERCIONES-DELECIONES EN EL ANÁLISIS FILOGENÉTICO DE SECUENCIAS GÉNICAS

DOLORES GONZÁLEZ

Departamento de Sistemática Vegetal, Instituto de Ecología, A.C.
Apartado Postal 63, Xalapa, Veracruz, 91000, México

Resumen. El concepto general de homología filogenética significa que un carácter es cualquier aspecto del fenotipo o genotipo que varía dentro del grupo de estudio. En las secuencias cada base corresponde a una hipótesis de homología transformacional o táxica dependiendo si los nucleótidos varían o no en cada posición. Desde el punto de vista teórico, la dificultad de asignar un código apropiado para las indels estriba en que éstas se pueden interpretar como un artificio o como información filogenética (homología táxica y transformacional). Una revisión de los distintos códigos que se emplean para el análisis de las indels indicó que los seis códigos usados son: (1) ambiguas, (2) excluidas, (3) quinto estado, (4) presencia/ausencia, (5) diferente longitud/varios estados y (6) diferente longitud/varios caracteres. Cada uno de estos códigos se evaluó en la reconstrución de hipótesis filogenéticas con una matriz de 20 secuencias del gen de ARN ribosomal del hongo fitopatógeno *Rhizoctonia solani*. Los efectos se midieron en términos de los cambios en homoplasia, topológicos, de resolución y en el número de árboles. Los efectos son distintos dependiendo del código empleado. El efecto más radical es causado por la codificación de las indels como diferentes caracteres para diferentes longitudes. Los análisis cladísticos con secuencias génicas deben incorporar una fase de evaluación de los efectos de distintos códigos para las indels dado que potencialmente son informativas.
Palabras clave: sistemática molecular, análisis de caracteres, homología, indels, análisis cladísticos.
Abstract. The general concept of phylogenetic homology visualizes that a character is any aspect of the phenotype or genotype that varies inside a group. In gene sequences each position corresponds to a hypothesis of transformational or taxic homology depending if the nucleotides vary or not in each position. It is theoretically difficult to designate a proper code to an indel. This is because an indel can be interpreted as an artifact or as phylogenetic information (taxic or transformational homology). A review of the distinct codes employed for the analysis of the indels shows that the six codes used are: (1) missing, (2) excluded, (3) fifth state, (4) presence/absence, (5) different length/several states and (6) different length/several characters. Each code was evaluated in the phylogenetic hypotheses of the phytopathogenic fungi *Rhizoctonia solani* with a matrix of 20 sequences of the ribosomal ARN genes. The effects of the distinct codes were measured in changes of homoplasy, topology, resolution and tree number. The effects differ depending on the code used. The most radical effect was caused by the coding of the indels as different length/several characters. The cladistic analyses with gene sequences should include an evaluation of the effects of the different codes for the indels, since they may be potentially informative.
Key words: molecular systematics, character analysis, homology, indels, cladistic analysis.

El objetivo central de la sistemática filogenética es construir clasificaciones basadas en hipótesis de interrelaciones entre grupos hermanos y que estén apoyadas por caracteres homólogos. Potencialmente un carácter es cualquier aspecto del fenotipo o genotipo que es comparable y varía dentro del grupo de estudio. Las fuentes de caracteres son diversas como la morfología, la fisiología, la ultraestructura y las secuencias génicas entre otros. Todos los tipos de caracteres pueden contener información histórica a algún nivel jerárquico particular (Sober, 1988; Hall, 1994). Lo que cuenta es que el carácter sea variable,

es decir, que muestre mayor variación entre la colección de unidades que dentro de las unidades. También, la variación debe ser heredable e independiente de otros caracteres. En resumen, un carácter informativo es un sistema de por lo menos dos homólogos transformacionales discretos empíricamente reconocidos como estados (Stevens, 1991; De Luna y Mishler, 1997).

Los caracteres y estados derivados de las secuencias génicas se pueden establecer a varios niveles de organización estructural. Al nivel más amplio, el gen puede ser considerado como un carácter y las condiciones alternativas del gen constituyen los estados. Por ejemplo, en un estudio con alozimas, el gen para LDH puede ser considerado el carácter y los estados serían los diferentes electromorfos. A su vez, regiones particulares dentro de un gen pueden ser considerados el carácter, en cuyo caso los estados serían por ejemplo la presencia o ausencia de un sitio de restricción. Al nivel más fino, cada posición en la secuencia de un gen también puede representar un carácter y los nucleótidos en esa posición son los estados. En todos los niveles, cada alternativa variable representa un estado y por lo tanto implica una hipótesis de homología filogenética. En las secuencias génicas si los nucleótidos son diferentes en la misma posición entonces cada uno corresponde a una hipótesis de homología transformacional. Si los nucleótidos no varían en una posición entonces corresponde a una hipótesis de homología táxica (Mindell, 1991).

El análisis de las secuencias de nucleótidos se inicia con su alineamiento. Con este procedimiento se establecen las hipótesis de homología táxica y transformacional a nivel de cada nucleótido. Dado que a menudo existen diferencias en la longitud de las secuencias entre taxa, para la misma colección puede haber alineamientos alternativos (Lake, 1991; Mindell, 1991; Waterman *et al.*, 1991; Knight y Mindell, 1995). Existen varios métodos para alinear las secuencias y establecer la correspondencia entre cada nucleótido, algunos son manuales y otros automáticos (Needleman y Wunsch, 1970; Swofford y Olsen, 1990; Mindell, 1991; Waterman *et al.*, 1991). Como consecuencia de maximizar la correspondencia entre nucleótidos frecuentemente es necesario insertar posiciones vacías intercaladas en las secuencias génicas de algunos taxa. Estos espacios se han denominado "gaps" y se han interpretado como inserciones o deleciones ("indels"). En lo sucesivo a estas posiciones vacías o "gaps" se les llamará indels. Una vez que ya se ha aceptado un alineamiento se han establecido los caracteres y los estados. La siguiente fase de análisis la constituye la búsqueda de árboles filogenéticos. Estos análisis con secuencias génicas se han llevado a cabo bajo varios enfoques tanto fenéticos como filogenéticos (Felsenstein, 1984, 1988; Swofford y Olsen, 1990; Nei, 1991; Williams, 1993).

El análisis cladístico de secuencias génicas presenta aspectos metodológicos adicionales a los que son conocidos para el análisis de datos morfológicos. En la literatura ya se han señalado los problemas asociados a la representatividad de la variación interna de las unidades de estudio. Se ha cuestionado el uso de un solo organismo para representar un taxon (Baverstock y Moritz, 1990; Miyamoto y Cracraft, 1991). También se han dado recomendaciones para el muestreo de genes apropiados, ya que algunos se caracterizan por tasas de mutación muy lentas o muy rápidas (Friedlander *et al.*, 1992; Graybeal, 1994; González, en revisión). Del mismo modo, se han examinado diferentes estrategias para asignar pesos a caracteres o a estados. Esto obedece a que se ha encontrado que las transiciones son más frecuentes que las transversiones (Wheeler, 1990; Williams y Fitch, 1990; Knight y Mindell, 1993, 1995). También se han generado varios métodos para codificar la variación de las secuencias. La forma más común es considerar a los cuatro nucleótidos como cuatro estados potenciales en cada posición: adenina, guanina, citosina y timina. Otra forma de recodificación de las secuencias génicas es la traducción de los nucleótidos a aminoácidos. De este modo cada carácter puede tener hasta 20 posibles estados que corresponden a los 20 aminoácidos que componen las proteínas. Un sistema menos común de codificación consiste en recodificar los cuatro nucleótidos a sólo dos estados. Esto se logra con la asignación de un código para las purinas y otro código para las pirimidinas (Lake, 1987). Sin embargo, un problema que no ha sido suficientemente explorado en la literatura es el de la codificación de las indels. En la mayoría de los casos es común que las indels se codifiquen como "?" y con esto se han eliminado de los análisis. A pesar de la gran cantidad de estudios filogenéticos con secuencias génicas es sorprendente que no se haya resuelto cómo codificarlas.

Desde el punto de vista teórico, la dificultad estriba en que una indel generada por el alineamiento se puede interpretar como un artificio o como información filogenética (homología táxica y transformacional). Desde el punto de vista metodológico, sea que una indel es un artificio o una hipótesis de homología, no se ha puesto atención en generar estrategias de codificación que adecuadamente representen una u otra alternativa. De hecho algunos programas de búsqueda de árboles y de optimización de caracteres (por ejemplo DNAml) ni siquiera brindan la oportunidad de considerar las indels como estados (Buckler y Holtsford, 1996) ya que éstas automáticamente

se interpretan como posiciones no informativas. En la presente contribución se analiza el problema de la codificación de las indels en los análisis cladísticos. Para ello, se hace una revisión breve del concepto de homología como base teórica para el análisis de caracteres al nivel de secuencias génicas. Se describe cómo surgen las indels en la matriz de datos y se presenta un resumen de los distintos códigos usados para su inclusión en los análisis filogenéticos. Finalmente, se hace una evaluación preliminar del efecto de cada uno de estos códigos en la reconstrución de hipótesis filogenéticas. Los efectos se miden en términos de los cambios en homoplasia, topológicos, de resolución y en el número de árboles. Para esta evaluación se usó una matriz de secuencias del gen de ARN ribosomal de veinte muestras del hongo fitopatógeno *Rhizoctonia solani* (González, 1992).

Homología a nivel de secuencias génicas

El término homología se ha referido a una gran diversidad de conceptos que tratan de describir una correspondencia histórica o relación mecanística entre los distintos rasgos de organismos (Wiley, 1981; Wagner, 1989). En particular el concepto de homología filogenética define la correspondencia histórica entre los caracteres de diferentes organismos (Patterson, 1982; Hall, 1994). En la sistemática filogenética se han propuesto criterios explícitos para examinar las hipótesis de homología (Patterson, 1982, 1988). Empíricamente, los homólogos se reconocen mediante los criterios de similitud, conjunción, heredabilidad e independencia. Inferencialmente, los homólogos se corroboran mediante la prueba de congruencia. Si el carácter es consistente con el patrón sugerido por otros caracteres en un cladograma parsimonioso entonces se hipotetiza que es homólogo (Patterson, 1982; De Pinna, 1991; Nelson, 1994; Roth, 1994).

En el caso de los datos moleculares se ha sugerido que deben existir conceptos diferentes de homología respecto a los que se aplican a los datos morfológicos. Se ha promovido el concepto de homología molecular como la correspondencia de similitud a varios niveles (nucleótido, gene, genoma) debida a ancestría común (Hillis, 1994). Algunos autores han argumentado que para identificar los caracteres homólogos, la similitud es el criterio más importante en comparación con la prueba de congruencia (Patterson, 1988). No obstante, otros autores han argumentado que epistemológicamente no hay diferencia entre el concepto de homología filogenética cuando se aplica a datos moleculares o cualquier otro tipo. La similitud es base empírica pero no puede ser argumento lógico para corroborar. El único criterio robusto para el examen de homología molecular es la congruencia entre caracteres (De Pinna, 1991; Mindell, 1991; De Luna y Mishler, 1997).

En la literatura de la sistemática molecular se ha recomendado el uso de genes ortólogos para análisis filogenéticos. Los genes se han denominado ortólogos si su similitud se debe a ancestría común y se han usado varios términos para las similitudes no históricas (parálogos, xenólogos, paraxenólogos, vease la revisión por Patterson, 1988). En cualquier caso, la prescripción de seleccionar genes ortólogos para estudios filogenéticos es irrelevante. Conceptualmente se ha debatido si la aplicación de estos calificativos sin relación a un cladograma específico es apropiado o no (De Pinna, 1991; Nelson, 1994; De Luna y Mishler, 1996). De hecho, los análisis filogenéticos comunmente no se basan en los genes como caracteres sino que cada posición en la secuencia es un carácter potencial. Es obvio entonces que después del análisis cladístico se puede revelar que un gen supuestamente ortólogo contiene varias posiciones no homólogas. Es decir, en una posición el mismo nucleótido es compartido por varios taxa cuyo origen no es por ancestría común (homoplasia). Paradójicamente, el análisis de un gen supuestamente parálogo puede identificar nucleótidos congruentes con otros caracteres moleculares o morfológicos entre taxa, es decir, tales genes podrían contener posiciones filogenéticamente homólogas. Esto señala lo ilusorio de la posición epistemológica sugerida por Moritz y Hillis (1990) de seleccionar sólo genes ortólogos para los análisis filogenéticos. Por un lado, esta selección no es posible antes del análisis cladístico ya que la historia misma del gene está bajo evaluación. Por otro lado, el foco del análisis no son los genes sino cada una de las posiciones como hipótesis independientes de homología. Sólo la congruencia es epistemológicamente robusta para la selección de caracteres informativos en el análisis filogenético de los datos moleculares.

La intercalación de espacios entre los nucleótidos de una secuencia es inevitable al maximizar la correspondencia entre secuencias durante el alineamiento. El problema conceptual con estas indels es decidir si son equivalentes a hipótesis de homología o son simplemente un artificio. Los principios conceptuales de homología filogenética se deben aplicar a todos los caracteres, sean morfológicos o moleculares.

Metodológicamente, el criterio de similitud se aplica de modo semejante a cualquier tipo de datos. En el caso de las secuencias génicas la similitud posicional es la base empírica para postular homología entre una o varias indels y un nucleótido presentes en la misma posición en dos o más taxa. Por ejemplo, se pue-

de argumentar que una indel en un taxon en comparación con un nucleótido en otro taxon satisface el criterio de similitud porque es la misma posición. Del mismo modo dos indels de dos taxa cumplen el criterio de similitud posicional. Sin embargo, inicialmente, no se sabe si esta similitud es filogenéticamente informativa o no, al igual que sucede con el caso de caracteres morfológicos. Entonces, todas las similitudes a nivel de cada posición (incluyendo las indels) se pueden postular como hipótesis de homología potencial. La congruencia entre cada hipótesis independiente resolverá si una posición en particular (por ejemplo, la que contiene una indel) puede mantenerse como hipótesis de homología filogenética. La decisión de si un nucleótido o una indel es informativa o no por lo tanto depende de incluirla en el análisis de parsimonia y ver si es congruente con otros caracteres (incluso otras indels). Por lo tanto es importante escoger un código apropiado no sólo para el análisis de los nucleótidos sino también para el de las indels potencialmente informativas.

Elaboración de la matriz de datos de secuencias génicas

Todos los pasos necesarios para construir una matriz de datos se denominan "análisis de caracteres" y en el caso de las secuencias esta fase incluye el alineamiento y la codificación. El alineamiento de las secuencias equivale al proceso mismo de la proposición de las hipótesis de homología a nivel de cada posición. Si la variación de las secuencias del gen que se está comparando entre varios taxa se debe únicamente a las substituciones (transiciones o transversiones) entonces la longitud de las secuencias es igual y el alineamiento es directo. Sin embargo, cuando las mutaciones resultan en inserciones o deleciones en alguna de las secuencias es necesario definir las posiciones comparables entre los taxa. Para ello, comunmente se deben introducir uno o varios espacios para mantener la similitud posicional entre las secuencias. En lo sucesivo, un espacio entre dos nucleótidos se denominará indel simple y dos o más espacios contiguos se denominarán indels múltiples. Las indels múltiples, a su vez, pueden ser de la misma o distinta longitud.

Las indels simples y múltiples se presentan en todos los genes y organismos aunque su frecuencia varía entre las diferentes regiones del genoma. Las indels son relativamente raras en los genes codificadores de proteínas, pero son comunes en las regiones que forman las asas de los genes de ARN ribosomal o en secuencias no codificadoras para proteínas. Generalmente, el alineamiento óptimo incluye indels que pueden ser de un nucleótido o de varios nucleótidos contiguos. En la figura 1 se muestra el alineamiento de las secuencias de cinco UTO's. En el alineamiento 1 se presentan las secuencias obtenidas directamente del gel mientras que en el alineamiento 2 se presentan las secuencias después de que se ha maximizado su similitud. Por ejemplo, los estados comparables en la posición 2 del alineamiento 1 son "cttcc" mientras que en el alineamiento 2 los estados son "c--cc" para cada uno de los cinco taxa.

Se han desarrollado diversos algoritmos para el alineamiento automático de secuencias que optimizan la similitud entre ellas penalizando a las substituciones y a las indels (Fitch y Smith, 1983). El valor de la penalización es arbitrario (Hein, 1989; Mindell, 1991). Si se asignan penalizaciones altas (por ejemplo, 10 substituciones a 1 indel) cada posición donde no hay cambio el costo es 0.0, si hay una substitución el costo es de 0.1 mientras que una indel vale 1.0. Los programas agregan el costo de todas las posiciones en un índice total de la comparación de cada par de secuencias y se selecciona el alinea-

Alineamiento 1

		posición 2
UTO 1	actaagcgtagctgct	c
UTO 2	ataagcgtagctagct	t
UTO 3	ataagcgtagcgct	t
UTO 4	actaagcgtagctaagct	c
UTO 5	actaagcgtagctgct	c

Alineamiento 2

UTO 1	actaagcgtagct--gct	c
UTO 2	a-taagcgtagcta-gct	-
UTO 3	a-taagcgtagc---gct	-
UTO 4	actaagcgtagctaagct	c
UTO 5	actaagcgtagct--gct	c

Figura 1. Dos alineamientos posibles de las secuencias génicas de cinco UTO's. El alineamiento 1 muestra las secuencias como son leídas directamente del gel. El alineamiento 2 muestra las secuencias después de que se ha optimizado su similitud total. El alineamiento óptimo incluye una serie de inserciones-deleciones que pueden ser de un nucleótido o de varios nucleótidos contiguos. Por ejemplo, los nucleótidos comparables de la posición 2 en el alineamiento 1 son cttcc. Después que se ha optimizado la similitud de las secuencias los nucleótidos comparables son c--cc.

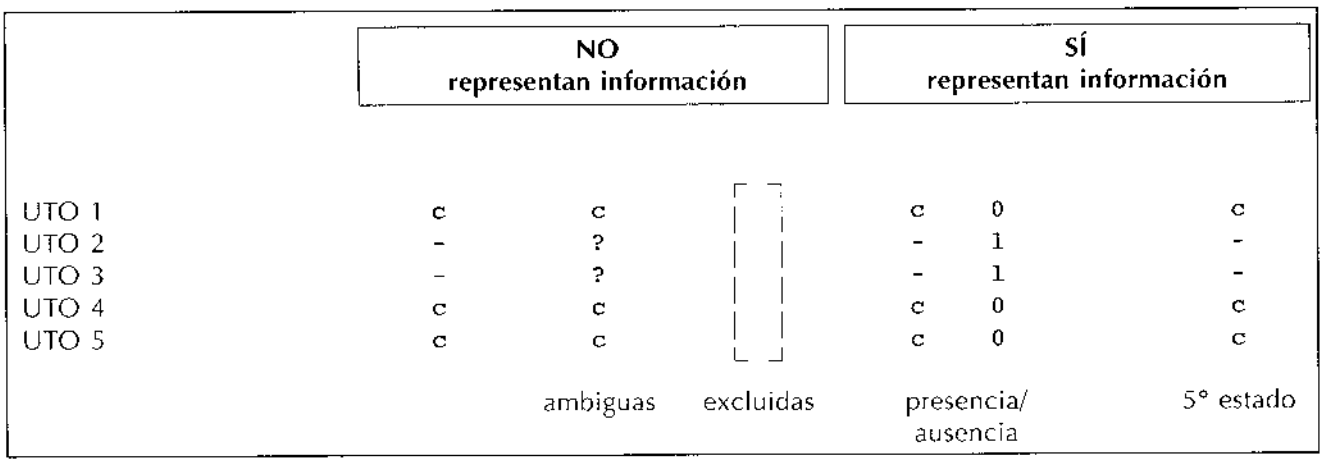

Figura 2. Enfoques alternativos para la codificación de las indels simples. Cuando se consideran como un artificio del alineamiento y cuando se consideran como una hipótesis de homología. En el primer caso se codifican como ambiguas (?) o se excluyen de la matriz. En el segundo se codifica como presencia/ausencia o como 5o. estado. Si se codifica como presencia/ausencia se adiciona una columna para representar la presencia o ausencia de la indel con la opción gapcode=missing en PAUP. Si se codifica como 5° estado sólo se selecciona gapcode=newstate en PAUP.

miento con el menor costo, es decir, donde se introducen el menor número posible de indels. Mientras que si se les da penalizaciones bajas (por ejemplo, 2 a 1) se favorece la inserción de indels. Los algoritmos calculan el costo de todos los desplazamientos posibles y escogen el alineamiento con el menor costo.

Enfoques filogenéticos para la codificación de las indels

Se revisaron los trabajos en cuatro revistas (Cladistics, Molecular Biology and Evolution, Systematic Biology, Systematic Botany) publicadas de enero de 1993

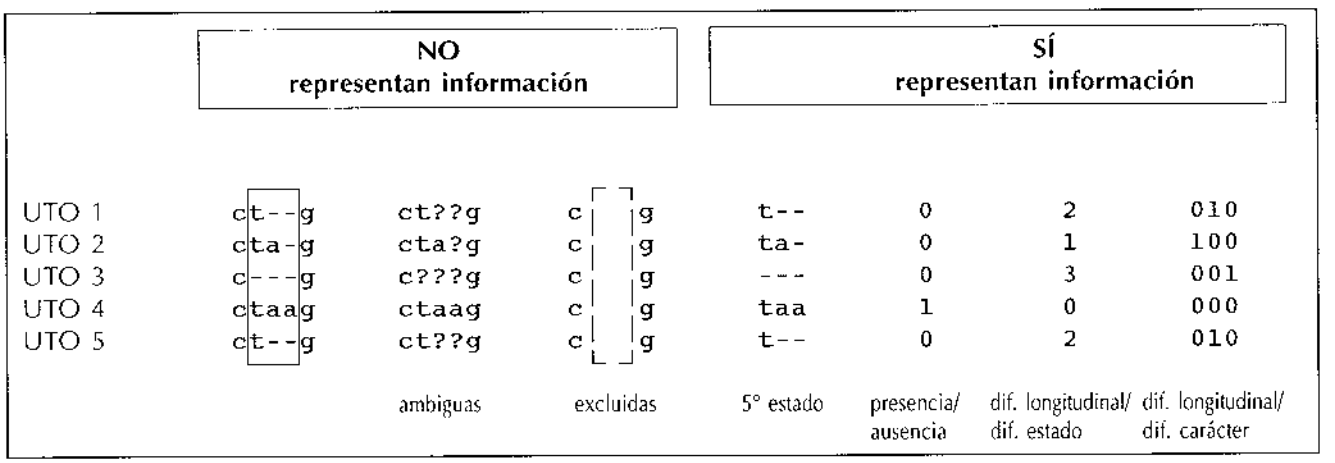

Figura 3. Enfoques alternativos para la codificación de las indels múltiples. Cuando se consideran como un artificio del alineamiento y cuando se consideran como una hipótesis de homología. En el primer caso se codifican como ambiguas o se excluyen de la matriz. En el segundo caso se codifica como 5o. estado, presencia/ausencia, diferente longitud/diferente estado y como diferente longitud/diferente carácter. Estos dos últimos códigos se han empleado cuando las indels son de varias posiciones contiguas pero de diferente longitud entre los UTO's. Cuando se codifican como diferente longitud/diferente estado se adiciona una nueva columna a la matriz de datos y cada una de las indels de distinta longitud se codifica con diferente estado. Si se emplea el código diferente longitud/diferente carácter, se adicionan varias columnas a la matriz de datos y cada indel de distinta longitud se codifica como la presencia/ausencia de un nuevo carácter. En ambos casos los análisis se efectúan con la opción gapcode=missing en PAUP.

a mayo de 1996. Cada estudio filogenético con secuencias se examinó para identificar el tipo de codificación de las indels. La figura 2 muestra las distintas formas de codificación de una indel simple mientras que la figura 3 muestra las de una indel múltiple de diferente longitud. El examen de la literatura reveló que se han usado dos distintos enfoques para el análisis de las indels. El primero las elimina como un artificio del alineamiento y por consiguiente sin ningún valor filogenético (por ejemplo, Ota y Nei, 1994; Pawlowski *et al.*, 1994). El segundo enfoque las considera implícitamente como hipótesis de homología (por ejemplo, Lloyd y Calder, 1991, Baum *et al.*, 1994; Vogler y DeSalle, 1994; Clark *et al.*, 1996).

Bajo el primer enfoque, dos alternativas comunmente empleadas implican que las indels son un artificio del alineamiento y que no representan información filogenética. La primera es codificarlas como ambiguas ("?") en la matriz de datos y la segunda es excluirlas de los análisis. Cuando se codifican como "?", se debe seleccionar la opción "gapcode=missing" en PAUP. En este caso los análisis optimizan cualquiera de las otras bases que aparecen en la misma columna o posición (por ejemplo, Winnepenninckx *et al.*, 1995; Murphy y Collier, 1996). En la mayoría de los trabajos, el tratamiento de las indels no es claro y frecuentemente se omiten inadvertidamente de los análisis filogenéticos al codificarlas como "?". En contraste, manualmente se pueden excluir las columnas que contienen indels aún si la mayoría de los taxa presentan un nucleótido y los análisis obviamente no toman en cuenta estas bases (Pawlowski *et al.*, 1994).

Bajo el segundo enfoque, los estudios que consideran a las indels como hipótesis de homología siguen cuatro alternativas de codificación (figura 3). La primera es codificar las indels en cada columna en la matriz de datos como un nuevo estado. Potencialmente una indel es un quinto estado y generalmente se representa por un guión (-). En este caso la búsqueda de cladogramas se efectúa con la opción "gapcode=newstate" en PAUP (por ejemplo, Stewart y Baker, 1994; Lafay *et al.*, 1995). La segunda alternativa es codificarlas como un carácter adicional. Esto se logra cuando por cada columna (primaria) donde existe una indel ("?") se agrega una nueva columna (secundaria) en la cual se registra la presencia o ausencia de las indels mediante un código binario (0,1). Para activar apropiadamente este código en PAUP los análisis se efectúan con la opción "gapcode=missing", el cual no elimina los nucleótidos de la columna primaria (por ejemplo, Steele y Vilgalys, 1994; Johnson y Soltis, 1994). Estas dos alternativas de codificación se han aplicado tanto a las indels simples como a las múltiples pero de la misma longitud. Algunos autores han argumentado que no es apropiado codificarlas como un quinto estado sobre todo las indels múltiples. Esto se debe a que cada una de las posiciones contiguas se interpreta como un evento independiente lo que ocasiona que las indels pesen excesivamente en los análisis. También se ha sugerido que es inapropiado codificarlas como presencia/ausencia puesto que hay indels múltiples que se sobrelapan (Eernisse y Kluge, 1993).

La tercera y cuarta alternativas se han utilizado en el caso de las indels múltiples y de diferente longitud entre los taxa. Algunos autores han optado por adicionar una nueva columna a la matriz de datos en la cual cada una de las indels de distinta longitud se codifica con diferente estado (Baum *et al.*, 1994). Otros han seguido la alternativa de adicionar varias columnas a la matriz de datos. Cada indel de distinta longitud se interpreta como un carácter y se codifica la presencia/ausencia como los estados (Vogler y DeSalle, 1994). Las columnas que contienen las indels se codifican con "?" y las búsquedas en PAUP se efectúan con la opción "gapcode=missing" ya sea que las indels de diferente longitud se codifiquen como un carácter o como diferentes caracteres.

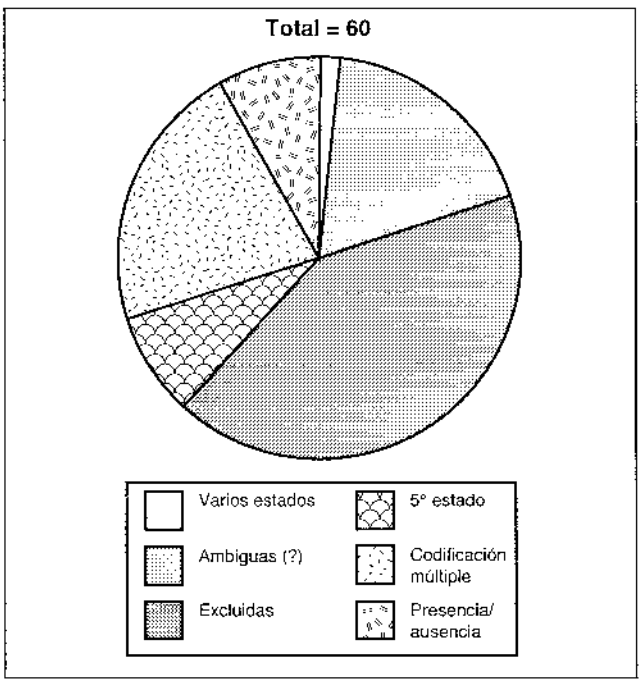

Figura 4. Proporción de los códigos usados para las indels en los análisis filogenéticos de secuencias génicas en 60 estudios taxonómicos. Las referencias se obtuvieron de cuatro revistas (*Cladistics, Molecular Biology* and *Evolution, Systematic Biology, Systematic Botany*) publicadas de enero de 1993 a mayo de 1996. La codificación múltiple incluye la asignación de más de un código para las indels en los análisis.

Evaluación del efecto de las indels en los análisis filogenéticos de secuencias génicas

La misma revisión bibliográfica también expuso que muy pocos trabajos exploran la asignación de más de un código para las indels (figura 4). En la mayoría, se codifican únicamente como ambiguas y otros las excluyen completamente de los análisis. Algunas veces no es explícito si las indels se codificaron como ambiguas o se excluyeron intencionalmente. No obstante, se identificaron trece estudios en donde se asignó claramente más de un código para las indels. Dos de estos estudios mezclaron en un solo análisis dos códigos para las indels localizadas en dos regiones distintas de la misma secuencia (Milinkovitch *et al.*, 1994; Csink y McDonald, 1995). Los once estudios restantes aplicaron los códigos en análisis por separado y se comparó el efecto en los cladogramas resultantes. En estos estudios, la combinación de códigos es diversa (Tabla 1). Por ejemplo, cinco estudios las codificaron como "ambiguas" y como "presencia/ausencia". Aunque por limitaciones del programa DNAml, dos de estos estudios codificaron las indels simples y múltiples con una base en lugar de "0,1" (Fitch *et al.*, 1995; Buckler y Holtsford, 1996).

Los resultados de la codificación múltiple también son variados (figura 5). Por un lado, cinco estudios calificaron el uso de los distintos códigos como irrelevante en relación a la topología (Beintema *et al.*, 1994; Ford *et al.*, 1995; Fitch *et al.*, 1995; Janczewski *et al.*, 1995; Crandall y Fitzpatrick, 1996). Por otro, se registró incremento en la homoplasia (Friedlander *et al.*, 1996) y también se han calificado como informativas (Eernisse y Kluge, 1993; Gielly y Taberlet, 1994; Johnson y Soltis, 1994; Vogler y DeSalle, 1994; Buckler y Holtsford, 1996). El reducido número de estudios que han examinado los códigos alternativos no permite por ahora anticipar cuáles son las tendencias en los cambios de los patrones de homoplasia o la topología de los árboles. También se desconoce qué otros efectos pueden resultar de los diferentes códigos para las indels. Con el propósito de iniciar el examen de este problema se hizo un ejercicio preliminar con los seis códigos conocidos en la literatura para el análisis de las indels. La matriz de datos disponible para este examen consiste en las secuencias del gen del ARN ribosomal para el complejo *Rhizoctonia solani* (González, 1992). La evaluación de los códigos se hizo sobre los efectos que tuvieron en los niveles de homoplasia, en los cambios topológicos, en la resolución y en el número de árboles.

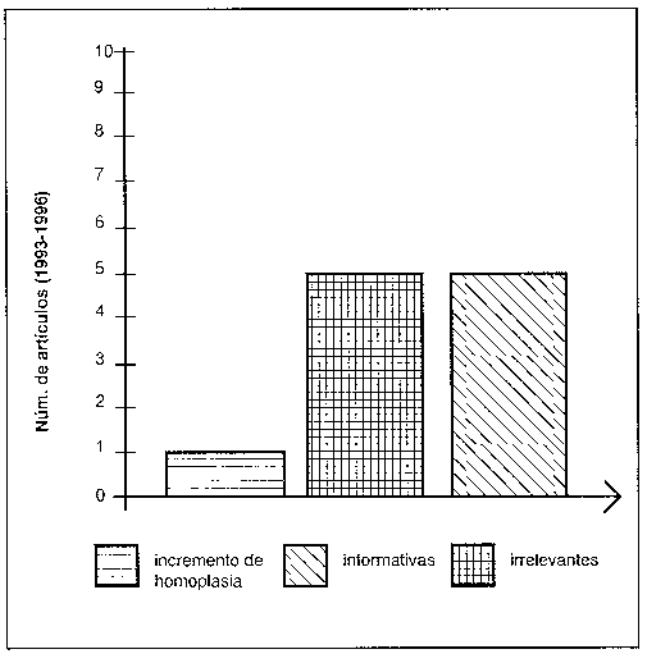

Figura 5. Efecto de la codificación múltiple de las indels en los estudios filogenéticos registrados en la literatura.

Análisis de las secuencias y codificación de indels. La región secuenciada del hongo fitopatógeno *R. solani* corresponde a la terminación 5' de la subunidad larga y una porción del espaciador interno 2 (ITS 2) del ADN ribosomal (González, 1992). Se secuenció un representante de 13 grupos intraespecíficos de diez grupos anastomósicos de este hongo (Ogoshi, 1987; Sneh *et al.*, 1991). Otros dos grupos intraespecíficos de un grupo anastomósico se representaron con 6 individuos para detectar polimorfismo. En total se secuenciaron 19 UTO's de *R. solani* y un grupo externo (*R. cerealis*). Las secuencias se alinearon manualmente con el programa ESEE (Cabot y Beckenbach, 1989). Se insertaron espacios para justificar diferencias en longitud. Dependiendo de los códigos usados para las indels el alineamiento de la región secuenciada produjo de 49 a 85 caracteres informativos, de los cuales 16 corresponden a indels. En la subunidad larga se localizaron nueve indels simples, mientras que en el ITS 2 se encontraron siete, cinco son simples y dos son múltiples de distinta longitud (figura 6). Las indels se codificaron de seis maneras: 1) ambiguas (?), 2) excluidas, 3) quinto estado (-), 4) presencia/ausencia (0,1), 5) diferente longitud/varios estados (0,1,2) y 6) diferente longitud/varios caracteres (0,1). La figura 7, muestra un fragmento de la matriz de las secuencias en la región del ITS 2 con las dos indels múltiples de distinta longitud tal como resultan del alineamiento. En la misma figura se ilustra la aplicación de cuatro códigos distintos y las columnas

generadas por cada código cuando se consideran como una hipótesis de homología.

Análisis filogenéticos. Las búsquedas de árboles parsimoniosos se realizaron con el programa PAUP 3.1.1 (Swofford, 1993). Los seis análisis correspondientes a cada código se efectuaron heurísticamente con diez réplicas y con las opciones "random taxon addition" y "TBR branch swapping". Se evaluó la estabilidad de los clados de los cladogramas resultantes en cada búsqueda con un análisis de "bootstrap" de 100 réplicas usando las opciones "simple taxon addition" y "TBR branch swapping". Un árbol de cada análisis se muestra en las figuras 8 y 9.

En el análisis número 1, la codificación de las indels como ambiguas generó 52 caracteres informativos y la opción "gapcode=missing" permite la optimización de las otras bases que aparecen en la misma posición. Este análisis resultó en un sólo árbol más parsimonioso de 105 pasos de longitud con un índice de consistencia (CI) de 0.543. La exclusión de las indels en el análisis número 2 elimina además la información de los nucleótidos que se presentan en las posiciones en las que aparece una indel. Como consecuencia sólo se generaron 49 caracteres informativos. Esta búsqueda encontró el mismo cladograma que el del análisis número 1 pero de 98 pasos de longitud y con un CI de 0.551. En el análisis número 3, la codificación de las indels como quinto estado (-) con la opción "gapcode=newstate" permite que se consideren todas las columnas en donde aparece una indel. Los 85 caracteres informativos generados es el mayor numero derivado de los seis códigos. Bajo este enfoque para codificar las indels se encontraron seis cladogramas de 183 pasos de longitud y con CI de 0.607 (figura 8).

En el análisis número 4 la codificación de las indels como un carácter adicional (presencia/ausencia) requiere agregar una columna por cada bloque de indels contiguas. Dado que se activa la opción "gapcode=missing" la(s) columna(s) con las indels sólo aporta(n) la información de los nucleótidos presentes. Esta codificación generó 65 caracteres informativos. El resultado de la búsqueda fueron cuatro cladogramas de 136 pasos (CI de 0.507) y el árbol de consenso es igual que el del análisis 3. La codificación de las indels múltiples como un carácter (diferente longitud como diferente estado) en el análisis número 5 generó 65 caracteres informativos. Se encontraron cuatro cladogramas de 139 pasos (CI de 0.518). El árbol de consenso de este análisis es el de mayor resolución comparado con los árboles de con-

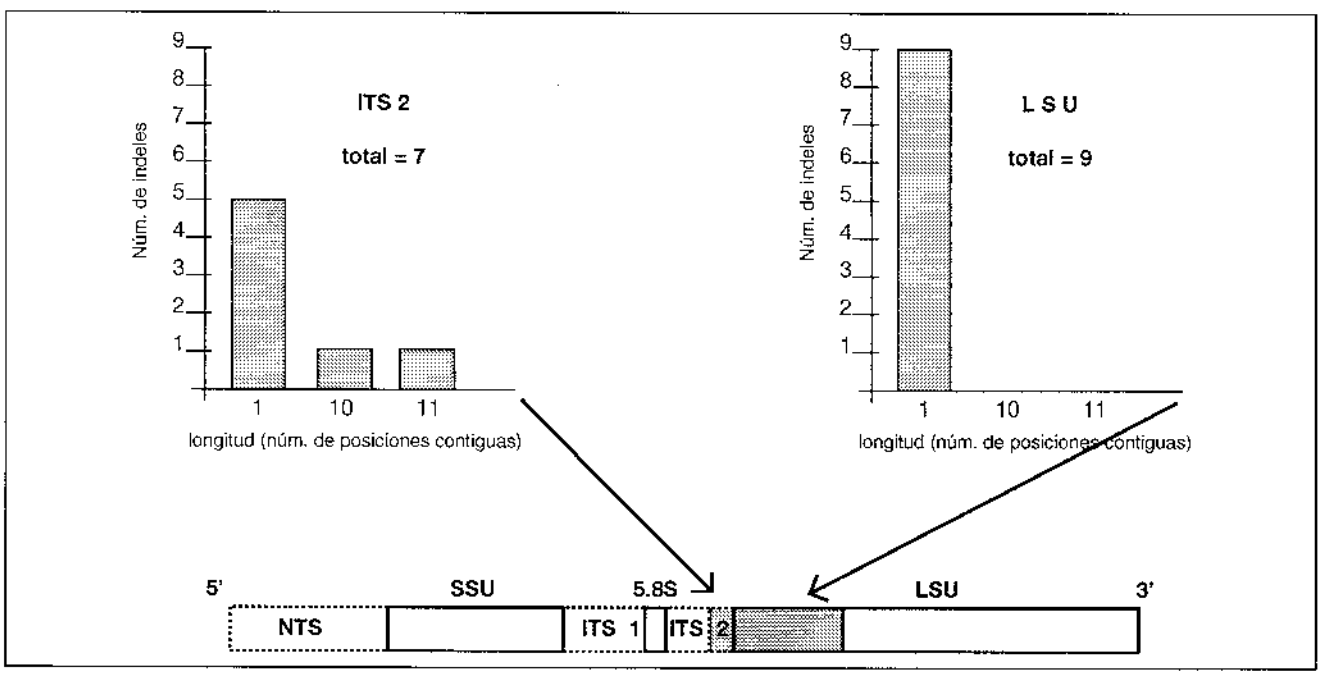

Figura 6. Tipo de las indels encontradas en la región secuenciada del ADN ribosomal de *Rhizoctonia solani* (región en gris). La estructura del gen incluye una región codificadora que incluye la subunidad larga (LSU), la subunidad corta (SSU) y la subunidad 5.8S. Una región no codificadora que es transcrita (ITS 1 y 2) y una región espaciadora que no se transcribe (NTS). Las indels de la subunidad larga son simples mientras que las indels del ITS 2 son simples y múltiples.

senso de las otras búsquedas donde se consideran a las indels como hipótesis de homología. En el análisis número 6 las indels múltiples se codificaron como varios caracteres (presencia/ausencia) dependiendo de su longitud. Este código generó 69 caracteres informativos y resultaron trece cladogramas de 149 pasos (CI de 0.490). El árbol de consenso de este análisis es el de menor resolución (figura 9).

Las figuras 8 y 9 muestran el efecto de los distintos códigos asignados a las indels sobre las relaciones filogenéticas del hongo fitopatógeno *R. solani*. Los seis diferentes códigos de las indels resultaron en obvias diferencias en el índice de consistencia, en el número de árboles y en la topología. Igualmente hay diferencias en los niveles de robustez revelados por el índice del "bootstrap" aunque la mayoría de los clados en los seis árboles son los mismos. Es difícil comparar los árboles en función de los niveles de robustez mediante el "bootstrap", pero mediante los índices de consistencia (CI) se puede tener una idea de las tendencias generales del efecto de los códigos para las indels. El problema es cómo evaluar si las diferencias entre los CI son significativas o son parte de la variación esperada debido a las diferencias en la matriz de datos usada para cada análisis. Empíricamente Sanderson y Donoghue (1989) encontraron en numerosos estudios cladísticos que los índices de consistencia están correlacionados con el número de taxa pero no con el número de caracteres. La evaluación de las diferencias entre los CI de los seis análisis presentes entonces debe tomar en cuenta que, aunque se realizaron con distintos números de caracteres informativos, cada análisis siempre incluyó 20 UTO's.

La revisión de Sanderson y Donoghue (1989) incluyó sesenta estudios y analizando el número de taxa y los CI reportados, ellos derivaron la regresión po-

SÍ
representan información

R. cerealis	T	----------	GAAA	A----------G	T	----------	A----------G	1 1	3 2	001	010
AG1-IA	?	??????????	????	?????????????	?	??????????	?????????????	? ?	? ?	???	???
AG1-IB	C	----------	GAAA	A----------	T	----------	A----------	1 1	3 3	001	001
AG1-IC	T	---------G	GAAA	A----------G	T	---------G	A----------G	1 1	2 2	010	010
AG2-1	T	----------	GAAA	----------G	T	----------	----------G	1 1	3 3	001	001
AG2-2	C	---------C	GAAA	A----------G	T	---------C	A----------G	1 1	2 2	010	010
AG3	T	----------	GAAA	A----------	T	----------	A----------	1 1	3 3	001	001
AG4HGI(7)	C	--------CT	GTAA	A-------AGGG	T	--------CT	A-------AGGG	1 1	1 1	100	100
AG4HGI(18)	C	--------CT	GTAA	A-------AGGG	T	--------CT	A-------AGGG	1 1	1 1	100	100
AG4HGI(30)	C	--------CT	GTAA	A-------AGGG	T	--------CT	A-------AGGG	1 1	1 1	100	100
AG4HGI(6)	C	--------CT	GTAA	A-------AGGG	T	--------CT	A-------AGGG	1 1	1 1	100	100
AG4HGII(44)	C	TTTCTACTCT	GAAA	AGTTTGGGAGGG	T	TTTCTACTCT	AGTTTGGGAGGG	0 0	0 0	000	000
AG4HGII(45)	C	TTTCTACTCT	GAAA	AGTTTGGGAGGG	T	TTTCTACTCT	AGTTTGGGAGGG	0 0	0 0	000	000
AG5	T	----------	GAAC	----------G	G	----------	----------G	1 1	3 3	001	001
AG6-I	T	----------	GAAT	----------G	T	----------	----------G	1 1	3 3	001	001
AG6-V	T	----------	GAAC	----------G	T	----------	----------G	1 1	3 3	001	001
AG7	T	----------	GAAA	A----------G	T	----------	A----------G	1 1	3 2	001	010
AG8	T	----------	GAAA	A----------	T	----------	A----------	1 1	3 3	001	001
AG9	T	----------	GAAA	A----------G	T	----------	A----------G	1 1	3 2	001	010
AGBI	G	ACATTTTTTT	TTCT	AAAAAAAATAAG	G	ACATTTTTTT	AAAAAAAATAAG	0 0	0 0	000	000
						5° estado		presencia/ ausencia	dif. long./ dif.estado	dif. long/dif. carácter	

Figura 7. Códigos de las indels usados en los análisis filogenéticos de la matriz de datos de secuencias de *R. solani*. Cuando se considera que no representan información filogenética, y cuando se considera que sí representa información filogenética.

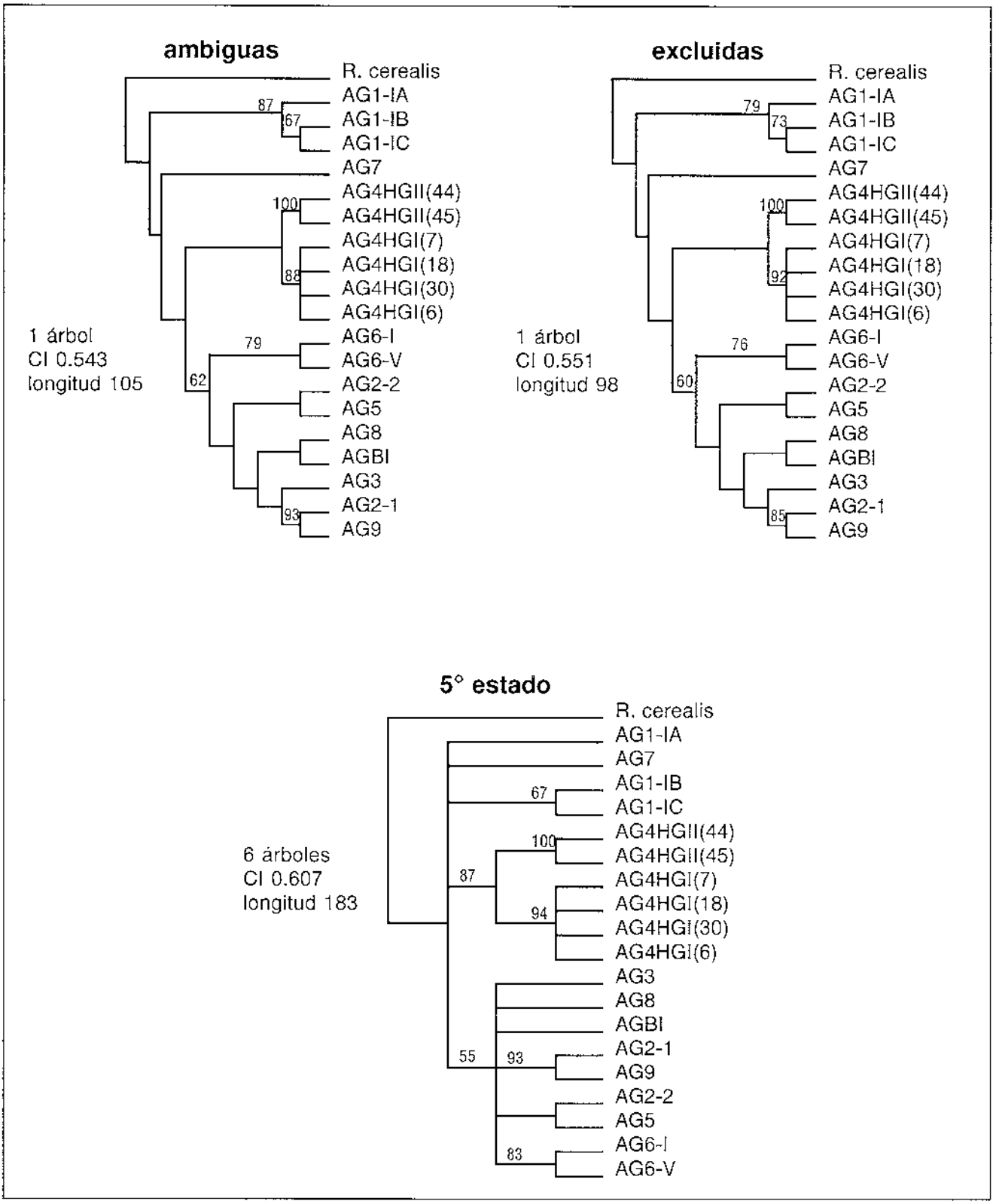

Figura 8. Efecto de codificar las indels como ambiguas, excluidas y 5o. estado en las relaciones filogenéticas de *R. solani*. El efecto se observa en los cambios en homoplasia, en los cambios topológicos, en los cambios en la resolución de las ramas y en los cambios en el número de árboles. Los números en las ramas del árbol representan el resultado de los análisis de "bootstrap". Sólo se incluyen los grupos que se mantienen en más del 50% de las réplicas.

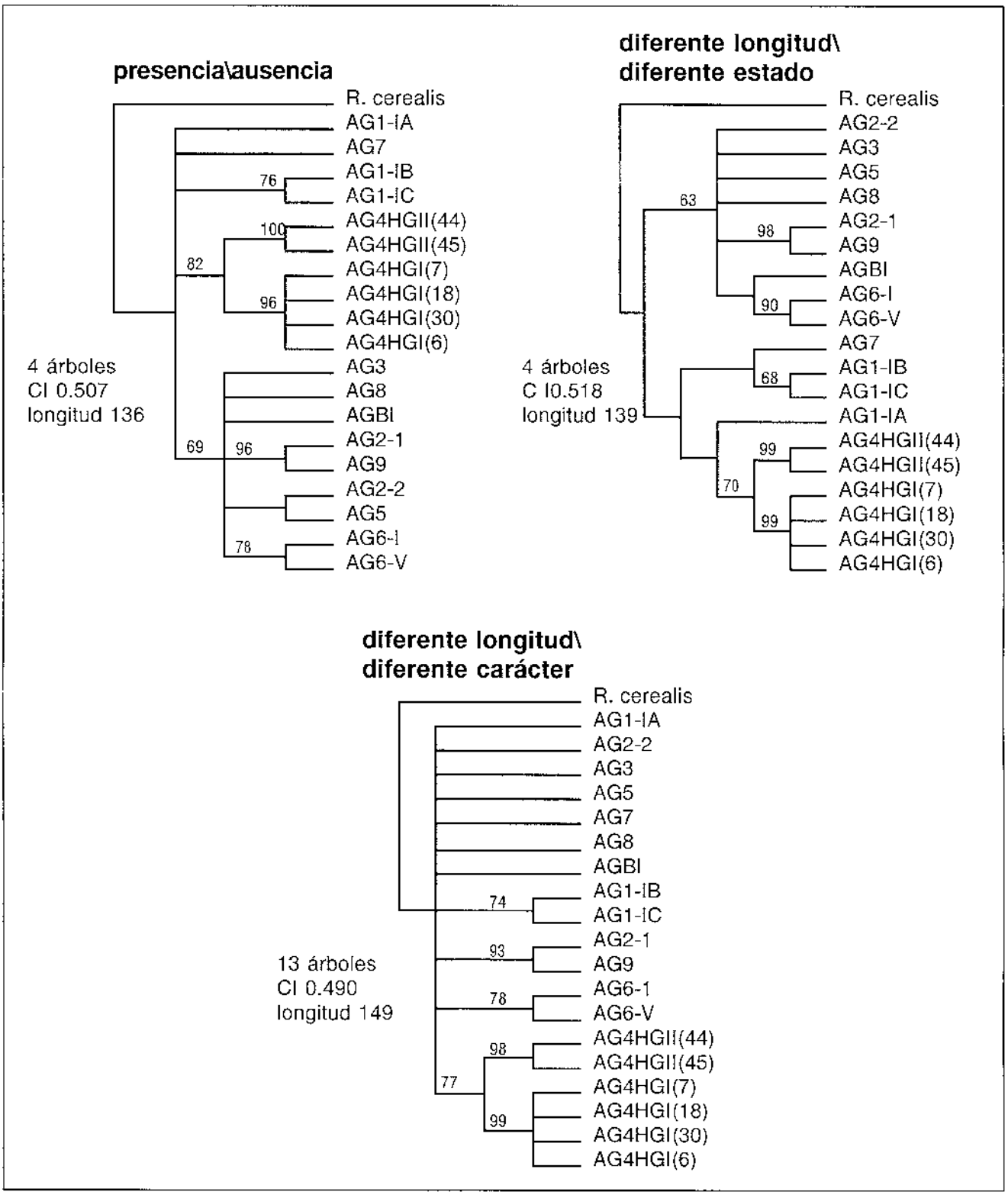

Figura 9. Efecto de codificar las indels como presencia/ausencia, diferente longitud/diferente estado y diferente longitud/ diferente carácter en las relaciones filogenéticas de *R. solani*. El efecto se observa en los cambios en homoplasia, en los cambios topológicos, en los cambios en la resolución de las ramas y en los cambios en el número de árboles. Los números en las ramas del árbol representan el resultado de los análisis de "bootstrap". Sólo se incluyen los grupos que se mantienen en más del 50% de las réplicas.

linomial CI = 0.90 - 0.022 (número de taxa) + 0.000213 (número de taxa)^2 para estimar el valor del índice de consistencia para un número de taxa dado. En el caso de 20 taxa, el valor del CI esperado es 0.545. Los valores de los cinco estudios revisados por Sanderson y Donoghue (1989) van de 0.35 a 0.68, mientras que los valores encontrados aquí en los seis análisis con *R. solani* oscilan entre 0.490 y 0.607. La desviación estándar a partir de estos once valores es de 0.09. Aunque no es una prueba formal, el hecho de que los valores del CI derivados de los seis códigos están comprendidos en la variación de una desviación estandar sugiere que no existen diferencias significativas. Sólo puede resaltarse la tendencia de que la codificación de las indels como diferentes caracteres para diferentes longitudes causa un aumento en el número de árboles, disminuye el índice de consistencia y se pierde la resolución presente en árboles derivados de los otros cinco códigos. En contraste, la mayor resolución, así como el menor número de árboles se obtienen cuando las indels se codifican como ambiguas o si se excluyen.

Conclusiones

Los análisis filogenéticos con secuencias génicas parecen simples a primera vista debido a que los caracteres y los estados ya están definidos en este tipo de datos. Sin embargo, esta simplicidad es engañosa, puesto que se requiere establecer homología táxica o transformacional a nivel de cada posición. La elaboración de hipótesis de homología a nivel de las secuencias génicas es un proceso complejo que incorpora las bases empíricas de similitud, variación discreta, conjunción, heredabilidad e independencia (tal como con cualquier otro sistema de caracteres). También, incluye una fase inferencial en la que las homologías potenciales se evalúan si son congruentes en un cladograma más parsimonioso. Además, la presente revisión ha resaltado las dificultades metodológicas para la codificación apropiada de las indels. Todo esto hace que el análisis de las secuencias génicas requiera no solo una base empírica amplia sino también un marco teórico claro.

La base teórica para el análisis cladístico de las

	ambiguas	excluidas	5° estado	presencia/ ausencia	dif. longitudinal/ dif. estado	dif. longitudinal/ dif. carácter
Ford *et al.*, 1995		X			X	
Crandall y Fitzpatrick, 1996	X	X		X		
Johnson y Soltis, 1994	X			X		
Eernisse y Kluge, 1993	X		X			
Friedlander *et al.*, 1996		X	X			
Buckler IV y Holtsford, 1996	X			X		
Janczwski *et al.*, 1995	X			X		
Fitch *et al.*, 1995		X		X		
Gielly y Taberlet, 1994	X			X		
Vogler y DeSalle, 1994	X		X			X
Beintema *et al.*, 1994	X			X		

Tabla 1. Combinación de códigos usados para las indels en los estudios donde se evaluó su significado filogenético. La combinación de códigos más usada fue ambiguas y presencia/ausencia. Sólo dos estudios evaluaron tres códigos simultáneamente.

secuencias génicas es el concepto de homología filogenética. Bajo este enfoque conceptual, si los nucleótidos en una posición no varían, estos son evidencia de homología táxica y si las bases son diferentes entonces se infiere una hipótesis de homología transformacional. Esto resalta que un carácter taxonómico es un sistema de por lo menos dos homólogos transformacionales o estados. Este principio conceptual se aplica tanto a las substituciones como a las indels. Tal como en el caso de otros caracteres, la congruencia entre hipótesis resolverá si las bases o las indels en una posición pueden mantenerse como evidencia de homología filogenética. Entonces la decisión de si una indel es informativa o no depende de incluirla codificada apropiadamente en los análisis de parsimonia.

La presencia de indels simples y múltiples es inevitable como resultado del alineamiento de las secuencias. La asignación de un código apropiado para las indels es un problema más serio de lo considerado por autores previos y hasta ahora éste ha sido un problema poco explorado. La revisión de la literatura reveló que se han usado seis códigos, dos de los cuales las eliminan tácitamente de los análisis y cuatro permiten su evaluación como hipótesis de homología. Comúnmente sólo las substituciones (transiciones y transversiones) se consideran sitios informativos, pero en este trabajo se ha argumentado que las indels también representan hipótesis de homología potencial. Congruente con esta posición, se deberían codificar las indels para identificar tales hipótesis (táxica o transformacional) y usarlas junto con las substituciones para reconstruir patrones de ancestría común. En principio, cualquiera de los cuatro códigos que consideran a las indels como informativas debería preferirse. Pero por ahora, la mejor manera de representar la variación de las indels parecería ser codificarlas como un carácter adicional con varios estados dependiendo de su longitud (figura 3).

Muy pocos estudios han examinado los efectos del uso de más de un código para las indels. Los seis códigos conocidos se aplicaron para evaluar su efecto en las relaciones filogenéticas entre poblaciones del hongo fitopatógeno *R. solani*. La inclusión de las indels en esta matriz de secuencias no afecta significativamente el nivel de homoplasia en seis análisis diferentes. El efecto más radical es sobre la resolución de los clados y el número de árboles. La codificación que más afectó fue la de diferentes caracteres para diferentes longitudes de las indels múltiples. En contraste, la mayor resolución así como el menor número de árboles se obtienen cuando las indels se codifican como ambiguas o se excluyen de la matriz. Estos resultados obviamente son preliminares e incompletos ya que derivan de una sola matriz de datos. El reducido número de estudios que han usado códigos alternativos no permite conocer cuáles son las tendencias en los cambios de los patrones de homoplasia o la topología de los árboles. También se desconoce qué otros efectos pueden resultar de los diferentes códigos para las indels. Por esto, se están reanalizando varias matrices de secuencias publicadas para evaluar el efecto de recodificar las indels, ya que éstas no se han incluido en la mayoría de los análisis filogenéticos.

Agradecimientos

Agradezco a Efraín De Luna y dos revisores anónimmos sus sugerencias y críticas al presente manuscrito. Este trabajo se realizó gracias al apoyo económico del Instituto de Ecología, A. C., mediante la cuenta 902-14 y al proyecto PACIME-CONACyT No. 1837P-N9507.

Literatura Citada

Baum D.A., Sytsma K.J. y Hoch P.C. 1994. A phylogenetic analysis of *Epilobium* (Onagraceae) based on nuclear ribosomal DNA sequences. *Systematic Botany* **19:** 363-388.

Baverstock P.R. y Moritz C. 1990. Sampling design. pp. 13-24. En: Hillis D.M. y Moritz C., edrs. *Molecular Systematics*. Sinauer, Sunderland, MA.

Beintema J.J., Stam W.T., Hazes B. y Smidt M.P. 1994. Evolution of arthropod hemocyanins and insect storage proteins (Hexamerins). *Molecular Biology and Evolution* **11:** 493-503.

Buckler IV E.S. y Holtsford T.P. 1996. *Zea* Systematics: Ribosomal ITS evidence. *Molecular Biology and Evolution* **13:** 612-622.

Cabot E.L. y Beckenbach A.T. 1989. Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. *Computer Applications in the Biosciences* **5:** 233-234.

Clark A.G., Leicht B.G. y Muse S.V. 1996. Length variation and secondary structure of introns in the Mlc 1 gene in six species of *Drosophila*. *Molecular Biology and Evolution* **13:** 471-482.

Crandall K.A. y F. Fitzpatrick Jr. 1996. Crayfish molecular systematics: Using a combination of of procedures to estimate phylogeny. *Systematic Biology* **45:** 1-26.

Csink A.K. y Mc Donald J.F. 1995. Analysis of *copia* sequence variation within and between *Drosophila* species. *Molecular Biology and Evolution* **12:** 83-93.

De Luna E. y Mishler B.D. 1997. El concepto de homología filogenética y la selección de caracteres taxonómicos. *Boletín de la Sociedad Botánica de México.*

De Pinna M.C.C. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7:** 367-394.

Eernisse D.J. y Kluge A.G. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Molecular Biology and Evolution* **10:** 1170-1195.

Felsenstein J. 1984. The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. Pp. 169-191 en: Duncan T. y Stuessy T.F., edrs. *Cladistics: Perspectives on the reconstruction of evolutionary history*. Columbia University Press, New York.

Felsenstein J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22:** 521-565.

Fitch D.H.A., Bugaj-Gaweda B. y Emmons S.W. 1995. 18S ribosomal RNA gene phylogeny for some Rhabditidae related to Caenorhabditis. *Molecular Biology and Evolution* **12:** 346-358.

Fitch W.M. y Smith T.F. 1983. Optimal sequence alignments. *Proceedings of the National Academy of Science USA*. **80:** 1382-1386.

Ford V.S., Thomas B.R. y Gottlieb L.D. 1995. The same duplication accounts for the PgiC genes in *Clarkia xantiana* and *C. lewisii* (Onagraceae). *Systematic Botany* **20:** 147-160.

Friedlander T.P., Regier J.C. y Mitter C. 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. *Systematic Biology* **41:** 483-490.

Friedlander T.P., Regier J.C., Mitter C. y Wagner D.L. 1996. A nuclear gene for higher level phylogenetics: phosphoenolpyruvate carboxykinase tracks mesozoic-age divergences within Lepidoptera (Insecta). *Molecular Biology and Evolution* **13:** 594-604.

Gielly L. y Taberlet P. 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. *Molecular Biology and Evolution* **11:** 769-777.

González D. 1992. Classification of the plant pathogenic fungus *Rhizoctonia solani* (Basidiomycotina: Tullasnelales) using ribosomal DNA sequence data. MS thesis. Duke University, Durham, North Carolina.

González D. El uso de secuencias génicas para estudios taxonómicos. *Boletín de la Sociedad Botánica de México*. En revisión.

Graybeal A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. *Systematic Biology* **43:** 174-193.

Hall B.K., ed. 1994. *Homology. The hierarchical basis of comparative biology*. Academic Press, San Diego.

Hein J. 1989. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences when the phylogeny is given. *Molecular Biology and Evolution* **6**: 649-668.

Hillis D.M. 1994. Homology in Molecular Biology. Pp. 339-368 en: Hall B.K., edrs. *Homology: the hierarchical basis of comparative biology*. Academic Press, San Diego, Ca.

Janczewski D.N., Modi W.S., Stephens J.C. y O'Brien S.J. 1995. Molecular evolution of mitochondrial 12S RNA and cytochrome b sequences in the pantherine lineage of Felidae. *Molecular Biology and Evolution* **12:** 690-707.

Johnson L.A. y Soltis D.E. 1994. matK DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. *Systematic Botany* **19:** 143-156.

Knight A. y Mindell D.P. 1993. Substitution bias, weighting of DNA, sequence evolution, and the phylogenetic position of fea's viper. *Systematic Biology* **42:** 18-31.

Knight A. y Mindell D.P. 1995. Weighting of nucleotide sequences: A reply. *Systematic Biology* **44:** 112-116.

Lafay B., Smith A.B. y Christen R. 1995. A combined morphological and molecular approach to the phylogeny of asteroids (Asteroidea: Echinodermata). *Systematic Biology* **44:** 190-208.

Lake J.A. 1987. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Molecular Biology and Evolution* **4:** 167-191.

Lake J.A. 1991. The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution* **8:** 378-385.

Lloyd D.G. y Calder V.L. 1991. Multiresidue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. *Journal of Evolutionary Biology* **4:** 9-21.

Milinkovitch M.C., Meyer A. y Powell J.R. 1994. Phylogeny of all major groups of cetaceans based on DNA sequences from three mitochondrial genes. *Molecular Biology and Evolution* **11:** 939-948.

Mindell D.P. 1991. Aligning DNA sequences: Homology and phylogenetic weighting. pp. 73-89 En: Miyamoto M.M. y Cracraft J., edrs. *Phylogenetic analysis of DNA sequences*. Oxford University Press, New York.

Miyamoto M.M. y Cracraft J. 1991. *Phylogenetic analysis of DNA sequences*. Oxford University Press, New York.

Moritz C. y Hillis D.M. 1990. Molecular systematics: Context and controversies. pp. 1-10. En: Hillis DM y Moritz C., Edrs. *Molecular Systematics*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Murphy W.J. y Collier G.E. 1996. Phylogenetic relationships within the aplocheiloid fish genus Rivulus (Cyprinodontiformes, Rivulidae): Implications for Caribbean and Central American biogeography). *Molecular Biology and Evolution* **13:** 642-649.

Needleman S.B. y Wunsch C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48:** 443-453.

Nei M. 1991. Relative efficiencies of different tree-making methods for molecular data. pp. 90-128. En: Miyamoto M.M. y Cracraft J., Edrs. *Phylogenetic analysis of DNA sequences*. Oxford University Press, New York.

Nelson G. 1994. Homology and Systematics. pp. 101-149. En: Hall B.K., ed. *Homology The hierarchical basis of comparative biology*. Academic Press, San Diego.

Ogoshi A. 1987. Ecology and pathogenicity of anastomosis and intraspecific groups of *Rhizoctonia solani* Kühn. *Annual Review of Phytopathology* **25:** 125-143.

Ota T. y Nei M. 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH

gene family. *Molecular Biology and Evolution* **11:** 469-482.

Patterson C. 1982. Morphological characters and homology. Pp. 21-74 En: Joysey K.A. y Friday A.E., Edrs. *Problems of Phylogenetic Reconstruction.* Academic Press, New York.

Patterson C. 1988. Homology in classical and molecular biology. *Molecular Biology and Evolution* **5:** 603-625.

Pawlowski J, Bolivar I., Guiard-Maffia J. y Gouy M. 1994. Phylogenetic position of foraminifera inferred from LSU rRNA gene sequences. *Molecular Biology and Evolution* **11:** 929-938.

Roth V.L. 1994. Within and between organisms: replicators, lineages, and homologues. Pp. 301-337 En: Hall B.K., Ed. *Homology The hierarchical basis of comparative biology.* Academic Press, San Diego.

Sanderson M.J. y Donoghue M.J.. 1989. Patterns of variation in levels of homoplasy. **43:** 1781-1795.

Sneh B., Burpee L. y Ogoshi A. 1991. *Identification of Rhizoctonia species.* APS press, St. Paul, Minnesota.

Sober E. 1988. *Reconstructing the past. Parsimony, Evolution and Inference.* MIT Press, Cambridge.

Steele K.P. y Vilgalys R. 1994. Phylogenetic analyses of Polemoniaceae using nucleotide sequences of the plastid gene matK. *Systematic Botany* **19:** 126-142.

Stevens P.F. 1991. Character states, morphological variation, and phylogenetic analysis: A review. *Systematic Botany* **16:** 553-583.

Stewart D.T. y Baker A.J. 1994. Patterns of sequence variation in the mitochondrial D-loop region of shrews. *Molecular Biology and Evolution* **11:** 9-21.

Swofford D.L. 1993. *PAUP: phylogenetic analysis using parsimony, version 3.1.1.* Illinois Natural History Survey, Champaign.

Swofford D.L. y Olsen G.J. 1990. Phylogeny reconstruction. Pp. 411-501. En: Hillis D.M. y Moritz C., Edrs. *Molecular Systematics.* Sinauer, Sunderland, Ma.

Vogler A.P y DeSalle R. 1994. Evolution and phylogenetic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis. Molecular Biology and Evolution* **11:** 393-405.

Wagner G.P. 1989. The biological homology concept. **20:** 51-69.

Waterman M.S., Joyce J. y Eggert M. 1991. Computer alignment of sequences. Pp. 59-72. En: Miyamoto M.M. y Cracraft J., edrs. *Phylogenetic analysis of DNA sequences.* Oxford University Press, New York.

Wheeler W.C. 1990. Combinatorial weights in phylogenetic analysis: A statistical parsimony procedure. *Cladistics* **6:** 269-275.

Wiley E.O. 1981. *Phylogenetics. The theory and practice of phylogenetic systematics.* John Wiley & Sons, New York.

Williams D.M. 1993. DNA analysis: methods. Pp. 102-123. En: Forey P.L., Humpries C.J., Kitching I.L. y Scotland R.W., edrs. *Cladistics. A Practical Course in Systematics.* Oxford University Press, New York.

Williams P.L. y Fitch W.M. 1990. Phylogeny determination using dynamically weighted parsimony method. *Methods in Enzymology* **183:** 615-626.

Winnepenninckx B., Backeljau T. y DeWachter R. 1995. Phylogeny of protostome worms derived from 18S rRNA sequences. *Molecular Biology and Evolution* **12:** 641-649.