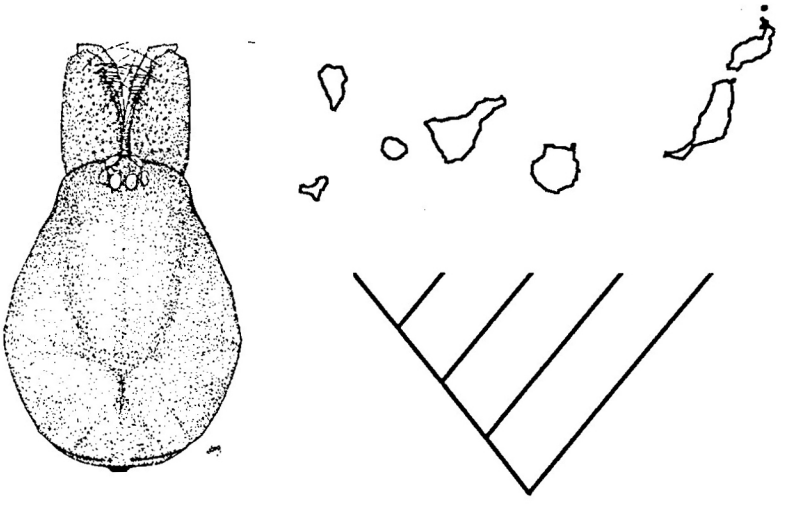
Departament de Biologia Animal
Facultat de Biologia
Universitat de Barcelona

Tesi Doctoral

COLONITZACIÓ I RADIACIÓ DEL GÈNERE *Dysdera* (ARACHNIDA, ARANEAE) A LES ILLES CANÀRIES



Miquel Àngel Arnedo Lombarte

1998

3.3. RECONSTRUCCIÓ FILOGENÈTICA

3.3.1. Conceptes bàsics

Existeix un ordre a la natura que es manifesta d'una manera jeràrquica. S'assumeix que aquesta ordenació jeràrquica de la diversitat biològica és el resultat de l'evolució, és a dir, de la descendència amb modificació (Eldredge i Cracraft 1980). La reconstrucció o inferència filogenètica és el procés pel qual mitjançant l'aplicació d'un conjunt de tècniques i metodologies, obtenim una hipòtesi de les relacions genealògiques o evolutives entre els taxons objectes d'estudi. Aquesta hipòtesi de relació es representa en forma d'un diagrama de branques anomenat **arbre filogenètic** o **filogènia**. L'àrbre filogenètic es composa de dos elements principals: els **nodes** i els **internodes** o **branques** (**fig. 8**). Els nodes són de dos tipus: el **nodes terminals**, als quals només arriba una sola branca i que representen als taxons estudiats, i els **nodes interns**, als quals conflueix més d'una branca i que representen els avantpassats hipotètics dels nodes que deriven d'ells. La multiplicació de branques en els nodes interns s'interpreta com el fenomen d'especiació, o formació de dos taxons a partir d'un taxó ancestral. Les branques, les línies que connecten dos nodes, representen els canvis produïts entre dos nodes o, el que és el mateix, l'evolució produïda entre els dos taxons connectats. Les branques poden ser **terminals**, si connecten un node intern amb un node terminal,

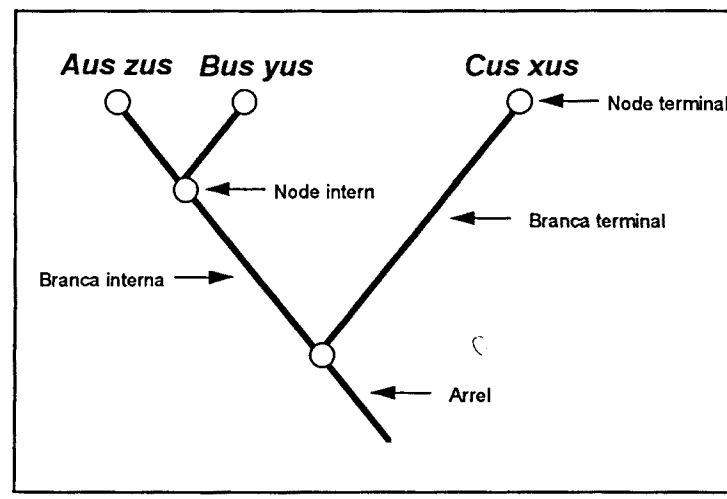


Figura 8.- Arbre filogenètic a on s'indiquen els diferents elements que el componen.

internes, si connecten dos internodes, o **arrels**, si només s'uneixen a un node. Les arrels representen l'origen o el punt de partida del grup de taxons d'interés i la seva funció és la de donar direccionalitat a l'àrbre i poder així parlar de nodes anteriors o avantpassats i nodes posteriors o descendents. Si l'arbre no té arrel s'anomena **xarxa** o, simplement, arbre sense arrel o desarrelat. D'altra banda, el **grup intern** o ***ingroup***, és el conjunt de taxons estudiats per l'investigador, el **grup extern** o ***outgroup*** és qualsevol del taxons utilitzats en una anàlisi no inclòs en els taxons d'estudi. Un cas particular d'*outgroup* és el **grup germà** o ***sister group***, que és el grup genealògicament més proper al grup considerat o, dit d'una altra manera, el grup que comparteix un avantpassat comú més recent amb l'*ingrup.*

La metodologia anomenada **cladística** o **cladisme** ha esdevingut el paradigma en la inferència filogenètica (Kluge i Wolf 1993). El cladisme va ser originalment introduït a la biologia per l'entomòleg alemany Willi Hennig, que l'any 1950 publicà el llibre *Grundzüge einer Theorie der phylogenetischen Systematik*, el qual ha estat cabdal a la història de la biologia comparativa i la ciència evolutiva (Janvier 1984). Tanmateix, aquesta obra restà pràcticament ignorada fins a la seva traducció a l'anglès, apareguda l'any 1966 sota el títol *Phylogenetic systematics.* Les idees de Hennig van ser formulades en el context d'una petita revolució dins la taxonomia que tenia com a objectiu dotar-la d'una major objectivitat. La taxonomia tradicional havia estat acusada de no ser una autèntica ciència degut precisament a la gran subjectivitat dels criteris que utilitzava (Quicke 1993). Paral.lelament a l'aparició de la sistemàtica filogenètica, va ser proposada una altra metodologia amb uns objectius similars però amb una filosofia radicalment diferent: la **fenètica** o **taxonomia numèrica** (Sneath i Sokal 1973). Finalment, es pot reconèixer una tercera escola taxonòmica, anomenada **taxonomia evolutiva** o **eclèctica** (Simpson 1961, Mayr i Ashlock 1991), resultat de l'adopció dels principis de la teoria sintètica de l'evolució a la taxonomia. L'aportació de Hennig a aquesta 'revolució taxonòmica' va ser el de postular el principi de què les classificacions havien de reflectir exactament les relacions filogenètiques entre els taxons a classificar i més concretament l'ordre dels brancatges en elles observats. Alguns taxònoms evolutius van adduir que el brancatge era només un dels aspectes de la

filogènia, i que l'altre, l'anagènesi (és a dir, les transformacions o canvis acumulats entre dos brancatges), era completament ignorat. E. Mayr va proposar el nom de cladisme (del grec *Klados*: branca) per a aquesta taxonomia, per distingir-la de l'autènticament filogenètica que, segons el mateix Mayr, inclouria ambdós components. Evidentment, considerar a la filogènia com a la base de la classificació havia d'anar lligat a la proposició d'una metodologia per a recuperar les relacions filogenètiques. Ha estat precisament aquest component el que amb més força s'ha desenvolupat, esdevenint la metodologia pràcticament universal per a la reconstrucció filogenètica.

Potser la forma més simple de caracteritzar el cladisme en front d'altres metodologies taxonòmiques és a través del tractament que cadascuna d'elles fa dels caràcters. Un caràcter és una característica, una part observable, un atribut d'un organisme, que pot ser adequadament descrit o definit (Wiley i col. 1991). En fenètica, els caràcters son tractats tots com a iguals, i es parla de **semblança global** (*overall similarity*). El conjunt de caràcters considerat, mitjançant una determinada funció de similitud (o dissimilitud), queda substituït per un determinat valor que representa la proximitat (o distància) entre els diferents taxons. En taxonomia evolutiva es distingeixen dos tipus de caràcters: els **homòlegs** i els **homoplàsics**. Dos caràcters son homòlegs si estan relacionats evolutivament, de manera que un prové de la transformació de l'altre o tots dos provenen de la transformació d'un tercer. Al conjunt de caràcters homòlegs se l'anomena **sèrie de transformació** (Wiley i col. 1991). Tanmateix, aquesta nomenclatura està bastant en desús, i a la actualitat hom tendeix a referir-se majoritàriament a la sèrie de transformació com el caràcter, i als caràcters com els **estats** del caràcter. D'altra banda, dos caràcters són homoplàsics si no estan relacionats evolutivament, és a dir, si provenen de transformacions independents d'un caràcter anterior. Aquesta definició d'homoplàsia pot refinar-se diferenciant-ne tres tipus diferents (**fig. 9**): **convergències** o transformacions a un mateix caràcter en línies evolutives molt distanciades, **paral.lelismes** o transformacions al mateix caràcter en línies molt properes i **reversions**, transformacions d'un caràcter a una forma anterior. Els caràcters basats en seqüències nucleotídiques presenten una forma addicional d'homoplàsia. Com

que només existeixen cinc caràcters (o estats) possibles (A, C, G, T, -), la probabilitat de compartir el mateix caràcter per canvis independents és extremadament alta. Si es compara una determinada posició nucleotídica (=sèrie de transformació o caràcter) en taxons molt divergents o amb taxes de canvi molt elevades, la superposició de canvis (*multiple hits*) haurà esborrat qualsevol informació filogenètica d'aquestes. L'existència de substitucions múltiples en un determinat conjunt de posicions pot posar-se de manifest observant el nombre de canvis respecte a la divergència entre els taxons. A partir de certs valor de divergència, el nombre de canvis ja no augmenta degut a què aquests es donen en posicions que ja havien canviat anteriorment i, per tant, ja havien estat contabilitzades. Aquest fenòmen s'anomena **saturació**. Finalment, en el cladisme es distingeixen dos tipus diferents de caràcters homòlegs: **plesiomorfies**, o **caràcters plesiomòrfics**, i les **apomorfies**, o **caràcters apomòrfics**. Si dos caràcters són homòlegs, de manera que un es correspon a la transformació evolutiva de l'altre, el primer constitueix una apomorfia i el segon a una plesiomorfia. Si el caràcter és compartit per més d'un taxó, ens referim a ell com a **sinapomorfia** o **simplesiomorfia**, segons els casos. Si, al contrari, el caràcter només està present en un dels taxons, l'anomenem **autapomorfia**. A la **figura 10** s'han representat els diferents tipus de caràcter esmentats. Així per exemple el caràcter cercle fosc és homoplàsic, donat que ha sorgit de la transformació independent del caràcter cercle blanc en dues línies diferents. Tanmateix, cercle fosc i cercle blanc són homòlegs.

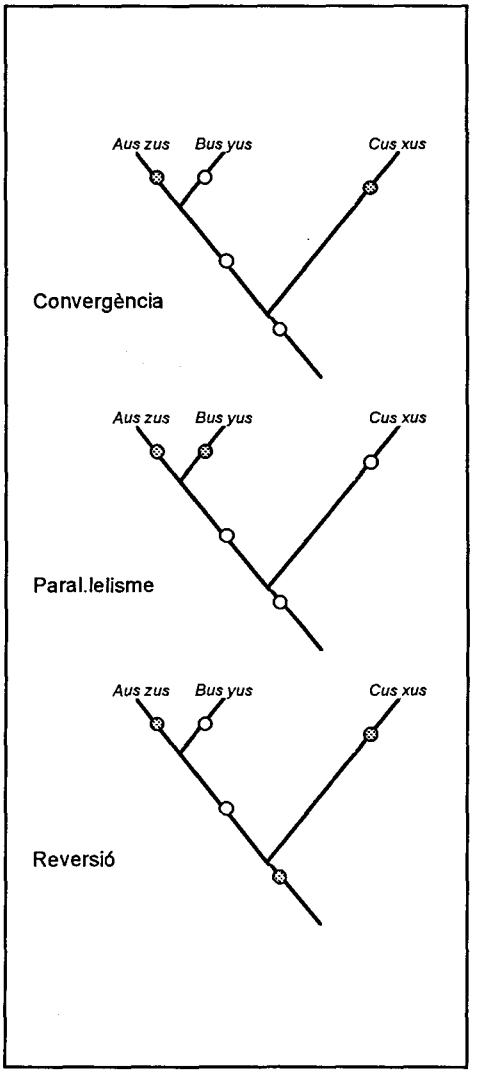


Figura 9.- Representació gràfica dels principals tipus d'homoplàsies

ja què un prové de la transformació de l'altre. De la mateixa manera, tots els caràcters en forma de quadrat i tots els caràcters en forma de triangle son homòlegs entre ells, és a dir formen sèries de transformació, o, en la nomenclatura més utilitzada, son estats del mateix caràcter. Ara bé, en el cas dels triangles, els triangles blancs son plesiomòrfics respecte als foscs, que, per tant, son apomòrfics. Igualment, els quadrats foscs son derivats, sinapomorfs, dels quadrats blancs, simplesiomorfs, malgrat que alhora, son plesiomòrfics respecte el quadrat negre, que deriva d'ells i és, per tant, una autapomorfia.

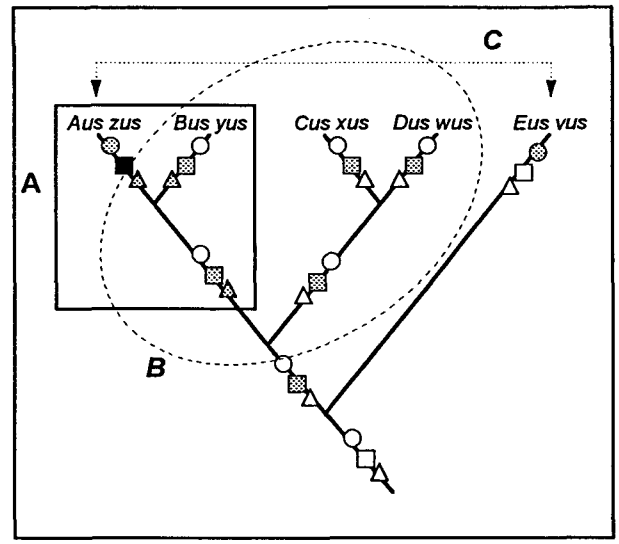


Figura 10.- Filogènia hipotètica de cinc espècies on es mostrn els tipus de caràcters i grups.

La importància de distingir entre els diferents tipus de caràcters esmentats es deu a què cadascun d'ells defineix un tipus d'agrupació. Així, un conjunt de taxons que comparteixen una apomorfia, és a dir una sinapomorfia, constitueixen un grup **monofilètic** (grup A, **fig. 10**). Un conjunt de taxons amb una plesiomorfia comuna, una simplesiomorfia, formen un grup anomenat **parafilètic** (grup B, **fig. 10**), i, finalment, un conjunt de taxons agrupats en base a la presència d'un caràcter homoplàsic, formen un grup **polifilètic** (grup C, **fig. 10**). Les autapomorfies, obviament, no defineixen cap agrupació. Els grups parafilètics i els polifilètics son grups artificials. Només els grups monofilètics, o **clades,** constitueixen grups naturals i, per tant, són els únics que tenen importància filogenètica (Janvier 1984). La incapacitat de l'aproximació fenètica i l'evolutiva o eclèctica de detectar l'existència de sinapomorfies i, per tant, la seva incapacitat de reconèixer la presència de grups monofilètics, les desacredita com a metodologies de reconstrucció filogenètica.

L'elaboració de la filogènia d'un conjunt de taxons és equivalent al procés de recuperar tots els grups monofilètics formats per aquests, el que alhora suposa

establir les sinapomorfies que els defineixen. Per tant, els caràcters representen les evidències que permeten recuperar les relacions filogenètiques dels taxons. La problemàtica de l'establiment de la filogènia d'un grup queda reduïda a l'avaluació d'un conjunt dels seus caràcters per tal de reconèixer les sinapomorfies. Al procés de selecció i avaluació dels caràcters dels taxons estudiats se l'anomena **anàlisi cladística**. El resultat d'aquest anàlisis és un **cladograma**, que és un diagrama en forma d'arbre on es representen la distribució dels diferents caràcters i els grups que defineixen. El reconeixement dels brancatges com a fenòmens d'especiació i dels nodes interns com a avantpassats, converteix el cladograma en una **filogènia**. Finalment, l'establiment dels possibles processos, la inferència del context biogeogràfic o temporal i altres explicacions o implicacions de la configuració obtinguda, constitueixen l'**escenari**. La gran majoria d'autors tendeixen a sinonimitzar cladograma amb filogènia, igualment anàlisi cladística i anàlisi filogenètica. Tanmateix, estrictament parlant, el cladograma és quelcom més ampli, i un mateix cladograma pot correspondre a més d'una filogènia, com es representa a la **figura 11**. Les tres filogènies representades són compatibles amb el cladograma. A la filogènia A, l'espècie *Cus xus* és avantpassat de *Bux yus* que alhora ho és de *Aus zus*. A la filogènia B, *Cus xus* és avantpassat de l'ancestre

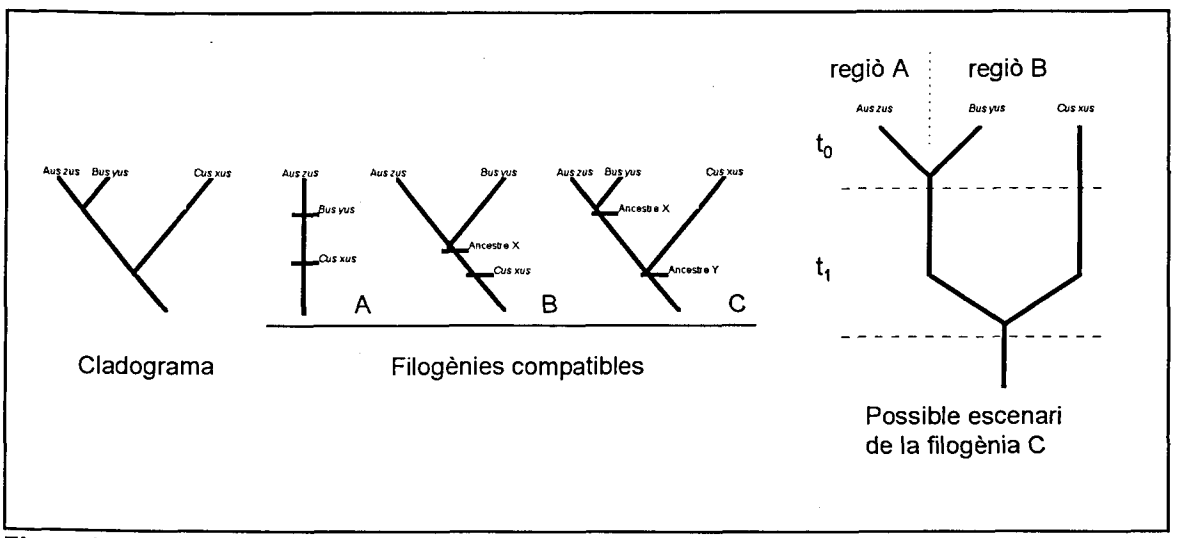


Figura 11.- Diferències entre cladograma, filogènia i escenari. Les tres filogènies representades (A-C) són compatibles amb el cladograma.

comú a *Aus zus* i *Bus yus*. Finalment, a la filogènia C, *Aus zus* i *Bus xus* comparteixen un avantpassat comú que no ho és de *Cus xus*, malgrat comparteixen un altre avantpassat amb aquesta. L'escenari representat és compatible amb la filogènia C, a on s'hipotetitza un temps geològic de separació de *Cus xus* de les altres espècies i un fenomen de vicariància com a explicació de l'origen de *Aus zus* i *Bus yus*.

Ara bé, com es poden reconèixer les sinapomorfies de la resta de caràcters? L'assignació de quins són els caràcters (o estats de caràcter) sinapomòrfics es realitza mitjançant tres passos successius. El primer pas consisteix en establir les sèries de transformació, o els estats dels caràcters, en base a les anomenades **homologies primàries** (De Pinna 1991). Aquestes, que es corresponen al concepte més o menys clàssic d'homologia, han de considerar-se com una primera hipòtesi que ha de ser corroborada en els passos següents. Molts autors actuals prefereixen reservar la paraula homologia per referir-se només a les sinapomorfies, i es refereixen a aquest primer pas com a 'establiment de les hipòtesis d'agrupació' a través de la identificació topogràfica i la identificació dels caràcters (o estats de caràcter) (Brower i Schawaroch 1996). S'ha de tenir en compte a l'hora de definir les sèries de transformació (o estats dels caràcters) l'anomenat **principi auxiliar de Hennig** (Hennig 1953, 1966) que diu: 'mai assumir convergència ni evolució paral.lela, sempre assumir homologia en absència d'evidència contrària'. Des d'un punt de vista operacional, es poden aplicar els criteris de Remane (1956) per establir aquestes primeres hipòtesis d'homologia: (1) similitud de posició (2) qualitat especial (semblança en la microestructura i desenvolupament) (3) continuïtat entre formes intermèdies. El següent pas és determinar la **polaritat** dels caràcters (o dels estats de caràcter), és a dir, quins són els caràcters (o estats de caràcter) primitius, i quins els derivats. Hennig va proposar quatre regles, una de principal i tres d'accesories o auxiliars, per determinar la polaritat dels caràcters:

1. *Regla del grup extern* (esquema d'argumentació o argumentació dels caràcters).

   De dos o més caràcters (o estats de caràcter) homòlegs trobats en un grup monofilètic, aquell que també es trobi al grup germà és l'estat plesiomorf, i el que es trobi només en el grup intern és l'apomorf (Wiley i col. 1991).

2. *Regla de la progressió*: Hennig assumia que els taxons del grup estudiat més primitius es trobarien més propers al centre d'origen del grup, i per tant els caràcters (o els estats) mostrats per aquests serien els plesiomorfs (Janvier 1984):
3. *Regla de la correlació de les sèries de transformació.* La hipòtesi de polaritat d'una sèrie de transformació, o estats d'un caràcter, es corrobora per la congruència amb la distribució d'altres sèries o caràcters.
4. *Regla del criteri ontogenètic.* Malgrat formulat originàriament com a criteri auxiliar, alguns autors han considerat aquest mètode com el més important perquè presenta una evidència directa de la polaritat dels caràcters, mentre que els altres són evidències indirectes (Nelson i Platnick 1981, Patterson 1982). Nelson (1978) va reformular aquesta regla com: 'donada una determinada transformació de caràcter (o estats d'un caràcter), des d'un caràcter (estat) més general a un de menys general, el més general és el primitiu, i el menys general és el derivat'. Per a d'altres autors (Patterson 1982), és l'ordre d'aparició dels caràcters (o els estats de caràcter) durant l'ontogènia de l' organisme allò que es correspon amb la seva polaritat.

Un altre criteri addicional que pot ser utilitzat és la **seqüència estratigràfica** (Gingerich 1979), en la qual la polaritat s'obté a partir de de l'observació de l'aparició dels caràcters en el registre fòssil.

El tercer i últim pas consisteix en diferenciar les autèntiques homologies de les homoplàsies. El criteri principal és el de la **congruència dels caràcters**, del qual la regla 3 anteriorment esmentada és una extensió. L'assumpció bàsica és que les homologies haurien de mostrar un patró congruent entre elles, degut a que són el resultat de la descendència amb modificació. Per contra, la distribució de les homoplàsies hauria de ser incongruent amb la resta de caràcters, ja que aquestes no tenen relació amb la jerarquia resultant de les relacions genealògiques. Un exemple de l'aplicació d'aquest criteri pot trobar-se a la **figura 12**. Suposem que tenim tres taxons i quatre sèries de transformació o caràcters, de les quals s'ha establert quina és la seva polaritat i se sap per tant quines agrupacions suggereixen. En comparar-les, es pot veure que dos d'elles donen suport a una

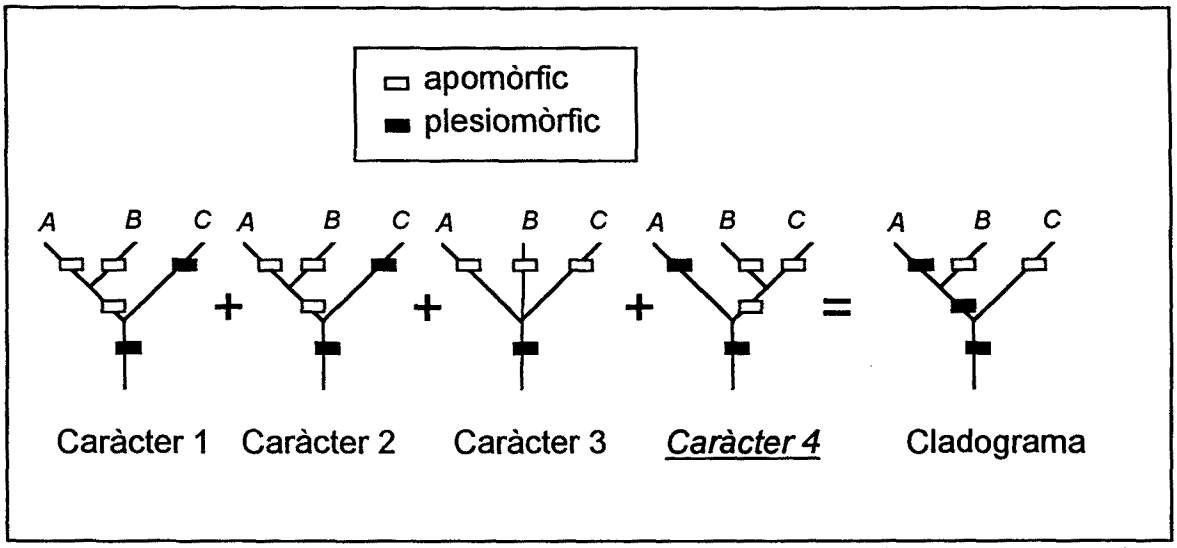


Figura 12.- Test de congruència del caràcters com a criteri per a la distinció entre homologies i homplàsies.

determinada agrupació, una tercera no és informativa i la quarta és discordant. S'infereix d'aquesta última incongruència, que el quart caràcter és homoplàsic. Finalment es mostra la reinterpretació del caràcter sobre el cladograma resultant.

3.3.2. Aproximacions actuals a la inferència filogenètica

Fins aquí s'han discutit els principis i la metodologia del que podria considerar-se com cladisme Hennigià o filogenètica sistemàtica clàssica. Tanmateix, el desenvolupament posterior d'alguns d'aquests principis degut a la seva implementació als ordenadors i a l'aparició de nous tipus de caràcters, principalment les dades moleculars i més concretament les seqüències nucleotídiques i ha donat lloc a l'aparició de

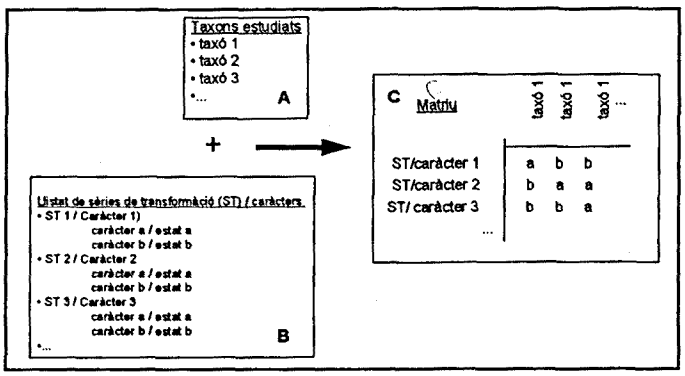


Figura 13.- A: llistat de taxons, B: llistat de sèries de transformació o caràcters, C: matriu de caràcters.

diferents metodologies d'inferència filogenètica. Aquestes, malgrat compartir com a base comuna el reconeixement dels grups monofilètics com a únics indicatius de relacions filogenètiques, difereixen molt en la manera en que aquests són establerts. A l'actualitat hi ha una certa tendència a identificar el cladisme amb una d'aquestes metodologies, la parsimònia, la qual cosa és del tot errònia. El punt de partida de tots aquests mètodes és el mateix: la **matriu de caràcters (fig. 13)**. Aquesta és el resultat de l'assignació a cada taxó objecte d'estudi del caràcter (o estat de caràcter) que mostra de la llista de sèries de transformació (o caràcters) prèviament establerta, utilitzant el principi auxiliar de Hennig (1966) i mitjançant l'aplicació dels criteris clàssics de determinació (Remane 1956).

Els principals mètodes d'anàlisis filogenètica utilitzats actualment són: (1) l'anàlisi de compatibilitat, (2) l'anàlisi de parsimònia, (3) l'anàlisi de màxima versemblança (*likelihood*) i (4) l'anàlisi basada en matrius de distancies.

3.3.2.1. Anàlisi de compatibilitat

Originalment proposat per LeQuesne (1964), posteriorment desenvolupat per Estabrook i Meacham (Estabrook i col. 1976a,b; Meacham 1981, Meacham i Estabrook 1985) i implementat en el programa d'ordinador CLINCH (Fiala i Estabrook 1977), aquest mètode es basa en la congruència dels caràcters, de manera que només els caràcters que recolzen agrupacions de taxons compatibles entre elles reflecteixen la filogènia. L'assumpció és que, com que només hi ha una filogènia dels taxons, els caràcters originats a través d'ella no poden ser incongruents. Ans al contrari, els caràcters que suposen agrupacions de taxons incompatibles amb la resta són caràcters que no estan correlacionats amb la filogènia i per tant no han de ser considerats. L'anàlisi es converteix doncs en intentar determinar el nombre màxim de caràcters compatibles entre ells, el que s'anomena un ***clique***. La pràctica d'eliminar els caràcters no compatibles de la reconstrucció filogenètica ha estat molt criticada. En molts casos el nombre de

caràcters eliminats pot arribar a excedir el nombre de caràcters utilitzats. Alguns proponents d'aquesta aproximació han desenvolupat noves metodologies per reintroduir alguns dels caràcters eliminats en posteriors estadis de l'anàlisi (Sharkey 1989). Tanmateix, la crítica més important a aquest mètode és que els caràcters descartats poden de fet reconciliar-se amb l'àrbre simplement assumint un major nombre de transformacions.

3.3.2.2. Anàlisi de parsimònia

La utilització de la parsimònia en filogènia va ser originàriament introduïda com a criteri de construcció d'àrbres en algorismes computacionals (Edwards i Cavalli-Sforza 1963, Camin i Sokal 1965, Kluge i Farris 1969, Fitch 1971) i posteriorment justificada des d'una òptica filosòfica i epistemològica (Farris 1969, 1974, 1983; Eldredge i Cracraft 1980, Wiley 1981, Kluge 1985, Sober 1988). Com en la metodologia anterior, s'assumeix que l'existència de congruència entre els caràcters és el resultat de l'existència d'una genealogia, és a dir, la presència d'un avantpassat comú. Ara bé, l'homoplàsia no pot ser explicada en aquests termes i algun altre procés, o més probablement conjunt de processos, ha de ser invocat per explicar la seva existència. Alhora, no existeix cap raó per a suposar que aquests processos hagin d'estar relacionats entre ells i, per tant, les homoplàsies no tenen perquè tenir una distribució congruent ni entre elles, ni respecte a la resta de caràcters. Un principi bàsic en ciència és l'anomenada 'navalla d'Ockham', la qual proposa que: 'donades vàries solucions possibles a un problema, la solució més econòmica és la que té preferència sobre les altres' (Scotland 1992). En el present context, la hipòtesi més parsimoniosa és aquella que explica la totalitat de les dades, amb el menor nombre possible d'invocacions de processos alternatius *a posteriori* (Wiley 1981). Un error freqüent és interpretar que la parsimònia suposa que l'evolució ha hagut de donar-se a través dels camins més senzills possibles. La implementació de la parsimònia en la reconstrucció filogenètica no té res a

veure amb aquesta afirmació, de la mateixa manera que tampoc implica que les homoplàsies siguin fenòmens poc freqüents (Farris 1983). Al minimitzar el nombre d'homoplàsies, estem maximitzant el nombre de congruències entre els caràcters, que s'assumeix, són el resultat de les relacions genealògiques d'aquests. Operacionalment, la utilització del criteri de la parsimònia pot interpretar-se de la següent manera: si els caràcters d'una determinada sèrie de transformació (estats del caràcter) són homòlegs, això implicarà que el número de transformacions (canvis) entre ells és el menor possible (que és igual al número de caràcters, o estats, menys u), ja que la seva presència és el resultat de compartir un avantpassat comú; al contrari, si algun dels caràcters de la sèrie de transformació (estats de caràcter) és homoplàsic, això implica que existeix almenys una transformació addicional i, per tant, el nombre de transformacions (canvis) és superior, almenys en un, al mínim. Per tant, en minimitzar el nombre de canvis en el total de les sèries de transformació (o caràcters), estem minimitzant el nombre de caràcters homoplàsics. La implementació del criteri de la parsimònia als ordenadors és força directa, perquè aporta un criteri d'avaluació dels diferents arbres: el cladograma més parsimoniós és aquell amb el menor nombre de canvis. Un dels avantatges de la incorporació del criteri de parsimònia a l'anàlisi cladística i la seva implementació computacional és que fa innecessària la polarització dels caràcters *a priori*, éssent aquesta un dels resultats de la pròpia anàlisi. Aquesta propietat va ser proposada per Farris (1982), qui caracteritzà explícitament la polaritat en termes d'arrelament del cladograma. El raonament és el següent: l'aplicació del criteri de parsimònia suposa contar el nombre de transformacions de caràcter en l'àrbre a avaluar, el que alhora implica assignar un determinat caràcter (o estat) als internodes corresponents als avantpassats hipotètics; això es realitza de manera que el caràcter (o estat) assignat minimitzi el nombre de transformacions. A aquest procés de reconstrucció dels caràcters (o estats) dels avantpassats hipotètics se l'anomena **optimització**. Com que l'anàlisi de parsimònia assigna el caràcter (o estat) als ancestres hipotètics de l'àrbre, si es fixa la seva arrel queda determina automàticament la direcció de transformació dels caràcters (o estats) i, per tant, la seva plesiomorfia relativa (Farris 1982). En conseqüència,

l'única cosa necessària per a polaritzar els caràcters (i per tant separar les apomorfies de les plesiomorfies) és situar l'arrel de l'àrbre, la qual cosa es pot fer introduint en l'estudi un o més taxons externs (*outgroups*), analitzant tots els taxons conjuntament, i finalment situant l'arrel en l'internode que uneix als taxons interns amb els externs. Aquesta metodologia permet alhora avaluar la monofília del grup intern, afegint a l'anàlisi diversos *outgroups* i comprovant si l'assignació de l'arrel és compatible amb la mateixa. La relació entre polaritat, *outgroups* i arrelament ha estat recentment sistematitzada per Nixon i Carpenter (1993). Aquesta propietat no és exclusiva de la parsimònia, i és aplicable a qualsevol metodologia que permeti l'assignació de l'arrel a qualsevol internode de l'àrbre, sense canviar el valor d'aquest sota el criteri d'optimització o l'algorisme de construcció escollit.

La principal diferència de la parsimònia respecte a la compatibilitat, és que els caràcters homoplàsics no són eliminats, sinó simplement minimitzats en el context de tota l'evidència existent.

Un dels avantatges de la parsimònia sobre altres mètodes és que la seva aplicació és 'agnòstica' en relació als detalls del procés evolutiu, i reflecteix simplement el suport empíric de les diferents hipòtesis filogenètiques (Farris 1983). Com s'explica més endavant, certes aproximacions a la inferència filogenètica comporten l'assumpció explícita d'un model evolutiu. Tanmateix, alguns autors han assenyalat que la parsimònia, malgrat que de forma implícita, assumeix una sèrie de condicions. Segons Chippindale i Wiens (1994) aquestes condicions són: (1) la independència dels caràcters, (2) la independència dels taxons, és a dir, l'absència d'hibridització, introgressió o transferència lateral, i (3) taxes similars de canvis al llarg de les branques de l'arbre. La necessitat de la independència dels caràcters deriva directament de la concepció hipotètica-deductiva de la inferència filogenètica. La congruència entre els caràcters és la base del suport d'una hipòtesi sobre una altra, si aquesta congruència és producte de la dependència o la relació condicional dels caràcters aleshores queda en entredit. La segona condició és resultat de què, quan les relacions entre els taxons són tocogenètiques (reticulars) les sinapomorfies deixen de ser indicadors inequívocs de descendència d'un avantpassat comú (Nixon i Wheeler 1990). La tercera condició ha estat proposada

per autors que defensen una aproximació estadística al problema de la reconstrucció filogenètica i està basada en el comportament de la parsimònia en simulacions per ordenador amb dades moleculars i assumint determinats models evolutius. Felsenstein (1978) demostrà computacionalment amb quatre taxons que si la taxa de canvis de les branques terminals de dos taxons és molt més gran que l'existent a la resta, la parsimònia sistemàticament uneix aquests taxons, independentment de les veritables relacions genealògiques. Aquesta situació s'anomena **'atracció de les branques llargues'** (*long branch attraction*). D'aquests resultats se'n deriva que la parsimònia assumeix una quantitat de canvis similar entre les branques. Investigacions posteriors han establert que, per a què la parsimònia recuperi l'arbre real, ha d'existir una quantitat similar i constant de canvis entre els caràcters (estats), entre les sèries de transformació (caràcters) i entre els taxons (Yang 1996). L'existència en casos reals d'aquestes situacions encara no ha estat demostrada.

Es poden assenyalar una sèrie de crítiques a la parsimònia. La primera és que no hi ha cap bona raó per assegurar que, en l'evolució, la divergència dels caràcters sigui sempre més comuna que els fenòmens que produeixen l'homoplàsia (Panchen 1979). Com s'ha explicat anteriorment, l'aplicació de la parsimònia no implica que l'homoplàsia sigui un fenomen infreqüent, simplement que aquesta no té per què ser congruent amb la distribució de la resta de caràcters. Evidentment, poden existir casos en els que hi hagi un seguit d'homoplàsies congruents entre elles com a resultat d'un determinat procés comú, com pot ser una adaptació a un medi concret. Per exemple, en reconstruir una filogènia de vertebrats podem tenir caràcters com l'absència de pèl i la presència d'aletes amb una distribució congruent que uneixi peixos amb dofins. Tanmateix, l'addició de més caràcters o inclús la reconsideració dels mateixos (l'estructura anatòmica de les aletes) acabarà per descobrir la incongruència d'aquests caràcters envers molts d'altres. L'adaptació a un medi pot transformar caràcters, però no pot esborrar l'existència dels heretats ni crear-ne *de novo*. La majoria de crítiques, però, provenen de l'aplicació de l'estadística a la reconstrucció filogenètica. Una característica important d'un estimador estadístic és l'anomenada **consistència**. Un estimador estadístic és

consistent si tendeix cap al valor veritable d'una quantitat a mesura que augmenta el nombre de dades (Felsenstein 1988a). Mitjançant la utilització de dades nucleotídiques generades a partir d'un cert model d'evolució de les seqüències (simulació), s'ha demostrat que la parsimònia com a mètode de reconstrucció filogenètica pot ser inconsistent (Felsenstein 1978, Penny i col. 1992). La resposta a aquestes crítiques ha estat destacar la natura teòrica dels models assumits per generar les dades i la manca d'exemples reals que demostrin aquest comportament (Farris 1983). D'altra banda, Yang (1996) assenyala que el criteri de parsimònia utilitza només una part de la quantitat d'informació disponible i que altres mètodes, i en concret la màxima versemblança, utilitzen més informació amb la mateixa quantitat de dades. Això és degut a què en parsimònia només són utilitzats aquells caràcters que suporten certes agrupacions de taxons, mentre que en distàncies o màxima versemblança els invariables poden ser també informatius.

3.3.2.3. Anàlisi de màxima versemblaça (maximum likelihood)

Representa un canvi radical en l'aproximació a la inferència filogenètica, la qual és considerada com un problema essencialment estadístic. Aquesta postura és absolutament irreconciliable amb les anteriorment exposades, en especial amb la parsimònia, ja que suposa un canvi de paradigma. L'ús de la parsimònia en la reconstrucció filogenètica ha estat defensada sota una aproximació hipotètico-deductiva (Popper 1968a.b). En sistemàtica, aquesta es basa en què com que mai no podrem saber quina és la filogènia real dels taxons, tot el que podem fer és adoptar un criteri (parsimònia) per avaluar els mèrits relatius de les hipòtesis proposades (cladogrames), de forma que la hipòtesi escollida sigui la que exhibeixi una major congruència en la distribució de les sinapomorfies, esdevenint aquestes els tests de falsificació de les hipòtesis (Eldredge i Cracraft 1980) (tanmateix, veure Panchen 1992 per a una crítica de la validesa d'aquesta connexió).

La màxima versemblança com a mètode d'inferència filogenètica va ser

introduït per Cavalli-Sforza i Edwards (1967) i desenvolupat i popularitzat per Felsenstein (1981). Com s'ha esmentat anteriorment, aquesta aproximació es basa en plantejar la reconstrucció filogenètica en termes de fer una estima (filogènia) d'una quantitat desconeguda, en presència d'incertesa i utilitzant un model probabilístic del procés evolutiu (Felsenstein 1988a). La màxima versemblança és el mètode més general per a l'obtenció d'estimacions estadístiques. A partir de l'adopció de cert model d'evolució (M) i d'uns caràcters determinats dels taxons (D), la versemblança d'un arbre (T) és la probabilitat de les dades (caràcters), donat aquest arbre i el model assumit, P(D; T, M), i considerada com a funció de l'àrbre (Felsenstein 1988a). Fixat un arbre, la probabilitat de tots els conjunts de dades ha de sumar u. Tanmateix, quan les dades es mantenen constants i s'avaluen diferents arbres, la suma d'aquestes no té per què ser u i, per tant, són considerades versemblances en lloc de probabilitats. L'anàlisi consisteix doncs en obtenir l'arbre amb el màxim valor de versemblança. En general, en l'avaluació de cada arbre no només es té en compte el brancatge, sinó també la llargada de les branques i, en molt casos, l'estimació de part dels paràmetres del model adoptat (Yang 1996). Un corol.lari del mètode és l'existència d'un model subjacent a les dades. Si bé aquests han estat desenvolupats per a dades moleculars, són inexistents, i probablement impossibles, en dades morfològiques de manera que aquesta aproximació queda limitada a la utilització de caràcters moleculars. Aquesta és una de les primeres crítiques que se li poden fer al mètode de la màxima versemblança. Tanmateix, la crítica més important a aquesta aproximació i la més freqüentment esgrimida en la seva contra, és l'adopció d'un model concret d'evolució, que passa per ser del tot indemostrable i, en el millor dels casos, representa una simplificació de la realitat amb uns efectes sobre el resultat final del tot imprevisibles (Carpenter 1992). La resposta a aquesta crítica ha estat la demostració, a través de simulacions, que l'àrbre correcte es recupera en molts casos malgrat que el model assumit sigui erroni (Nei 1996, Yang 1996). Alguns autors també proponents de la visió estadística de la reconstrucció filogenètica han criticat aquest mètode perquè si bé és veritat que l'àrbre amb el màxim valor de versemblança donat un cert espai de probabilitat dels paràmetres és el millor estimador de la filogènia, deixa de ser-ho

quan els arbres comparats pertanyen a diferents espais de probabilitats (Nei 1996). Aquesta situació es dóna quan els paràmetres del model no romanen constants en les comparacions entre seqüències, és a dir, quan el procés evolutiu no és constant, la qual cosa pot considerar-se com a força probable.

3.3.2.4. Anàlisi basada en matrius de distàncies

Aquests mètodes constitueixen els últims romanents del feneticisme en sistemàtica. Es basen en el supòsit de què la distància entre dos taxons, entesa com a mesura de la seva dissimilitud, es relaciona directament amb la seva relació filogenètica. Malauradament, aquesta assumpció és només vàlida en absència d'homoplàsies. Precisament, i com s'ha esmentat amb anterioritat, la manca de discriminació entre homologies i homoplàsies fa que aquesta metodologia no sigui apta per a la reconstrucció filogenètica, almenys amb dades morfològiques. Tanmateix, la possibilitat d'utilitzar models evolutius amb els quals poder corregir les estimacions de les distàncies en funció de l'homoplàsia existent, els ha reservat un lloc en la inferència filogenètica a partir de dades moleculars. A diferència de les altres aproximacions, que es basen en caràcters discrets, aquests mètodes transformen mitjançant l'aplicació d'una certa funció, la matriu de caràcters en una matriu de distàncies (taxons a columnes i a files), a on cada cel.la representa un mesura del grau de dissimilitud entre els dos taxons implicats. La funció per calcular la distància entre els taxons corregeix alhora aquest valor per a l'homoplàsia existent, calculada a partir de l'establiment d'un model d'evolució de les dades (Williams 1992). El pas següent es construir una topologia, un arbre, que pot fer-se de dues maneres: (1) utilitzant mètodes algorísmics, que estableixen un protocol determinat per anar unint els taxons en funció de les seves distàncies, p. ex. UPGMA, Neighbor joining (Saitou i Nei 1987), etc.. o (2) utilitzant un criteri d'optimització per avaluar diferents arbres, p. ex. el mètode dels mínims quadrats (Cavalli-Sforza i Edwards 1967, Fitch i Margoliash 1967), de mínima evolució (Rzhetsky i Nei

1992), etc...

Les mateixes crítiques i contrarèpliques expressades en l'apartat anterior són igualment aplicables a aquest mètode, a excepció del problema dels espais de probabilitat dels paràmetres.

3.3.3. Justificació de la metodologia d'inferència filogenètica adoptada

En el present estudi s'ha adoptat el criteri de parsimònia per a la inferència de les relacions filogenètiques entre les espècies considerades. Com s'ha exposat en l'apartat anterior, existeix en l'actualitat una gran quantitat de mètodes i criteris d'anàlisi filogenètica. Malgrat que la descripció d'aquests mètodes ha estat, per raons d'espai, necessàriament succinta, és suficient per fer palès que, en darrer terme, l'elecció d'un o d'un altre respon en bona part al marc conceptual i filosòfic adoptat per l'investigador. Personalment, considero que l'aproximació de la parsimònia cladista està recolzada per una coherència epistemològica absent a les restants. D'altra banda, entenent, com es fa en aquest treball, que l'elecció d'una determinada filogènia ha de ser el resultat de l'avaluació de tota l'evidència possible, no puc acceptar utilitzar metodologies que descarten part d'aquesta evidència (p. ex. les dades morfològiques). Tanmateix, en escollir aquesta aproximació sóc conscient de l'existència d'arguments en contra, tant teòrics (Mayr i Ashlock 1991, Panchen 1992), com pràctics (Felsenstein 1988a, Yang 1996, Nei 1996, Swofford i col. 1996). Finalment, si bé crec que l'adopció de qualsevol altra aproximació està plenament justificada en el context d'una perspectiva diferent de la reconstrucció filogenètica, considero que la posició diguem-li eclèctica, especialment freqüent entre certs tipus d'investigadors, on l'estudi filogenètic esdevé un llistat de resultats obtinguts aplicant tants mètodes com sigui possible, està mancada de cap mena de legitimació conceptual.

3.3.4. Etapes de l'anàlisi cladística utilitzant màxima parsimònia

A continuació es comenten les diferents etapes de l'anàlisi cladística de parsimònia, amb les possibles opcions existents i la problemàtica concreta de cadascuna d'elles. Alhora, s'expliquen les diferències existents segons s'utilitzin dades morfològiques o dades moleculars, les quals únicament inclouran les seqüències nucleotídiques del DNA. Finalment, es discuteixen les diverses aproximacions al tractament de dades mixtes, en aquest cas concret moleculars i morfològiques.

3.3.4.1. Mostreig taxonòmic

El primer pas de l'anàlisi filogenètica és, obviament, la selecció dels taxons objecte d'estudi. Si bé en alguns casos la selecció dels taxons que formen l'*ingroup*, com en el present estudi on el nombre d'espècies és relativament accessible i està ben definit geogràfica o taxonòmicament, és trivial, en altres casos, principalment degut a la quantitat d'espècies implicades o al límits difosos del grup, aquesta no ho és en absolut. Wheeler (1992) ha demostrat a través de simulacions per ordenador que l'addició de taxons a una matriu de dades resulta en un augment de la precisió del cladograma. Tanmateix, és la selecció de l'*outgroup* la que sol plantejar més problemes. Malgrat que, contràriament al que semblava ser la idea més estesa, l'*outgroup* no té per què ser el grup germà del grup d'estudi (Nixon i Carpenter 1993), el coneixement previ de les relacions externes d'alguns organismes és suficientment pobre com per a què tot i així la selecció d'aquest constitueixi un problema. D'altre banda, quan els caràcters són del tipus seqüències nucleotídiques, la utilització d'un *outgroup* molt divergent pot resultar en l'assignament de l'arrel a l'atzar (*random outgroup*), generalment a la branca interna que acumula més canvis (Wheeler 1990a). Altrament, la funció de l'*outgroup* no té per què ser la d'arrelar l'àrbre. Hi ha ocasions, com en aquest treball, en què un dels objectius és justament testar la hipòtesi de la monofília del grup intern. La

severitat d'aquest test serà funció directa de l'exhaustivitat de la representació del grup extern.

3.3.4.2. Mostreig dels caràcters

Els caràcters representen l'evidència per a la reconstrucció filogenètica, l'absència d'evidència condueix a la incapacitat de caracteritzar part dels possibles grups monofilètics existents en l'estudi. L'efecte sobre l'àrbre és l'aparició de les anomenades **politomies**, nodes dels quals surten més de dues branques. En cladisme s'assumeix que les ramificacions en el cladograma són sempre dicotòmiques, ja que el procés d'especiació o cladogènesi dóna com a resultat dues espècies noves. D'altra banda, en un context hipotètico-deductiu, les hipòtesis més explícites són preferibles a les més ambigües, ja que representen afirmacions més susceptibles de ser refutades (falsejables) (Wiley 1981). A més, un cladograma dicotòmic pot resoldre l'ambigüitat o el conflicte en favor de l'evidència (Miyamoto 1985). El tipus de politomia resultant de la manca d'evidència provocada per un mostreig pobre dels caràcters s'anomena **suau** (*soft polytomy*). Aquesta pot ser igualment deguda a la presència de conflicte entre els grups suportats per dos o més caràcters. Al contrari, les anomenades politomies **dures** (*hard polytomies*), són aquelles en què la manca de suport de les agrupacions s'interpreta com a situació real d'especiació múltiple i, per tant, se les considera hipòtesis filogenètiques legítimes. Tanmateix, en general hom tendeix a preferir la interpretació 'suau', ja que: 'l'experiència demostra que una politomia avui ha desaparegut demà' (Coddington i Scharff 1996). En general, l'addició de nous caràcters a una matriu de dades resulta en un augment de la resolució filogenètica (Wheeler 1992), per tant, en principi, com més caràcters millor. D'altra banda, el nombre de caràcters necessari per a resoldre les relacions entre els taxons dependrà directament de la seva capacitat d'agrupament dels taxons, alhora que de la congruència entre ells. Un caràcter que no dóna suport a cap agrupació entre els taxons és **no informatiu**.

N'hi ha de dos tipus: els caràcters **invariables**, que són els que no canvien d'estat entre els taxons, i els autapomòrfics, que son exclusius d'un dels taxons. Quan hom utilitza dades morfològiques, aquests caràcters no informatius poden ser detectats i eliminats abans inclús de construir la matriu. En canvi, quan es treballa amb dades moleculars, aquests no poden ser caracteritzats fins que s'obté la matriu definitiva. En aquests casos, la quantitat de caràcters informatius depèn directament de la taxa de canvi de la seqüència nucleotídica analitzada. D'aquí es desprèn la importància de la selecció del fragment de seqüència a analitzar, el qual haurà de tenir un nivell de variabilitat adient per esbrinar les relacions entre els taxons estudiats. Afortunadament, els coneixements acumulats a través de l'estudi dels processos d'evolució molecular permeten tenir una informació prèvia força acurada de la utilitat de diferents gens per a resoldre diferents nivells filogenètics (Simon i col. 1994, Brower i DeSalle 1994). Els gens utilitzats en aquest estudi i la justificació de la seva elecció són comentats en l'apartat de resultats moleculars.

D'altra banda, com més diversa sigui la procedència de les dades, p. ex. caràcters morfològics, moleculars, comportamentals, etc.. major serà el poder explicatiu de l'arbre obtingut, ja que aquest serà el resultat de considerar una major part de l'evidència (Kluge i Wolf 1993). La qüestió de combinar caràcters de diferent naturalesa es discuteix amb més profunditat en apartats posteriors.

3.3.4.3. Construcció de la matriu de caràcters: definició i codificació dels caràcters[1]

Aquesta és potser l'etapa més important de l'anàlisi filogenètic, ja que representa el lligam entre les observacions i l'anàlisi, i influencia fortament els resultats (Pleijel 1995). Consisteix principalment en l'establiment de les homologies primàries (de

[1] En tots els apartats precedents, s'ha utilitzat indistintament els binomis de termes 'sèrie de transformació-caràcter', per una banda, i 'caràcter-estat de caràcter', per l'altre. Tanmateix, a partir d'aquest moment s'utilitzarà únicament la nomenclatura 'caràcter-estat de caràcter', tot i recordant que ambós són sinònimes.

Pinna 1991), el grau de corroboració de les quals determinarà el cladograma, distingint les autèntiques homologies (homologies secundàries) de les homoplàsies, i la seva codificació. Malgrat la importància d'aquesta fase, la situació actual sembla ser la de concentrar-se en els resultats de la manipulació de la matriu amb diferents programes d'ordenador, emfatitzant qualsevol elucubració extreta de la matriu per sobre de les observacions que són l'origen de la mateixa (Patterson i Johnson 1997). Tanmateix, si els caràcters han estat erròniament identificats o codificats tot el que d'ells es desprengui serà fals.

La delimitació dels caràcters es realitza a dos nivells: al de la seva identitat topogràfica i al de la identitat dels seus estats (Brower i Schawaroch 1996). En el primer, s'estableix el conjunt de característiques comparables entre els taxons. En el segon, fent ús dels criteris habituals per a la delimitació de les homologies potencials (Remane 1952, 1956; Patterson 1982, 1988, de Pinna 1991), les diferents manifestacions dels caràcters es classifiquen com a iguals (mateix estat) o diferents. Operacionalment, aquests dos passos corresponen a: (1) definir les columnes de la matriu de caràcters i (2) omplir les seves cel.les. Cal destacar que caràcters o estats de caràcter no són categories absolutes i el que en un estudi és un caràcter pot esdevenir un estat en un altre a una escala jeràrquica diferent. Sota aquesta perspectiva de l'establiment de les homologies primàries, el fet de treballar amb caràcters morfològics o caràcters moleculars planteja problemes molt diferents. Així, quan es treballa amb caràcters morfològics, la delimitació espaial del caràcter és relativament fàcil, mentre que la identificació dels estats, és a dir, decidir quines manifestacions del caràcter són genuïnament distintes i quines poden considerar-se iguals pot resultar força complex. Com que la identificació dels estats implica una 'discretització' de les dades, la utilització de caràcters continus comporta tot un seguit de problemes (Felsenstein 1988b, Chapill 1989, Stevens 1991). Els caràcters moleculars només tenen cinc manifestacions, representades pels quatre nucleòtids (A, C, G, T) i els *gaps* (-), per la qual cosa la definició dels estats resulta trivial. En canvi, l'assignació de les posicions que corresponen al mateix caràcter pot convertir-se en un veritable problema. El procés a través del qual s'estableixen les identitats topogràfiques de les posicions nucleotídiques s'anomena **alineació de les**

seqüències. En aquelles situacions on els fragments de la seqüència tenen el mateix nombre de nucleòtids entre els taxons estudiats, com és freqüentment el cas en gens codificadors de proteïnes, l'alineació no és problemàtica. Tanmateix, hi ha casos en que aquests poden diferir, degut a l'existència de fenòmens *indel*, és a dir, a la inserció o deleció d'algun nucleòtid. L'alineació d'aquestes seqüències passa per reconèixer a on han tingut lloc aquests fenòmens, incorporant espais (*gaps*) per tal de conservar l'homologia posicional. Aquesta és la situació habitual quan es treballa amb gens estructurals, com els gens ribosòmics o fragments no codificants. L'alineació pot fer-se de dues maneres: manual o automàticament. L'alineació manual es basa en el reconeixement visual d'alguns motius més o menys conservats i en la minimització de la incorporació de *gaps*. Una manera de millorar la qualitat de les alineacions manuals és la incorporació, en cas que existeixi, d'informació estructural. Així, per exemple, el gens ribosòmics tenen estructura secundària en la molècula transcrita, amb zones complementàries (*stems*) i zones de cadena senzilla (*loops*). El reconeixement de les zones complementàries en la seqüència pot ajudar molt a la determinació de les posicions homòlogues (Kjer 1995). Malgrat això, hi ha ocasions en què la divergència entre les seqüències és tan gran que l'assignació d'homologia es veu seriosament compromesa. Arribats a aquests punts, la millor opció és simplement eliminar aquestes zones de l'anàlisis posterior. L'alineació manual ha estat criticada per manca d'objectivitat i de repetitivitat, tant en l'assignació de posicions com en l'eliminació de certes zones de l'anàlisi (Gatesy i col. 1993). Els alineadors automàtics es basen en l'optimització d'una certa funció de les coincidències de nucleòtids, les substitucions entre aquests i la introducció dels *gaps*. L'addició d'aquests últims ha de penalitzar-se, ja que si no qualsevol parell de seqüències podria ser perfectament alineada insertant suficients *gaps* (Hillis i col. 1996). S'ha proposat que l'alineació de les seqüències hauria de formar part del procés d'inferència filogenètica (Sankoff i col. 1973). Aquesta aproximació ha estat implementada en alguns algorismes d'alineació automàtica (Wheeler i Gladstein 1994, Wheeler 1996), els quals a través de l'optimització d'una certa funció, determinen l'alineament o alineaments que globalment minimitzen les diferències

entre les seqüències. Malgrat que els alineadors automàtics ofereixen *a priori* un criteri objectiu d'alineació, la necessitat d'incorporar-hi els valors de diferents paràmetres, com el cost de les substitucions entre els nucleòtids, o la penalització per l'addició bé de nous *gaps* o bé de l'augment de la mida dels *gaps* previs, introdueix novament una certa subjectivitat. Tanmateix, el punt més negatiu de la utilització d'aquests algoritmes, i la raó principal que ens ha dut a rebutjar-los en aquest estudi, és l'elevat cost computacional que suposen, i que comporta en alguns casos la inviabilitat de l'obtenció d'una alineació fiable. D'altra banda, el principal avantatge d'aquests mètodes d'alineació és la seva repetitibilitat.

La **codificació** dels caràcters consisteix en l'assignació d'un símbol alfabètic o numèric a cadascun dels estat definits per a cada caràcter. Generalment, els estats dels caràcters morfològics es codifiquen amb números, ja sigui començant pel 0 o per l'1, mentre que per als nucleotídics s'utilitzen les inicials de la base (A, C, G, T) o el signe (-) si són *gaps*. Atenent al nombre d'estats, els caràcter poden ser **binaris**, si només tenen dos estats, o **multiestats**, si en tenen tres o més. Evidentment, els caràcters moleculars són sempre multiestats. La codificació, que a primera vista pot semblar una operació directa i senzilla, pot plantejar molts problemes, i es converteix en un dels passos a on es prenen decisions que poden tenir una gran influència en el resultat final. Els punts més conflictius apareixen, d'una banda, en el tractament de les **dades incertes** (***missing data***) i, de l'altra, en la **ponderació dels caràcters** (***character weighting***).

3.3.4.3.1. Missing data

Les dades incertes o *missing data* corresponen a situacions a on no existeix cap estat per a definir una determinada situació en algun dels taxons d'estudi. Això pot ocórrer per tres raons (Platnick i col. 1991b), (1) **l'estat és inexistent**: el taxó en qüestió té un dels estats definits, però no és observable en l'individu concret que s'estudia, ja sigui per que es troba en un mal estat de preservació, perquè està en

un estadi de desenvolupament a on el caràcter no existeix, o perquè uns determinats nucleòtids no s'han pogut seqüenciar, etc...(2) **el caràcter és inexistent**: el caràcter definit no existeix en el taxó o (3) **el caràcter és polimòrfic**: el taxó presenta més d'un estat possible per a aquest caràcter. Malgrat tots tres casos representen situacions molt diferents, reben, en molts casos, el mateix tractament a l'anàlisi: 'estat desconegut' o '?'. D'aquesta manera, com que cap d'ells no té l'estat definit no influencien el resultat final de l'anàlisi. Per a aquest caràcter en concret, el taxó podrà ser assignat a qualsevol grup sense canviar el nombre de passos del cladograma. Si bé el primer tipus de *missing data* no planteja majors problemes, ja que pot ser codificat estudiant un altre individu o, si més no, la codificació com a interrogant reflecteix exactament la seva situació, això no és així en els altres casos. En el cas dels caràcters polimòrfics, codificar-los com a interrogants amaga l'existència d'una transformació entre estats (Nixon i Davis 1991). Això suposa que el nombre de passos de l'arbre final serà una subestima i, per tant, que aquest podria no ser el més parsimoniós si aquest caràcter s'hagués codificat amb un estat concret. Les solucions proposades a aquest problema són (1) intentar establir a partir d'altres fonts d'evidència quin seria l'estat derivat pel taxó polimòrfic (Wiley i col 1991), (2) separar el taxó en diferents taxons terminals cadascun amb un dels estats possibles, (3) utilitzar un programa d'ordenador que implementi el tractament de dades polimòrfiques. Així, cal esmentar que si bé alguns programes permeten introduir els possibles estats d'un caràcter polimòrfic (PAUP 3.1 Swofford 1993; PAUP* 4.0 Swofford en prep., PHYLIP Felsenstein 1993, NONA 1.5.1, PEE-WEE 2.5.1, PHAST 1.1, SPA 1.1 Goloboff 1996a,b,c,d), altres només admeten la seva codificació com a interrogants (HENNIG86 Farris 1988, MEGA Kumar i col 1994). Tanmateix, només PAUP, NONA, PEE-WEE, PHAST i SPA utilitzen realment el polimorfisme com a caràcter sistemàtic, i tracten el taxó polimòrfic com si fós un node intern que portés a tants taxons terminals com estats existents, els quals són obligats a romandre monofilètics (Goloboff 1996b). Finalment, cal esmentar que existeix una codificació sistematitzada pels polimorfismes nucleotídics: R=AG, Y=CT, M=AC, K=GT, S=CG, W=AT, H=ACT, B=CGT, V=ACG, D=AGT, N=ACGT. Els caràcters inexistents plantegen una

problemàtica diferent. Degut a què durant l'anàlisi s'assigna a cada node intern l'estat probable més parsimoniós, al node que porta als taxons als quals els hi manca aquest caràcter se li assigna d'igual forma un estat, tot i que aquest és impossible (Platnick i col. 1991b). Això pot dur a l'obtenció de resultats no esperats ni desitjats quan s'analitzen les dades amb programes de parsimònia . La raó d'això, és que segons quin estat s'hagi assignat, el nombre de passos final pot variar, descartant resolucions per ser menys parsimonioses quan en realitat ho són igual o, el que és el mateix, escollint solucions que poden no ser les més parsimonioses (Maddison 1993).

Tanmateix, aquest efecte no es dóna sempre. Si només hi ha un taxó terminal amb el caràcter inaplicable o si tots els taxons 'inaplicables' formen un clade o, al revés, si tots els taxons amb el caràcter present formen un clade, no hi haurà cap problema. En realitat, aquest fenomen és un cas particular d'un problema més ampli que té a veure amb la definició dels caràcters i dels seus estats. El següent exemple ajudarà a entendre aquest punt (modificat de Pleijel 1995). Donada la variabilitat representada en la **figura 14**, es podrien definir els caràcters de fins a quatre maneres diferents (A-D):

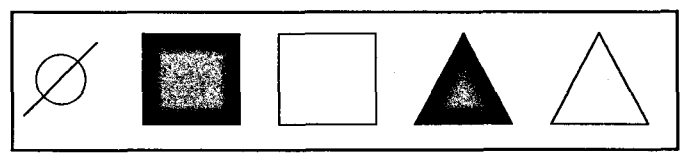


Figura 14.- Distintes manifestacions d'un caràcter: absent, quadrat negre, quadrat blanc, triangle negre, triangle blanc.

A (o codificació composta).

1. Caràcter X: absent (0); present, quadrat i negre (1); present quadrat i blanc (2); present, triangular i negre (3); present, triangular i blanc (4).

B.

1. Caràcter X: absent (0); present, quadrat (1); present, triangular (2).
2. Caràcter Y: absent (0); present, negre (1); present blanc (2).

C.

1. **Caràcter X: absent (0); present (1).**
2. **Caràcter Y: quadrat (0); triangular (1).**
3. **Caràcter Z: negre (0); blanc (1).**

D (o codificació reductiva).

1. Caràcter X: absent (0); present (1).
2. Caràcter X quadrat: absent (0); present (1)
3. Caràcter X triangular: absent (0); present (1)
4. Caràcter X negre: absent (0); present (1)
5. Caràcter Z blanc: absent (0); present (1)

Es pot observar fàcilment que només el tipus de delimitació **C** (en negreta) introdueix caràcters inexistents a la matriu, per tant, la utilització de qualsevol dels altres tipus evitaria el problema. Tanmateix, l'elecció d'un tipus o un altre de delimitació es basa en diferents criteris, i totes elles poden induir a algun error en el resultat. Entre els criteris que poden utilitzar-se per escollir una forma o un altre es destaquen (a part dels *missing data*):

1. *Independència dels caràcters* (Pleijel 1995, Wilkinson 1995a). Com s'ha esmentat anteriorment i per les raons adduides, la parsimònia assumeix la independència dels caràcters. Sota aquesta perspectiva, el tipus A, o codificació composta (Wilkinson 1995a), és l'únic que minimitza la possibilitat de dependència entre aquests. En canvi, en el tipus B i, de forma molt més exagerada, en el tipus D, o codificació reductiva (Wilkinson 1995a), l'estat absent es repeteix en alguns caràcters, quan, obviament, s'està referint a la mateixa característica. L'efecte d'aquesta multiplicació de l'estat en l'anàlisi és que els grups suportats per aquest estaran artificialment afavorits enfront de la resta.
2. *Independència de la jerarquia* (Pleijel 1995, Wilkinson 1995a). Com també ha estat comentat, l'homologia dels caràcters és un concepte relatiu que depèn directament de l'escala filogenètica que s'estigui estudiant. Així, una determinada característica que a cert nivell està o no present, pot tenir diferents formes si s'observa a una escala més fina. La codificació reductiva (D) és l'única que evita haver de redefinir els caràcters cada cop que s'afegeixen a la matriu nous taxons amb diferents combinacions d'aquests.
3. *Recuperació i contrast de la informació* (Pleijel 1995, Wilkinson 1995a). Aquest

criteri fa referència a l'eficiència de cadascun dels tipus esmentats per incorporar les observacions sense pèrdua de dades, alhora que a la forma més apropiada per a ser testada per congruència. Tenint en compte que la congruència es mesura entre els caràcters i no entre els estats definits dins un caràcter, els quals s'assumeixen, les codificacions A-C comporten un menor contrast de les dades. Contràriament, aquest es maximitza en la codificació reductiva, a on tota la variabilitat pot ser contrastada. D'altra banda, l'existència d'estats múltiples en les codificacions A-C fa que la informativitat d'aquests sigui menor, ja que el suport a un determinat grup per part d'algun dels estats dependrà de la situació de la resta d'estats. Això no passa en la codificació reductiva.

4. *Reconstrucció dels estats dels avantpassats* (Wilkinson 1995a). En el cas de la codificació composta, només un dels estats definits pot ser assignat a un avantpassat. En canvi, si s'utilitza la codificació reductiva, aquest caràcter pot separar-se en més d'un, de forma que l'avantpassat pot presentar-ne una combinació no observada en cap dels taxons terminals

Com es pot comprovar, existeixen criteris a favor i en contra de cadascun d'aquests mètodes de delimitació dels caràcters. Si bé les codificacions més reductives presenten avantatges en quant a contrast, estabilitat i reconstrucció evolutiva, aquestes poden afectar negativament l'anàlisi sobrevalorant alguns dels caràcters, i els grups que aquests suporten. Cap mètode destaca de manera absoluta sobre els altres, i cal considerar les diferents alternatives a l'hora de construir els caràcters.

En el context de les dades moleculars, un problema similar al representat pels caràcters inaplicables o inexistents es dóna amb el tractament dels *gaps*. Així, pot observar-se que en moltes anàlisis moleculars publicades, els *gaps* son tractats com si fossin *missing data*. Tanmateix, assumint que la incorporació d'aquests *gaps* a les seqüències ha estat el resultat de la seva alineació, és a dir, de l'establiment de les hipòtesis d'homologia posicional, sembla il.lògic no considerar-los com un cinquè estat (Wheeler 1993). D'altra banda, tractar els *gaps* com a cinquè estat té

l'inconvenient de què cadascuna de les posicions d'una inserció-deleció múltiple (que inclou més d'un *gap*) es considerada com a un caràcter independent, quan certament poden no ser-ho (Maddison i Maddison 1992). El resultat d'incloure caràcters dependents en una anàlisi és la sobrevaloració dels grups que aquests suporten.

Deixant de banda els problemes que pugui representar l'agrupació de situacions clarament diferents sota el mateix codi, la pròpia existència d'interrogants en la matriu pot tenir efectes negatius en l'anàlisi filogenètica (Nixon i Davis 1991, Platnick i col. 1991b). Un d'aquests és la multiplicació del nombre d'arbres igualment parsimoniosos. Degut a l'absència en els taxons amb *missing data* dels caràcters que suporten els grups, aquests poden ajuntar-se a un grup o un altre sense variar el nombre de passos final. Malgrat que a primer cop de vista la millor solució podria semblar minimitzar el nombre d'interrogants eliminant els taxons que en són responsables, això podria comportar alteracions importants en les relacions inferides. Aquesta seria la situació si, per exemple, el taxó eliminat presenta en els caràcters coneguts combinacions absents en els altres taxons (Wilkinson 1995b).

3.3.4.3.2. Ponderació dels caràcters (character weighting) i dels canvis entre estats (transformation costs)

Tota l'evidència comparativa té valor potencial en la inferència filogenètica. Tanmateix, són tots els caràcters igualment bons com a indicadors de les relacions filogenètiques? o, de la mateixa manera, són totes les transformacions entre els estats igualment probables evolutivament? L'opinió més generalitzada és que ni tots els caràcters aporten la mateixa evidència, ni tots els canvis d'estat poden considerar-se de la mateixa manera (Farris 1969, Williams i Fitch 1989, Wheeler 1990b). En aquest context sorgeix la necessitat d'introduir a l'anàlisi les possibles consideracions, hipòtesis o assumpcions, sobre el valor relatiu tant dels caràcters com de les transformacions dels seus estats. Per tal de facilitar la discussió, ens referirem al valor relatiu assignat a un determinat caràcter enfront dels demés com

el **pes** (weight) del caràcter, i al valor del canvi d'un estat particular a un altre com a **cost** de la transformació. Alguns autors, tot i acceptant els valor diferents de certs caràcters i transformacions, consideren que la ponderació d'aquests és una forma d'incorporar a l'anàlisi models evolutius, que representen assumpcions que l'allunyen de les seves bases empíriques (Siebert 1992). Sota aquest punt de vista, l'anàlisi cladística s'hauria de dur a terme sense ponderar ni els caràcters ni els canvis d'estats. D'altra banda, no ponderar és operacionalment equivalent a ponderar uniformement, la qual cosa no deixa de ser una assumpció externa a l'anàlisi (Wheeler 1983). A aquesta aproximació se l'anomena **parsimònia uniformement ponderada** (*equally* o *uniformly weighted parsimony*), en oposició a la **parsimònia amb ponderació diferencial** (*differentially or non-uniformly weighted parsimony*). En qualsevol cas, el problema principal consisteix en determinar quins són els pesos i/o els costs a assignar. Així, la utilització d'una ponderació uniforme pot justificar-se en la manca d'evidència en favor de l'aplicació d'un pes (o cost) determinat a uns caràcters (o canvis) en front d'uns altres, com és freqüentment el cas quan es treballa amb dades morfològiques. La ponderació de caràcters i canvis pot ser *a priori*, si es realitza a partir d'assumpcions o observacions prèvies a l'anàlisi, o *a posteriori*, si s'utilitza evidència derivada d'una anàlisi inicial.

Cost de les transformacions

Les relacions entre els estats d'un caràcter, és a dir, el cost de tots els canvis possibles entre ells, pot representar-se en una **matriu de costs** o **matriu de passos** (*step-matrix*) (Sankoff i Rosseau 1975). La **figura 15** il.lustra un exemple de matriu de costs. A les files es representa l'estat inicial, previ al canvi, del caràcter, i en columnes l'estat final, resultat de la

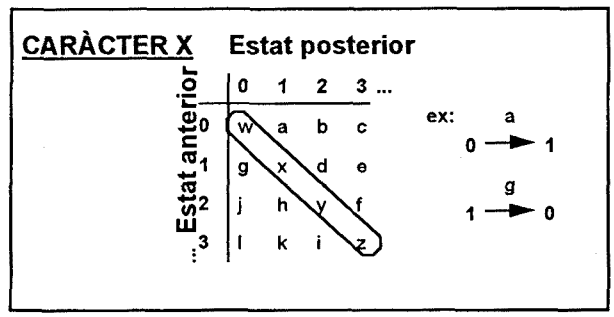


Figura 15.- Exemple de matriu de costs.

transformació. A les cel.les hi figura el valor numèric assignat a la transformació entre dos estats concrets. Aquestes matrius són **simètriques** si la transformació d'un estat *i* a un estat *j* té el mateix cost que el canvi invers, en cas contrari la matriu resultant és **asimètrica**. Sota el criteri de parsimònia, només aquells estats de caràcter que defineixen grups són informatius. Això queda reflectit en la matriu de costs no assignant cap valor a la transformació d'un estat a ell mateix. Amb tot, els autors que defensen l'adopció extrínseca de models evolutius proposen obtenir la matriu a partir de models probabilístics que especifiquen les probabilitats relatives tant de les transformacions entre els estats com del manteniment del mateix estat (Maddison i Maddison 1992). Així, en aquests casos, la diagonal pot tenir assignats costs. Cal destacar que les matrius de costos derivades a partir de models probabilístics converteixen a la reconstrucció més parsimoniosa dels estats ancestrals, també en la màxima probabilitat Bayesiana de l'estima d'aquests caràcters (D.R. Maddison 1990, Maddison i Maddison 1992). Sota aquesta perspectiva probabilística i utilitzant dades moleculars, s'ha suggerit la possibilitat de calcular la matriu de costos a partir de les estimes de màxima versemblança dels canvis entre estats (Yang 1994). L'avantatge d'aquesta aproximació és que computacionalment la implementació de la parsimònia és molt més eficient que la de la màxima versemblança i permet cerques més acurades de l'arbre òptim (Swofford i col. 1996).

Malgrat que, contràriament a allò comentat en línies anteriors, hom prefereix no adoptar un model probabilístic de l'evolució dels caràcters, encara pot incloure certes limitacions als canvis entre els estats basant-se en cert coneixement apriorístic de les seves característiques. En el cas més senzill, que correspondria a una situació a on no existeix cap mena de coneixement sobre la possible relació entre els estats, qualsevol estat pot canviar directament a qualsevol altre. Els caràcters tractats d'aquesta forma s'anomenen **desordenats** o **no additius**. L'algorisme pel càlcul del nombre de passos quan tots els caràcters son desordenats va ser desenvolupat per Fitch (1971), i a la seva aplicació se la coneix com a **parsimònia de Fitch**. Una altra relació entre els estats pot ser la representada per una sèrie de transformacions lineals, de forma que el canvi entre

estats contigus sigui menys costós que entre estats que no ho són. Un exemple a on podria aplicar-se aquesta assumpció seria la d'un caràcter definit com a 'pilositat', on els estats fossin: glabre, 'poc pilós', 'molt pilós'. No seria estrany plantejar la possibilitat de què la transformació de 'molt pilós' a 'glabre', passi necessàriament per la presència de l'estat 'poc pilós'. Els caràcters que incorporen aquesta assumpció s'anomenen **ordenats** o **additius**. Els caràcters binaris representen el cas més simple de caràcters ordenats. Kluge i Farris (1969) i Farris (1970) van proposar l'algorisme pel càlcul del nombre de passos de l'arbre més parsimoniós donada aquesta premissa, al qual hom es refereix com a **parsimònia de Wagner**. Dos casos particulars d'aquesta última són les anomenades **parsimònia de Dollo** (Farris 1977) i la de **Camin-Sokal** (Camin i Sokal 1965). En el primer cas, a l'assumpció de caràcters ordenats se li suma la de l'afavoriment de les homoplàsies degudes a reversió per sobre de les degudes a convergència. Això s'aconsegueix fent més costoses les transformacions en un sentit que en l'altre (**fig. 16**). L'anàlisi de caràcters derivats d'enzims de restricció, constitueix un exemple d'aplicació d'aquestes restriccions: la pèrdua d'una diana de restricció (que suposa com a mínim el canvi d'un dels seus nucleòtids) és en general més fàcil que el guany d'un nova diana (en el pitjor dels casos suposa el canvi de quatre o sis nucleòtids, segons els tipus d'enzim). La parsimònia de Camin-Sokal assumeix l'ordenació dels caràcters, el coneixement sobre la seva polaritat, és a dir,

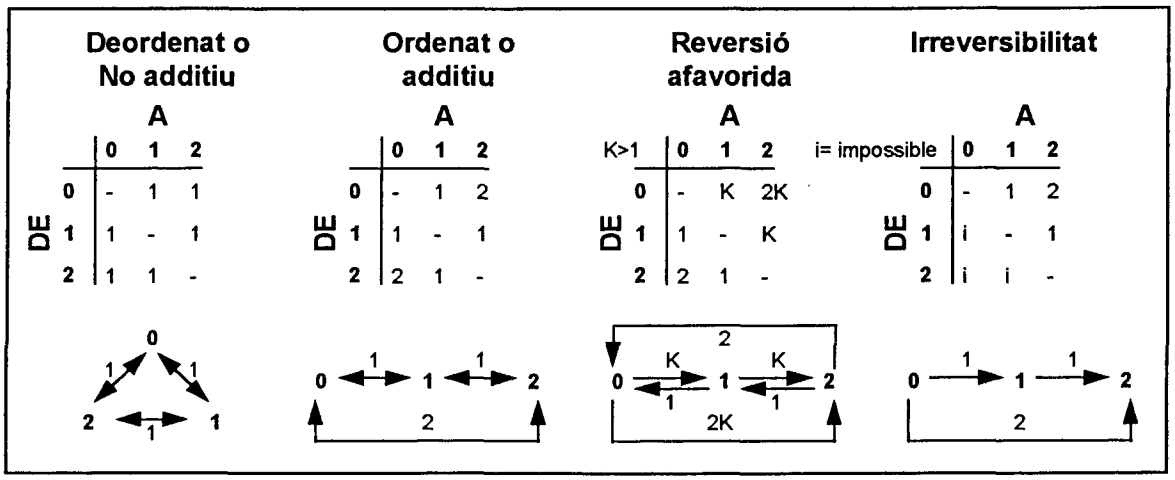


Figura 16.- Algunes relacions apriorístiques entre els estats d'un caràcter.

l'establiment de l'estat plesiomòrfic,i la irreversibilitat dels canvis. Aquesta darrera premissa s'incorpora a la matriu prohibint la transformació en un dels sentits (**fig. 16**). Aquest tipus de parsimònia és rarament utilitzat degut a la dificultat de justificació de la irreversibilitat, tant en dades morfològiques com en moleculars (Kitching 1992). Els diferents tipus d'assumpcions enumerades constitueixen de fet casos particulars d'allò que s'ha anomenat **parsimònia generalitzada** (Swofford i col. 1996). El desenvolupament d'algorismes capaços de calcular exactament el nombre de passos totals donada una certa matriu de costos (Sankoff i Cedregen 1983), permet incorporar a l'anàlisi un ventall amplíssim de models evolutius mitjançant l'assignació d'un valor de cost determinat a un tipus particular de transformació. En qualsevol cas, l'adopció d'un determinat model d'evolució, restricció o ponderació sobre les transformacions dels estats, ha de ser degudament justificada, alhora que formulada de manera explícita i clara (Wheeler 1986).

A diferència dels caràcters morfològics, els caràcters derivats de seqüències nucleotídiques tenen un nombre limitat d'estats possibles que és compartit potencialment per totes les posicions (=caràcters). En aquestes dades, la principal font d'homoplàsia són les substitucions múltiples en una determinada posició i la probabilitat de què aquestes hagin tingut lloc està directament relacionada amb la freqüència de canvi de la posició. De la mateixa manera, com més freqüent sigui un tipus concret de transformació (=substitució), més probable serà que amagui l'existència de canvis múltiples i, també, que hagi ocorregut independentment a diversos llinatges. Per tant, la resol.lució filogenètica d'aquest tipus de dades pot veures potenciada si es dóna un major cost a les transformacions menys freqüents (Williams i Fitch 1989). Una forma d'evidenciar l'existència de substitucions múltiples és mitjançant les anomenades **corbes de saturació**

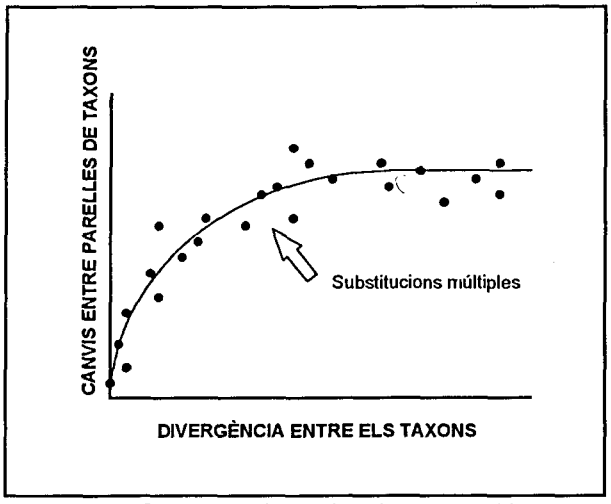


Figura 17.- Exemple de corba de saturació.

(**fig. 17**), que són gràfics a on es representa el nombre d'una determinada classe de transformació, present entre totes les comparacions possibles de taxons dos a dos, respecte a una mesura del nivell de divergència entre aquests. En absència de substitucions múltiples, el nombre de canvis augmenta a mesura que augmenta la divergència entre taxons; no obstant això, a partir de certs nivells de divergència el número de transicions roman constant. Aquesta situació és conseqüència de què els nous canvis s'acumulen sobre d'altres previs, de manera que no queden comptabilitzats en el nombre total. La informació sobre la freqüència relativa d'un tipus de canvi pot obtenir-se a partir del coneixement previ de com evolucionen les seqüències o bé a partir de l'observació de les característiques de les seqüències en l'estudi particular (Knight i Mindell 1993).

Convé esmentar que les transformacions entre els estats nucleotídics solen dividir-se en dos tipus (**fig. 18**). Els canvis entre bases de la mateixa naturalesa química, és a dir, entre pirimidines (**T**imina i **C**itosina) o entre purines (**A**denina i **G**uanina), s'anomenen **transicions (s)**, mentre que els canvis entre bases de naturalesa diferent (purina a pirimidina o al revés) reben el nom de **transversions (v)**. Malgrat que potencialment hi ha el doble de transversions que de transicions, s'ha observat que, en general, les transicions es donen molt més freqüentment que

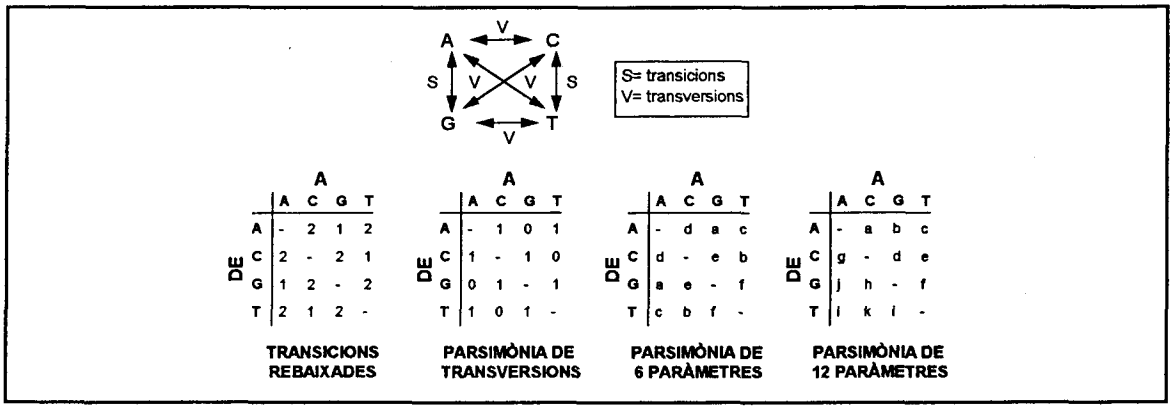


Figura 18.- Tipus de transformacions entre estats nucleotídics i exemples de matrius de costos en dades moleculars.

les transversions, especialment en DNA mitocondrial (Brown 1982, Li i col 1984, DeSalle i col. 1987), i per aquesta raó s'ha proposat que les transicions siguin

'rebaixades' i se'ls assigni un cost inferior respecta les transversions (Kocher i col. 1989) (**fig. 18**). El problema és decidir quin valor concret de cost és el correcte o, almenys, establir un criteri objectiu; amb tot, s'observa que en la majoria d'estudis la selecció d'aquests valors es força arbitrària. Una aproximació habitual es escollir uns quants esquemes de costos, amb les transicions de menys a més rebaixades, i veure quin efecte té cadascun d'ells sobre l'anàlisi. (p. ex.: 2:1, 3:1, 5:1, 10:1). Un mètode més objectiu ha estat proposat per Sturmbauer i Meyer (1992), que suggereixen utilitzar la taxa a la qual s'acumulen les transicions respecte a les transversions en comparacions entre dos taxons no afectades (les comparacions) per les substitucions múltiples. Això es tradueix a la pràctica en representar gràficament el nombre de transicions a l'eix de les Y, i el de transversions a les X, de totes les comparacions possibles de taxons dos a dos. La gràfica resultant té generalment una primera meitat de creixement conjunt de les dues variables seguit d'un estabilització del nombre de transicions; aquesta segona part correspon a l'aparició de substitucions múltiples i per tant no s'utilitza en el càlcul. El valor resultant s'obté de calcular el pendent de la zona de creixement conjunt a partir d'un model II de regressió lineal. La situació extrema de rebaixar l'impacte en l'anàlisi de les transicions està representada per l'anomenada **parsimònia de transversió** (**fig. 18**), *a on les transicions són completament eliminades* de l'anàlisi en assignar-les un cost zero. Tanmateix, aquesta aproximació pot comportar una pèrdua de resolució a les branques terminals dels arbres, ja que en molts casos les transicions son l'únic tipus de canvi que es produeix entre els taxons evolutivament més propers.

L'assumpció d'una major taxa d'acumulació de transicions enfront de transversions pot resultar, en algunes situacions, falsa. Així, s'ha demostrat per exemple que en alguns gens ribosòmics la proporció d'ambdós tipus de substitució és similar (Vawter i Brown 1993) o, com passa en alguns gens ribosomals mitocondrials, *la majoria de canvis corresponen a transversions* (DeSalle 1992). A més, en genomes amb un biaix important en la composició de bases cap a A i T, el nombre de transversions AT pot superar a qualsevol altra transformació (Dowton i Austin 1997). D'altra banda, la separació dels canvis entre bases en només dues

classes pot amagar relacions molt diferents entre els diferents tipus de canvis existents. Per tal d'evitar el possible biaix provocat per l'agrupació de canvis, es pot augmentar el nombre de classes considerades, ja sigui distingint els sis tipus diferents de transformacions (**fig. 18**) (**parsimònia de sis paràmetres**, Cunningham 1997a) o, afegint l'assimetria potencial dels canvis, establir fins a 12 classes (**fig. 18**) (**parsimònia de dotze paràmetres**, seguint la nomenclatura anterior). Han estat desenvolupades tot un seguit de metodologies per aplicar costos a les diferents transformacions. Aquestes metodologies deriven tota la informació de les dades originals i s'estalvien així assumir un model general d'evolució per a totes les molècules i els taxons (Wheeler 1990b). Entre aquests mètodes destaquen el dels **costos combinatorials** (Wheeler 1990b) (i modificacions posteriors, Rodrigo 1992) i el dels **costos EOR** (Knight i Mindell 1993). El primer es basa en la freqüència amb què els nucleòtids apareixen conjuntament a les posicions variables, de manera que com més gran sigui aquesta més probable serà que ambdós s'hagin intercanviat. Operacionalment, el mètode consisteix en assignar un valor (a_{ijk}) a cada coincidència de dos nucleòtids (*i*, *j*) en una posició derivat de la formula (d'aquí cost combinatorial):

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

a on n_k= nombre de nucleótids a la posició k i n= nombre de nucleòtids totals. Sumant els valors de totes les posicions a on apareixen conjuntament dos nucleòtids, obtenim una matriu simètrica de les associacions entre els quatre nucleòtids (A). El pas següent, és normalitzar la matriu, dividint cada valor per la suma dels valors de la columna a la qual pertany. La finalitat d'aquesta normalització és representar les freqüències de transformació des d'un nucleòtid concret fins a cada una de les tres alternatives. Aquesta normalització introdueix asimetria a la matriu, els valors dels canvis entre els nucleòtids és diferent en un sentit i en l'altre, de forma que es consideren 12 classes diferents de substitucions. Finalment, per passar de la matriu transformada d'associacions (T) a la de costos (W), s'aplica una transformació logarítmica:

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

Rodrigo (1992) ha criticat la normalització aplicada per Wheeler (1990b),

argumentant que aquesta es correspon a la pràctica amb una probabilitat condicionada. Per tal d'obtenir una autèntica probabilitat total, el denominador no ha de ser la suma de només les posicions variables, sinó la suma de totes les posicions, variables i invariables.

El mètode EOR (*expected / observed ratio*) proposa una aproximació diferent. La correcció no té en compte només la freqüència d'aparició de les dues bases, sinó també la composició nucleotídica de la seqüència. Per a cada parella de taxons es calcula el 'nombre esperat de cada tipus de canvi' multiplicant el nombre total de transicions o transversions (segons el canvi considerat) que presenten per la suma de les mitjanes de les proporcions de les dues bases implicades. El cost es deriva directament del quocient entre el valor obtingut anteriorment i el nombre de canvis 'observat'. Així, les substitucions que es donin amb freqüència del que s'espera tindran un cost menor que aquelles que es donin més rarament. Els punt febles d'aquesta aproximació són (Collins i col 1994): (1) l'absència d'un model explícit de procés a l'atzar d'on derivar el valor 'esperat' i (2) que els costos calculats per a les diferents classes tenen diferents escales i per tant la seva combinació en una mateixa matriu és incorrecta. Una crítica addicional, que pot fer-se extensiva a altres mètodes basats en comparacions entre els taxons dos a dos, és que els valors resultants d'aquestes comparacions no són independents evolutivament, de forma que els càlculs estan esbiaixats per la filogènia subjacent (Collins i col. 1994, Swofford i col. 1996). Per tal d'evitar aquest biaix es poden reconstruir els estats ancestrals a partir de l'arbre o arbres resultants d'una anàlisi preliminar. L'anomenada **ponderació dinàmica** (*dynamic weighting*) (Williams i Fitch 1989, Williams i Fitch 1990, Fitch i Ye 1991), deriva la matriu de costs a partir de la freqüència de les diferents substitucions calculades a partir d'un arbre inicial. Malgrat no ser un requeriment del mètode, aquest arbre és generalment l'arbre més parsimoniós aplicant pesos i costos uniformes. S'han proposat dos tipus de funció per transformar la matriu de freqüències en matriu de costos: lineal, fent l'invers de la freqüència, o quadràtic, fent l'invers del quadrat d'aquesta i, per tant, molt més extrem. A diferència d'altres mètodes, aquest permet incorporar també pesos diferencials als caràcters. El raonament és exactament el mateix que l'utilitzat pels

costos, amb la diferència que, enlloc d'una matriu, els 'pesos' queden incorporats a través d'un vector. La implementació d'aquesta aproximació és iterativa, de forma que després de calcular la matriu de costs i el vector de pesos, es realitza un nou anàlisi a partir del qual es calculen els nous valors de la matriu i el vector. Aquest procés es repeteix fins que l'arbre obtingut és igual a l'arbre anterior o fins que s'assoleix un cert nombre d'iteracions (20 generalment) sense l'estabilització del resultat. La principal crítica a aquest mètode i a tots els que deriven les freqüències de canvi a partir d'un arbre previ, és que els resultats son dependents de l'arbre inicial seleccionat. Un problema addicional de la ponderació dinàmica és el seu elevat cost computacional, encara que versions més simplificades d'aquesta poden dur-se a terme amb el programa MacClade 3.05 (Maddison i Maddison 1992), que permet el càlcul de les freqüències de cadascuna de les substitucions a partir d'un o més arbres subministrats per l'usuari, alhora que incorpora fins a quatre funcions de transformació de les freqüències dels canvis en costs: lineal, quadràtica i dues formes de logarítmica.

Finalment, cal esmentar que certs valors de les matrius de costos poden conduir a la violació de l'anomenada **desigualtat triangular**. Aquesta es defineix com:

$$d_{ij} \leq d_{ik} + d_{kj}$$

a on *i*, *j*, *k* representen els estats d'un caràcter, i *d* la distància o cost de la seva transformació. El no compliment d'aquesta desigualtat pot conduir a inconsistències lògiques. Per exemple, suposant que els tres estats siguin: A, C, G amb costos: AC=3, AG=1, CG=1, es comprova que 3>1+1 i que, per tant, violen la desigualtat triangular, el que implica que si en una certa posició només hi ha A i C, la reconstrucció més parsimoniosa és acceptar que l'estat ancestral era una G, i sembla il.lògic assignar a un avantpassat un caràcter no observat (Wheeler 1993). La violació de la desigualtat no és exclusiva dels caràcters moleculars, però com que és aquest el context en que més s'utilitzen matrius de costos, és més fàcil que ocorri. Hi ha tres situacions bàsiques on pot donar-se: si el cost de les transversions és superior al doble del de les transicions, si el cost dels *gaps* és molt baix (p. ex. si es tracten com a *missing data*) o si la matriu és asimètrica. Aquestes

situacions, si bé molts cops són inevitables, poden reduir el nombre de valors possibles i, per tant, seleccionar els esquemes amb més sentit filogenètic. Altres autors han argumentat que aquesta inconsistència pot tenir en certs casos una explicació biològica que la legitimi (Maddison i Maddison 1992, Fitch 1993; però veure Allard i Carpenter 1996).

Pes dels caràcters

Paral.lelament, i de manera similar a l'assignació de costos diferencials a les transformacions entre estats, els caràcters poden ser ponderats uns respecte als altres. Com ja s'ha comentat pels estats, els pes relatiu d'un caràcter enfront dels altres pot derivar-se d'una sèrie de consideracions externes i anteriors a l'anàlisi (*a priori*) o basar-se en resultats d'una anàlisi preliminar (*a posteriori*). En el cas particular de les dades morfològiques, els esquemes de ponderació *a priori* són en general de difícil justificació (Wheeler 1986, però veure Neff 1986 i Maddison 1993 per opinions contràries). Així i tot, s'han proposat mètodes de ponderació que utilitzen el criteri de compatibilitat entre els caràcters com a base per al càlcul dels seus pesos (Sharkey 1989). Entre els mètodes que deriven pesos a partir d'arbres anteriors destaca l'anomenada **ponderació successiva** (*successive weihting*) (Farris 1969; Carpenter 1988, 1994). Aquesta tècnica es basa en el concepte de **fiabilitat cladística** (*cladistic reliability*, Farris 1969), és a dir, en utilitzar el grau amb què un caràcter s'ajusta a una jerarquia com a criteri per a la seva valoració. Així doncs, abans de continuar l'exposició d'aquests mètodes de ponderació es necessari introduir els diferents índexos descrits i utilitzats a la literatura per a mesurar el grau d'ajust entre els caràcters i els arbres.

S'han descrit a la literatura diferents índexos per a reflectir l'encaix entre les dades i els arbres. Kluge i Farris (1969) van definir l'**índex de consistència (*ic*)** com el quocient entre el nombre mínim de passos (=transformacions entre estats) possibles d'un caràcter (*m*), que és igual al nombre d'estats menys u, i el nombre

de passos observats pel caràcter en un arbre determinat (*s*):

$$ic = m / s \quad , \quad IC = \Sigma m / \Sigma s$$

En el cas que l'ajust sigui perfecte, el nombre de passos observat serà igual al mínim teòric i el valor de l'índex serà 1. Com pitjor sigui l'encaix entre el caràcter i l'arbre, més passos addicionals seran necessaris per fer-lo ajustar, i el valor de l'índex baixarà, tendint a 0 a l'infinit. L'índex de consistència conjunt d'un arbre (*IC*) és el quocient entre la suma dels valors particulars de *m* i de *s* de cada caràcter. Degut a què les situacions a on el nombre de passos 'reals' excedeix el nombre mínim teòric s'expliquen per la presència d'homoplàsia, l'*ic* ha estat considerat com a una mesura d'aquesta. Amb tot, la influència de certes variables sobre l'*IC* pot comprometre el seu valor comparatiu entre diferents anàlisis. Així, l'*IC* està correlacionat amb el nombre de taxons, de forma que matrius amb més taxons tendeixen a tenir *IC* més baixos (Sanderson i Donoghue 1989); s'ha obtingut, però, una formula empírica per calcular l'*IC* esperat donat un cert nombre de taxons (*n*):

$IC_{esperat} = 0{,}90 - 0{,}022n + 0{,}000213n^2$, per $n<60$ (Sanderson i Donoghue 1989)

L'*IC* també resulta afectat per la proporció de caràcters ordenats en una matriu. Els caràcters ordenats, en general, encaixen pitjor en la definició de les agrupacions que no pas els desordenats perquè són més restrictius. La presència de *missing data* és una altra font de distorsió, ja que aquestes no incrementen mai la inconsistència dels caràcters. Finalment, els caràcters autapomòrfics o invariables inflen de forma artificial l'*IC*, perquè no suporten cap agrupació i, per tant, sempre encaixen perfectament. Si bé quan es treballa amb dades morfològiques no es solen introduir aquests tipus de caràcters a la matriu, això es inevitable en treballar amb dades moleculars, i cal eliminar aquests caràcters abans de calcular el valor de l'*IC*. En un altre ordre de coses, cal esmentar que Klassen i col. (1991) han demostrat que per a què una matriu de dades sigui filogenèticament informativa, els seu *IC* ha de ser major que l'*IC* d'una matriu aleatoritzada de la mateixa mida (n). Aquest valor pot obtenir-se de la formula:

$$IC_{random} = 2.937 \times n^{-0{,}9339}$$

L'anomenat **índex de retenció (*ir*)** (Farris 1989) va ser proposat per tal d'evitar alguns dels problemes associats a l'*IC* i expressa la proporció de similaritats a

l'arbre que són explicables com a sinapomorfies. En aquest cas, l'aproximació es realitza a través de l'avaluació de la quantitat de sinapomorfies en la matriu. Així, l'*ir* és el complementari del valor d'homoplàsia relativa present a la matriu respecte a la màxima possible. La formula pel seu càlcul és:

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

a on *g* és el nombre màxim de transformacions d'estats possibles d'un caràcter en qualsevol arbre (a la pràctica, el nombre de passos donada una politomia total, Wiley i col. 1991). Al valor *g - m* se'l denomina **variabilitat informativa** (Farris 1991). S'observa que quan *s* és igual a *m*, és a dir, en absència d'homoplàsia, l'*ir* és 1. Com més gran sigui l'homoplàsia del caràcter, menor serà el seu valor d'*ir*. L'índex de retenció de l'arbre (*IR*) es la suma dels valors de *g*, *s* i *m* per a cada caràcter. Valors alts de l'*IR* indiquen que els canvis s'acumulen preferentment a les branques internes, mentre que valors baixos evidencien que els canvis s'acumulen a les branques terminals (Siebert 1992). L'*IR* és doncs una bona mesura del suport basat en l'evidència dels grups, i a més, no és sensible a les autapomorfies ni als caràcters invariables, ja que en aquests *g* és igual a *s* i, per tant, l'*ir* és 0. Finalment, l'**índex reescalat de consistència (*CR*)** (Farris 1989) es defineix com el producte dels dos índexos anteriors. La seva implementació permet discriminar diferències en el valor de *g* per a caràcters que altrament mostren nivells similars d'homoplàsia (Quicke 1993).

La ponderació successiva assigna un pes diferencial als caràcters en funció del seu ajust a l'arbre resultant de l'anàlisi conjunta de tots els caràcters uniformement ponderats. Els caràcters de la matriu es jutgen a ells mateixos en base a la seva fiabilitat (Carpenter 1994). Originalment va ser descrit com a criteri addicional per a l'elecció d'un arbre en situacions on s'obtenien diferents arbres igualment parsimoniosos (Carpenter 1988). Això no obstant, s'ha constatat de què el resultat d'aquesta ponderació podria ser un arbre diferent en nombre de passos i en topologia a l'original (Platnick i col. 1991a, Brothers i Carpenter 1993). Molts autors han proposat que l'arbre obtingut mitjançant una ponderació adequada dels caràcters ha de ser considerat com a la millor hipòtesi de la filogènia,

independentment de què sigui menys parsimoniós sota la ponderació uniforme (Kluge i Farris 1969, Farris 1969, Platnick i col 1991a, Goloboff 1993; però veure Scharff i Coddington 1997). La ponderació successiva és un mètode iteratiu, de manera que a l'obtenció d'un arbre, o arbres, a partir de la ponderació inicial dels caràcters, segueix una nova assignació de pesos a partir d'aquests. El procés es repeteix fins que en dos anàlisis successius s'obtenen els mateixos arbres, amb el mateix nombre de passos i valors dels índexos. Aquest és el criteri anomenat d'**autoconsistència** (Farris 1969). Un arbre és autoconsistent si és él més curt sota els pesos dels caràcters que ell mateix implica, i és l'arbre que resol els conflictes entre els caràcters a favor d'aquells que són menys homoplàsics. L'aplicació pràctica d'aquesta tècnica pot comportar alguns problemes. La primera qüestió és l'elecció de l'índex a utilitzar per a mesurar l'ajust dels caràcters. Actualment, el més àmpliament emprat és el *CR*, que passa per ser el menys sensible a l'existència de diferents tipus de caràcters (diferent nombre d'estats, ordenats o no, etc.) a les matrius, alhora que pot assolir valors de 0 (cosa que no pot fer l'*IC*) i eliminar per tant completament un caràcter de l'anàlisi. Així i tot, alguns autors han proposat la utilització de l'*IC*, ja que el *CR* (a l'igual que l'*IR*) dóna un pes menor no només als caràcters més homoplàsics, sinó també a aquells amb una major variabilitat informativa (= *g* - *m*) (Goloboff 1991, 1993). Un problema addicional es planteja quan existeixen dos o més arbres igualment parsimoniosos per a les dades uniformement ponderades. En aquests casos pot escollir-se entre el valor més alt possible de l'índex del caràcter, la mitjana entre els valors de l'índex a cada arbre i el valor més baix. Un altre factor important té a veure amb la implementació computacional de la tècnica, en especial amb l'escala dels valors dels pesos. En el programa Hennig86 (Farris 1988), que només permet utilitzar el *CR*, s'utilitza el valor més alt de l'índex en cas de arbres múltiples, i els pesos assignats varien entre 0 i 10. A més, els valors decimals queden truncats i s'accepta només la part entera. NONA/PHAST/SPA permeten utilitzar qualsevol índex, sempre i quan aquest sigui implementat a través d'una petita rutina especificada per l'usuari, a partir d'un llenguatge de programació subministrat amb els programes. Per defecte els programes només permeten utilitzar l'*IC* com a funció d'assignació de pesos, el

millor ajust, i l'escala va de 0 a 100. Finalment, PAUP/PAUP* permeten escollir entre els tres tipus de funció (*IC*, *IR*, *CR*) i tots els possibles valors dels índexos donats alguns arbres. L'escala va de 0 a 1000, i ofereix la possibilitat d'arrodonir els valors enlloc de truncar-los. Els avantatges relatius d'una implementació envers les altres, així com l'efecte que poden tenir sobre l'elecció de l'arbre resultant, no han estat explícitament investigades. Si bé una escala d'1 a 1000 permet obviament una major discriminació, no queda clar si aquesta és realment desitjable degut al caràcter aproximat d'aquesta estimació dels pesos. D'altra banda, Scharff i Coddington (1997) han alertat de la possibilitat de què, en la ponderació successiva, caràcters complexos perfectament definibles i objectius rebin pesos baixos, mentre que caràcters amb interrogants o més pobrement definits es vegin potenciats. Aquests autors advoquen per avaluar críticament les implicacions per a l'evolució dels caràcters derivades de l'arbre ponderat resultant.

Goloboff (1993, 1995a), malgrat estar d'acord amb el criteri d'assignació de pesos utilitzat per a la ponderació successiva, argumenta en contra de la seva implementació iterativa. Així considera que (1) la solució final depèn del conjunt inicial de pesos, (2) pot donar-se el cas de què algun dels arbres inicials no sigui autoconsistent i (3) com que l'avaluació d'autoconsistència es realitza a partir d'una submostra d'arbres possibles, aquest mètode no permet trobar tots els arbres autoconsistents. Per tal d'evitar aquestes i altres limitacions de la ponderació successiva, es proposa una nova aproximació basada en el càlcul del pes dels caràcters simultàniament a la cerca dels arbres (Goloboff 1993, 1995a). Aquesta fita s'assoleix substituint el criteri d'optimització de la parsimònia clàssica, és a dir, la minimització del nombre de passos totals de l'arbre, per un de nou: la maximització d'una funció conjunta, còncava i decreixent de l'homoplàsia (*F*):

$$F = \sum f_i \text{ a on } f_i = k / (k + es_i + es_o)$$

L'**ajust** (*fit*) d'un caràcter (f_i) és funció còncava del nombre de passos addicionals ($es_i = s - m$) que mostra en un arbre determinat i del nombre de passos addicionals (es_o) deguts a polimorfismes en els taxons terminals. La constant de concavitat (k) determina el grau en al què els caràcters homoplàsics son rebaixats. Valors baixos de k provoquen l'augment de la concavitat de la funció, el que es tradueix en què

els caràcters homoplàsics tinguin poca influència. Valors alts de *k* linealitzen la funció (la relació entre el pes i l'homoplàsia a la parsimònia tradicional és lineal), aproximant la influència dels caràcters homoplàsics a la de la resta. Tanmateix, la selecció del valor de concavitat idoni per a una determinada anàlisi roman sense ser investigada i, de moment, és totalment subjectiva. L'assumpció principal d'aquest mètode és que els arbres que maximitzen aquesta funció de l'homoplàsia són els que maximitzen la fiabilitat dels caràcters. Aquesta aproximació està implementada en el programa d'ordenador PEE-WEE (*Parsimony and Implied Weights*) (v 2.50, Goloboff 1996c). A més de l'elecció del valor de *k*, hi ha altres problemes en la utilització d'aquest criteri. Per exemple, hom pot preguntar-se el sentit, des d'un punt de vista filogenètic, d'una diferència d'ajust de dècimes entre dos arbres (Coddington com. personal). En aquest context, convé esmentar que el propi autor ha implementat en el programa la capacitat per retenir arbres subòptims. Amb tot, la qüestió és més complexa, ja que no només cal examinar el valor de la reducció en ajust, sinó també les causes d'aquesta reducció, és a dir, els tipus de caràcters implicats en la mateixa (Goloboff 1995a). D'altra banda, quan s'utilitzen dades moleculars, la diferència entre el nombre de passos mínim i el nombre de passos màxim d'un caràcter pot ser molt gran, i provocar una reducció massa dràstica de l'ajust del caràcter. Per tal d'evitar aquest efecte s'ha proposat corregir la funció d'homoplàsia dividint el seu denominador per *m* (Gladstein i Wheeler en prep.).

Com ha estat esmentat en parlar de costos de les transformacions, certes peculiaritats de les dades moleculars, com ara els nombre fix d'estats compartits per tots els caràcters, així com el coneixement previ d'un seguit de processos que afecten l'evolució a nivell molecular, ofereixen un conjunt de criteris per a la ponderació, en aquest cas, dels caràcters. L'assumpció més generalitzada és que les posicions amb una taxa elevada de canvi han sofert, molt probablement, substitucions múltiples (=homoplàsia) i que, per tal d'evitar que tinguin una influència negativa sobre l'anàlisi, el seu pes ha de ser rebaixat (Hillis i col. 1993). En el cas de gens codificadors de proteïnes i degut a la degeneració del codi genètic, la freqüència de canvis a la tercera base dels codons acostuma a superar

la de la primera i segona base nucleotídiques; és una pràctica força generalitzada el donar un pes superior a les primeres i segones posicions. El valor concret del pes sol derivar-se de l'invers de la freqüència relativa de canvis, ja sigui per a cada posició o reunint la primera i segona en una sola classe en front de la tercera. Alhora, la quantitat de canvis per cada classe de posicions pot determinar-se mitjançant comparacions de taxons dos a dos o, per evitar la dependència dels caràcters, derivar-la de la reconstrucció dels estats en els avantpassats a partir d'un arbre preliminar (Williams i Fitch 1989, Maddison i Maddison 1992). Tanmateix, aquesta aproximació resulta una simplificació, ja que en certs casos un canvi a la tercera posició d'un codó pot comportar un canvi aminoacídic i, per contra, hi ha situacions en què canvis a la primera base no suposen necessàriament la substitució d'un aminoàcid. L'aplicació de pesos més 'realistes' esdevindria massa complexa (Simon i col. 1994). Quan la divergència entre els taxons és suficientment gran, existeix la possibilitat de traduir la seqüència nucleotídica en aminoàcids i realitzar l'anàlisi utilitzant-los com a caràcters. Recentment, Agosti i col. (1996) han proposat una opció intermitja per a la ponderació de posicions de gens codificants que consisteix en combinar els caràcters nucleotídics i aminoacídics en un sola matriu. D'aquesta manera s'aconsegueix indirectament el reforç de les configuracions derivades de les bases nucleotídiques que siguin congruents amb el canvis d'aminoàcids, en principi més conservatius. El major problema és que es suposa que els dos tipus de caràcters no són independents perquè es corresponen a dues formes de codificació de la mateixa informació, i, per tant, es viola un dels requisits per a l'aplicació de la parsimònia. Amb tot, aquesta autors argumenten que en la pràctica aquests caràcters es comporten com a independents i per tant l'aplicació del criteri de parsimònia és segur.

A diferència dels gens codificants, els gens ribosomals són fonamentalment estructurals i l'RNA transcrit a partir d'ells sofreix un procés de plegament que determina la seva estructura funcional. Aquesta és bàsicament la conseqüència de l'existència de regions de la seqüència que són complementàries d'altres a les quals s'uneixen. El resultat és una estructura espaial amb regions de cadena doble, els ***stems***, alternades amb regions de cadena senzilla, els ***loops***. La distinció de les

posicions ribosomals en dues classes és la base per a la ponderació diferencial aplicada a aquests gens. Així, d'una banda s'ha assenyalat que, degut al seu paper en el manteniment de l'estructura secundària, les posicions en els *stems* són més conservatives que les dels *loops* i, per tant caldria augmentar el seu pes en l'anàlisi. Estudis sobre les taxes de canvi de diferents regions dels gens ribosomals semblen però no recolzar aquesta afirmació (Simon 1991, Vawter i Brown 1993). Contràriament, s'ha suggerit que el pes de les posicions dels *stems* hauria de ser rebaixat a l'anàlisi perquè una mutació en una zona de l'*stem* requeriria una substitució compensatòria a la complementària per tal de mantenir l'estructura. Wheeler i Honeycutt (1988) consideren que, degut a la no independència de les posicions, els *stems* haurien de ser rebaixats en un factor de 1/2 respecte als *loops*. Simon (1991) considera que aquest valor és resultat d'una simplificació del procés ja que, degut a l'existència d'aparellaments no clàssics (p. ex. G-T o T-C), les compensacions poden no ser perfectes. Finalment, Dixon i Hillis (1993), a partir de dades empíriques, proposen no rebaixar els *stems* més d'un 80% respecte als *loops*.

Totes les aproximacions a l'assignació ja sigui de pesos o de costos en dades moleculars, pateixen la limitació de tractar els estats i les posicions com a classes (Carpenter 1994), cadascuna de les quals rep un valor de cost o pes. Aquesta uniformització dels estats i les posicions té l'inconvenient d'amagar les possibles diferències que existeixin entre els seus components. Així, per exemple, quan es divideixen les substitucions en transicions i tranversions i s'assigna un únic cost a cadascuna d'elles, s'està amagant el fet prou conegut de què la proporció de transicions respecte a transversions té un rang de variació molt ampli i més o menys relacionat amb el temps de divergència. Un altre desavantatge de la ponderació és que els estadísitics (=índexos) dels arbres obtinguts a partir de la incorporació d'un cert esquema de ponderació, tant en dades moleculars com en morfològiques, no poden ser comparats amb d'altres derivats d'esquemes diferents (Simon i col. 1994).

3.3.4.4. Tècniques per la construcció de l'arbre i l'optimització dels caràcters

Un cop la matriu de caràcters ha estat construïda i les assumpcions sobre els costos de les transformacions i/o el pes relatiu dels diferents caràcters han estat incorporades, el següent pas és trobar l'arbre més parsimoniós. La parsimònia, al igual que altres mètodes de reconstrucció filogenètica com la màxima versemblança o els mètodes de distàncies amb arbres additius, estableix un **criteri d'optimalitat**, és a dir, un criteri per, donat un conjunt d'arbres, decidir quin és el millor. En el cas de la parsimònia el criteri és: 'el millor arbre és aquell amb el menor nombre de **passos (*steps*)** (=transformacions, canvis, substitucions) possible'. Hom es refereix al nombre total de passos d'un arbre com a la **llargada de l'arbre (*tree length*)**. Quan s'utilitza un mètode que incorpora un criteri d'optimalitat, la problemàtica de trobar el millor arbre pot separar-se en dos aspectes: construir els arbres i avaluar-los (Swofford i col. 1996). L'avaluació d'un arbre consisteix en determinar primer la seva llargada (nombre de passos totals) i comparar-la aleshores amb la d'altres arbres. La llargada d'un arbre és la suma del nombre de passos de cada caràcter de la matriu multiplicat pel pes que li ha estat assignat. Així, donats dos caràcters, el primer amb 2 passos i un pes de 3, i el segon amb 4 passos i un pes de 1, la llargada total seria: (2 x 3) + (4 x 1)= 10. El càlcul del nombre de passos d'un caràcter en un arbre determinat, resulta de l'optimització del caràcter en aquest. S'anomena **optimització d'un caràcter** a l'assignació d'un estat concret del caràcter a cadascun dels nodes interns (avantpassats hipotètics) d'un arbre determinat, de manera que el nombre de transformacions necessàries per a explicar la distribució d'estats a l'arbre sigui el mínim possible. En certs casos l'optimització pot ser ambigua, és a dir, que l'assignació d'un estat o un altre a un determinat node és igualment parsimoniosa. Per tal d'escollir un dels estats possibles poden utilitzar-se una sèrie de criteris auxiliars. Així, per exemple, hom pot preferir que l'homoplàsia del cladograma sigui principalment explicable per reversions, el que s'aconsegueix afavorint que els canvis d'estat es produeixin cap a la base de l'arbre. A aquest

tipus d'optimització se l'anomena ACCTRAN (Swofford i Maddison 1987, Swofford 1990). Per contra, hom pot escollir que les convergències siguin la causa fonamental de l'homoplàsia i s'afavoreix l'optimització que concentra els canvis preferentment cap a les puntes del cladograma. Aquesta optimització rep el nom de DELTRAN (Swofford i Maddison 1987, Swofford 1990). L'optimització depèn alhora de les assumpcions sobre les relacions entre els estats que s'hagin incorporat, per exemple, l'optimització d'un caràcter serà diferent segons s'hagin considerat els seus estats com a ordenats o desordenats. Com ha estat explicat en l'apartat de costos de les transformacions, existeixen diferents algorismes computacionals que permeten calcular el nombre de passos d'un caràcter considerant certes assumpcions sobre les relacions dels seus estats. D'altra banda, cal tenir en compte que com que la llargada d'un arbre és la suma ponderada dels passos de cada caràcter, dos arbres poden tenir la mateixa llargada i ser topològicament diferents. Això s'explica perquè una topologia pot estalviar passos en certs caràcters i augmentar-los en d'altres, de manera que la suma final sigui la mateixa que la d'una altra topologia que ha estalviat o augmentat passos en caràcters diferents. La conseqüència d'això és l'obtenció de diferents arbres **igualment parsimoniosos**.

3.3.4.4.1. Cerques exhaustives

Sota la perspectiva plantejada, la cerca de l'arbre o arbres més parsimoniosos (=més curts) pot semblar força trivial. Només es necessita implementar un algorisme computacional que construeixi tots els arbres possibles amb el nombre de taxons estudiats, calcular la llargada de cadascun d'ells i seleccionar els més curts. Aquest tipus d'aproximació s'anomena **cerca exhaustiva** (per descripció de l'algorisme veure Swofford i col. 1996). Tanmateix, el problema no és, ni de bon tros, tan simple. La principal limitació d'aquesta aproximació és que el nombre d'arbres possibles augmenta de foma 'explosiva' amb el nombre de nodes terminals (taxons). Considerant un nombre *n* de taxons, el nombre d'arbres no arrelats

possibles ve donat per la formula:

$$\Pi\ (2_i - 1), \quad i = 1, 2, ..., n\text{-}2$$

Així, per a 7 taxons hi ha 945 arbres dicotòmics i sense arrel possibles, però si el nombre de taxons augmenta a 20, el nombre d'arbres possibles és de $2{,}2 \times 10^{20}$. La cerca exhaustiva només és factible si la matriu té 11 taxons o menys (34.459.425 arbres) (Swofford i col. 1996). Afortunadament, si el valor òptim d'un mètode no es mai inferior quan s'afegeix un nou element al càlcul, com en parsimònia, existeixen algorismes que permeten trobar el millor arbre o arbres sense d'haver d'avaluar tots el possibles (Swofford i col. 1996). Aquests algorismes es basen en el mètode del ***branch-and-bound***, aplicat per primera vegada al problema dels arbres filogenètics per Hendy i Penny (1982) i posteriorment desenvolupat i refinat per diversos autors (per descripció de l'algorisme veure Swofford i col. 1996). El *branch-and-bound* permet cerques amb un nombre major de taxons i amb menor temps computacional. Malgrat tot, en situacions on és factible, és preferible optar per utilitzar una cerca exhaustiva perquè d'aquesta manera s'obté informació sobre la distribució de la quantitat d'arbres existents per a cada nombre de passos, la qual cosa permet alhora avaluar arbres subòptims (Kitching 1992). Malauradament, la utilització del *branch and bound* té també un límit superior del voltant de 20-25 taxons, depenent del nombre i grau d'homoplàsia dels caràcters.

3.3.4.4.2. Cerques heurístiques

Des del punt de vista matemàtic, la cerca de l'arbre més curt donat un cert nombre de taxons i caràcters és un problema 'NPcomplet' (NP= polinomial no determinístic). Això vol dir que no és probable que existeixi un algorisme general que garanteixi solucionar totes les vessants del problema, de manera que el temps computacional vingui donat per una funció polinòmica de la mida del problema (Penny i col. 1992). Per tant, en absència d'aquestes algorismes, que reben el nom d'**eficients**, el temps invertit pels algorismes exactes per solucionar problemes NP-complets augmenta

extremadament ràpid a mesura que augmenta la complexitat de les dades. L'única aproximació possible en aquests casos és la utilització de mètodes de 'prova i error', que consisteixen en dos passos bàsics (1) construcció d'un arbre inicial i càlcul del seu nombre de passos i (2) modificació de l'arbre reordenant les branques (canviant la disposició dels taxons) per tal de trobar topologies més curtes (Maddison 1991). Aquests tipus de cerca s'anomenen **cerques heurístiques**. En general, l'obtenció de l'arbre inicial es realitza mitjançant un protocol d'**incorporació seqüencial** dels taxons (***stepwise addition***), així, es construeix un arbre preliminar seleccionant tres taxons i posteriorment es van afegint els taxons restants un a un. Segons el criteri utilitzat per a la selecció dels taxons inicials i la seqüència posterior d'incorporació, podem definir diverses variants del protocol:

1. *As is*. Els tres taxons inicials són els primers de la matriu i la incorporació dels taxons posteriors segueix estrictament l'ordre d'aquesta.
2. *Random*. Els taxons inicials i l'ordre d'incorporació es determinen a l'atzar.
3. *Simple* (Farris 1970). Es selecciona un taxó, generalment el definit com a *outgroup*, i es calcula la distància de *Manhattan* de tots els taxons restants respecte al seleccionat. Els dos amb la distància més curta formaran, juntament amb el primer, l'arbre preliminar. L'ordre d'incorporació dels taxons restants es realitza en funció de la seva distància al primer (de menor a major). En el cas que dos o més taxons tinguin el mateix valor, es selecciona un d'ells a l'atzar.
4. *Closest*. S'examinen tots els arbres possibles de tres taxons i es selecciona el més curt. Posteriorment, s'afegeix cadascun dels taxons restants a totes les branques possibles de l'arbre i es reté el taxó i la disposició que fan l'arbre més curt. Això es repeteix fins que s'han incorporat tots els taxons. Aquest és el mètode més costós computacionalment.

Un cop tenim l'arbre o arbres inicials, el següent pas consisteix en **intercanviar l'ordre de les seves branques** (***branch swapping***). S'han descrit diferents tipus de reordenacions que es diferencien per la forma en què s'intercanvien les branques (Swofford i col. 1996). Les més utilitzades pels programes d'inferència filogenètica són (**fig. 19**):

1. *Nearest neighbour interchange (NNI)*. Es selecciona una branca interna i s'intercanvien entre ells els dos grups que queden a banda i banda, per a cada branca hi ha dos canvis possibles. Això es porta a terme per totes les branques internes.
2. *Subtree pruning and regrafting (SPR)*. Es trenca una branca interna de manera que s'obtinguin dos subarbres, un amb una branca lliure i l'altre sense. Posteriorment, el subarbre amb la branca lliure s'uneix, a través d'aquesta, a les diferents branques de l'altre subarbre. S'avaluen totes les possibles combinacions de trencament de branques i connexió de subarbres.
3. *Tree bisection and reconnection (TBR)*. Similar a l'anterior però, en aquest cas, cap dels dos subarbres te una branca lliure. Posteriorment, es connecten les branques de cadascun dels subarbres amb una nova branca. S'avaluan totes les combinacions possibles de trencaments i reconnexions.

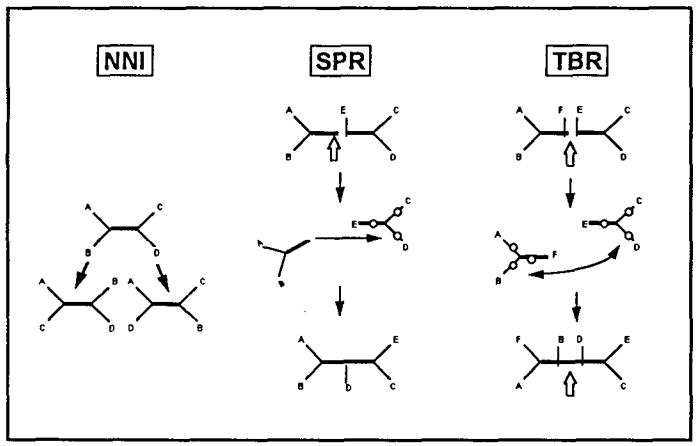


Figura 19:- Esquemes dels diferents mètodes d'intercanvi entre branques (modificat de Page 1993).

La relació entre el conjunt d'arbres que pot generar cada tipus de reordenació és NNI ⊂ SPR ⊂ TBR (Maddison 1991). Es comprova doncs que les reordenacions TBR són les més exhaustives, malgrat que són també les computacionalment més costoses.

Cal tenir en compte, que les cerques heurístiques no garanteixen obtenir l'arbre o arbres més curts, sinó que donats uns arbres inicials poden trobar tots els arbres més curts derivats de les seves reordenacions; el que es coneix amb el nom

de **mínim local**. Si algun dels mínims locals trobats coincideix amb el mínim global, aleshores haurem trobat, almenys, algun dels arbres més parsimoniosos. Tanmateix, mai no es pot tenir la certesa de què això sigui així. Un problema afegit és la possibilitat de què els arbres més parsimoniosos formin part de diferents illes. Una **illa** es defineix com a una col.lecció d'arbres parsimoniosos 'connectats' i separats d'altres grups d'arbres igualment parsimoniosos per arbres intermitjos més llargs (Maddison 1991). La connexió què es fa referència és el fet que aquests arbres estan separats els uns dels altres per un sol intercanvi de branques. Evidentment, el nombre i la topologia dels arbres d'una illa estaran determinats pel tipus d'intercanvi definit (NNI, SPR o TBR). Si la cerca utilitza només un arbre inicial, només serà possible trobar una illa, ja que les illes estan separades per arbres més llargs i en els reordenaments només es retenen els arbres més curts. Maddison (1991) va demostrar empíricament que si els arbres obtinguts d'una anàlisi tenen un índex de consistència inferior a 0,67 i el nombre de taxons és molt més gran que 20, molt probablement existeix més d'una illa. Una manera de trobar diferents illes és utilitzar més d'un arbre inicial; la gran majoria de programes permeten aquests tipus de cerques. Una de les més eficients és generar molts arbres inicials utilitzant el protocol d'incorporació de taxons *random* (p. ex. 100 arbres inicials) i aplicar posteriorment intercanvis de tipus TBR a cadascun d'ells. Com que, en principi, cada arbre serà diferent als altres, la convergència dels diferents resultats cap a un conjunt d'arbres comuns confereix una major confiança a l'anàlisi. Una altra manera de moure's 'entre illes' és utilitzar arbres que en principi siguin més llargs, amb l'esperança de què les seves reordenacions ens permetin trobar arbres més parsimoniosos addicionals. Finalment, una altra opció és utilitzar intercanvis de branques més complexes per tal d'augmentar la mida de la mostra d'arbres investigats. Així, el programa PEE-WEE i afins (Goloboff 1996a,b,c,d) incorporen algorismes d'intercanvis múltiples entre branques. Tanmateix, aquests tipus d'intercanviadors de branques encara estan en període experimental i no hi ha cap referència sobre la seva eficàcia.

3.3.4.5. Fiabilitat dels arbres i els clades obtinguts

Donat un mètode d'inferència filogenètica amb un criteri explícit d'optimalitat, la millor hipòtesi de les relacions filogenètiques entre els taxons estudiats és aquell arbre que maximitza el criteri esmentat. En el cas concret de la parsimònia, l'arbre o arbres que representen la millor hipòtesi de la filogènia donada una matriu de caràcters són aquells amb el menor nombre de passos possibles. Tanmateix, hi ha tot un seguit de situacions que poden portar a posar en dubte que l'arbre obtingut, en el seu conjunt o en alguns dels clades que conté, reflecteixi realment les relacions filogenètiques dels organismes o, almenys, que ho faci ostensiblement millor que altres arbres. En aquest context, hom pot interrogar-se sobre el grau de confiança o fiabilitat que li mereix la interpretació de l'arbre resultant com a hipòtesi de filogènia. En realitat aquest grau de confiança o fiabilitat pot separar-se en dos aspectes: (1) té l'arbre en el seu conjunt un senyal o estructura filogenètica significatius? i (2) quin és el grau de suport o d'estabilitat dels diferents clades continguts en aquest arbre? Abans de continuar, cal recordar que l'aplicació de certs conceptes estadístics, com els intervals de confiança o la significació d'un determinat test, es basen en l'existència d'una distribució de probabilitat coneguda, o almenys assumida, de les dades. En inferència filogenètica, aquesta distribució de les dades pot ser coneguda només en cas que s'accepti un determinat model evolutiu, responsable de la generació d'aquestes distribucions. La parsimònia es caracteritza, enfront d'altres mètodes de reconstrucció filogenètica (màxima versemblança, mètodes de distàncies), per la no incorporació explícita en l'anàlisi d'un determinat model evolutiu. Malgrat tot, existeixen una sèrie de mètodes matemàtics que permeten el càlcul de certs estadístics en absència, o desconeixement, de la distribució de probabilitat de les dades. Aquestes metodologies s'anomenen en conjunt **tècniques de remostreig**, ja que mitjançant mostreig repetit de les dades originals, s'obté un conjunt de rèpliques d'aquestes anomenades pseudorèpliques, a partir de les quals s'elabora la distribució de l'estadístic d'interés. Aquestes mostres s'elaboren a partir de l'**aleatorització**

(***randomization***) de les dades originals. En el cas concret de les matrius de caràcters, aquest remostreig pot realitzar-se a nivell de columnes (caràcters) o a nivell de files (taxons). Les tècniques d'aleatorització computacional intensiva poden dividir-se en tres tipus bàsics: ***jackknife***, ***bootstrap*** (un subtipus dels anomenats mètodes de Monte Carlo) i **permutacions** (Siddall 1995). Cadascuna d'elles difereix de les altres per la manera en què les dades originals són aleatoritzades (**fig. 20A-C**). Així, en *jackknife* les pseudorèpliques s'obtenen per eliminació d'una (de primer ordre) o d'un nombre *n* (d'ordre *n*) de les dades inicials. En el *bootstrap* les pseudorèpliques s'obtenen seleccionant a l'atzar i amb reemplaçament les dades inicials, fins a obtenir el mateix nombre de dades que en la mostra original; en la pseudorèplica, algunes dades originals estaran repetides mentre que d'altres seran absents. Finalment, en les permutacions les dades són simplement reordenades i es trenca d'aquesta manera qualsevol possible covariació entre elles.

3.3.4.5.1. Senyal, estructura o informació filogenètica d'un arbre

Com que tots els mètodes de reconstrucció filogenètica assumeixen implícitament que existeix una relació entre els taxons susceptible de ser representada en forma d'arbre, fins i tot en el cas en què els estats haguessin estat assignats amb un dau, qualsevol d'aquests mètodes generaria un o més arbres. La parsimònia maximitza la congruència entre els caràcters, però no discrimina les fonts de congruència. Així, en absència d'un procés unificador (=evolució), aquesta congruència pot produir-se per simple atzar. La pregunta és doncs si la matriu de dades mostra una congruència entre els caràcters major a l'esperada per l'atzar. Hi ha dues tècniques principals per intentar respondre a aquesta qüestió:

1. *Skewness*. La distribució de la llargada dels arbres possibles per a una determinada matriu (nombre d'arbres de cada llargada) està fortament esbiaixada cap als arbres amb més pasos. L'estadístic g_1 (Sokal i Rohlf 1981) mesura el grau de biaix (*skewness*) d'una distribució, i és negatiu quan la distribució està esbiaixada cap a l'esquerra (=menys arbres curts

que llargs). Com més gran sigui la congruència dels caràcters, sigui quina sigui la raó, més esbiaixada serà la distribució, ja que només un petit grup d'arbres serà molt més curt que la resta. Huelsenbeck (1991) i Hillis (1991) han proposat d'utilitzar l'índex g_1 com a mesura del grau d'estructura filogenètica d'una matriu: com més negatiu sigui aquest valor més gran serà el senyal filogenètic d'aquesta. A més, a partir dels valors de g_1 obtinguts per a matrius generades a l'atzar, es pot definir un test de significació per veure si el valor de la matriu original se separa significativament del valor per a matrius sense senyal filogenètica, és a dir, d'aquells on la congruència entre els caràcters es deu només a l'atzar (Hillis i Hulsenbeck 1992). Amb tot, Källersjö i col. (1992) han demostrat que l'índex g_1 por estar més influenciat per la freqüència dels estats dins dels caràcters que per la congruència entre aquests, a més de ser insensible al nombre de caràcters. Altres crítiques són que els valors de g_1 no són prou coneguts com per a establir acuradament els llindars de significació de l'estadístic i, d'altra banda, la qüestió de si una diferència significativa respecte del valor generat a l'atzar implica necessàriament una estructura jeràrquica significativa (Trueman 1993).

2. *Test AFC (*segons Källersjö i col 1992, = *PTP, permutation tail probability).* Seguint un raonament similar a l'anterior, la congruència entre els caràcters en una matriu amb senyal filogenètica ha de ser més gran que la simplement produïda per l'atzar. Archie (1989) i, independentment, Faith i Cranston (1991) van formular un test (Archie-Faith-Cranston) de significació basat en una aproximació diferent: la permutació dels estats de cadascun dels caràcters de la matriu original permet d'obtenir matrius amb les mateixes característiques que l'original (igual nombre de caràcters, igual nombre d'estats i *missing data* per cada caràcter), però amb congruència entre els caràcters deguda només a l'atzar. Com que el grau de congruència entre els caràcters és el responsable últim de què uns arbres siguin més curts que uns altres, es pren la llargada de l'arbre més parsimoniós com a mesura de la congruència. La distribució de probabilitats (de fet freqüències) s'estableix a partir del càlcul de l'arbre més parsimoniós per a un nombre *N*

de matrius permutades (=pseudorèpliques) (**fig. 20C**). La matriu original serà significativament diferent de les permutades, es a dir, tindrà senyal filogenètica, si la probabilitat de què la llargada de l'arbre més parsimoniós per les dades originals cau dins de la cua inferior de la distribució de probabilitats. Formalment (Farris 1991):

$\alpha' = 1 - E / (N + 1)$, on $\alpha' <$ nivell de significació escollit (p. ex. 0,05)

a on α' és la taxa d'error i *E* és el nombre de casos (observats i permutats) amb llargades tan o més curtes que la llargada observada. S'han observat situacions a on s'ha assolit significació, al nivell del 0,05, en només 20 permutacions (Trueman 1993). És important recordar que aquest test només permet respondre la pregunta de si la congruència de la matriu es més gran que l'esperada per l'atzar, però no si aquesta congruència és deguda a l'evolució (filogenètica) o a qualsevol altre procés subjacent (p.ex. adaptació ecològica). Si bé aquest test ofereix certs avantatges sobre el *skewness* (no sensibilitat a les freqüències dels estats, influència del nombre de caràcters), s'ha observat que pot trobar una alta significació a matrius amb una estructura filogenètica ambigua (Källersjö i col 1992) o, fins i tot, ser significatiu quan aquesta simplement no existeix (Alroy 1994). Per tal d'evitar aquest efecte, Källersjö i col (1992) han proposat d'utilitzar un altre índex: el suport total (TS), que resulta de sumar els suports de Bremer (veure l'apartat següent) de tots els clades de l'arbre considerat. Trueman (1996), malgrat estar d'acord amb la utilització del test en la seva formulació original proposa de no incloure els *outgroups* en el test, ja que d'aquesta manera s'evita un biaix cap al rebuig de la hipòtesi nul.la (sense diferències significatives), que pot aparèixer en cas contrari.

3.3.4.5.2. Grau de suport dels clades de l'arbre/arbres obtingut/s

No tots els clades que apareixen en un arbre tenen el mateix suport. Aquest suport pot entendre's com el grau d'estabilitat d'un clade davant de pertorbacions de les

dades existents (p. ex. revisió i recodificació d'un caràcter o eliminació del mateix) o de la incorporació de noves (p. ex. noves seqüències o caràcters comportamentals). Una primera aproximació a aquest suport és veure el nombre de caràcters que recolza una branca. Un criteri addicional seria avaluar els tipus de caràcters implicats, p. ex. caràcters morfològics complexes o simples, o transicions enfront de transversions. Un refinament a la simple llargada de les branques seria considerar només les sinapomorfies dels clades que no siguin homoplàsiques, és a dir, aquelles que no hagin ocorregut enlloc més de l'arbre (Kluge 1989). Malgrat tot, s'ha suggerit que aquest tipus de consideracions poden ser errònies, ja que l'homoplàsia no té perquè estar uniformement distribuïda al llarg de l'arbre, sinó que es pot concentrar en certes parts d'aquest (Sanderson 1995). Entre els índexos proposats com a mesura d'estabilitat dels clades esmentarem els següents:

1. *Bootstrap*. Felsenstein (1985a) va proposar d'utilitzar el test estadístic del *bootstrap* (Efron 1979, 1982, 1987) per estimar els límits de confiança de les branques internes d'un arbre filogenètic. Des d'un punt de vista purament estadístic, el *bootstrap*, com a mètode de remostreig, és un mètode per a trobar variances, intervals de confiança i d'altres característiques de certs estadístics quan no es coneix, o és difícil d'establir, la distribució de les mostres. Per a fer això s'assumeix que la mostra inicial representa fidelment les característiques de l'univers al qual representa i aleshores, en lloc d'obtenir més mostres d'aquest univers, es generen noves mostres a partir de l'original, per remostreig. En el cas del *bootstrap*, cada pseudorèplica s'obté mitjançant remostreig amb reemplaçament de les dades originals i té la mateixa mida de mostra que les dades originals. El nivell de confiança d'una hipòtesi pot ser estimat pel percentatge de vegades que la hipòtesi és suportada per les pseudorèpliques construïdes per *bootstrap*. En el context de la inferència filogenètica, la matriu de caràcters representa a la mostra inicial. Les pseudorèpliques són matrius amb el mateix nombre de taxons i caràcters però en les quals els caràcters han estat remostrejats amb reemplaçament, de manera que alguns estaran repetits vàries vegades i d'altres seran absents (**fig. 20B**). A continuació, s'obté l'arbre més

parsimoniós per a cadascuna de les pseudorèpliques. El suport d'una determinada branca serà el percentatge de cops que el clade que aquesta defineix apareix en el conjunt d'arbres de les pseudorèpliques **(proporció de *bootstraps*, BP)**. El primer problema que planteja la utilització del *bootstrap* com a mesura de suport és, precisament, el saber exactament què és el què està mesurant. Així, Felsenstein (1985a) proposa originalment que el *bootstrap* seria la probabilitat de què una branca interna fos recuperada en anàlisis posteriors amb noves dades independents. Posteriorment, Felsenstein i Kishino (1993) suggereixen que la proporció de *bootstraps* s'interpreta millor com la probabilitat de que una branca determinada estigui representada en la filogènia veritable del grup. Ambdues interpretacions són força diferents, la primera es coneix com a **repetibilitat (*repeatibility*)** i la segona com a **exactitut (*accuracy*)**. Un tercer concepte que de vegades pot produir confusions, és el de **precisió (*precision*)**. La precisió, només es refereix al grau amb que un nombre limitat de pseudorèpliques s'aproximen al valor de BP obtingut a través d'un nombre infinit d'aquestes. La precisió és una funció simple del nombre de pseudorèpliques, que s'ha calculat han de ser entre 400-2000 per a obtenir una estima acurada de les BP reals (Hedges 1992). Hillis i Bull (1993) han examinat aquests dos possibles sentits de les BP utilitzant simulacions d'ordinador i filogènies de laboratori i conclouen que les BP constitueixen (1) una mesura molt poc acurada del grau de repetibilitat de la branca o clade considerat i (2) són mesures molt esbiaixades de l'exactitud d'un clade o branca, que subestimen els valors alts i sobreestimen els petits. A més, aquest biaix és depenent del nombre de taxons, del nombre de caràcters i de la topologia de l'arbre i, per tant, no poden ser directament comparats entre estudis. Malgrat tot, aquests autors consideren que els BP són útils com a mesura de la confiança relativa en els diferents clades de l'arbre. S'han proposat diverses correccions del càlcul de les BP per tal de corregir per el biaix (Bull i col. 1993, Rodrigo 1993, Zharkikh i Li 1995), però la seva implementació computacional està molt limitada o és molt costosa. Existeixen un llistat molt llarg de crítiques a la

utilització dels *bootstraps* en reconstrucció filogenètica, basades principalment en la violació de la majoria de les seves assumpcions. Potser la més fonamental és la de què, per a què el *bootstrap* sigui vàlid, els caràcters sujectes a remostreig han de ser independents els uns dels altres i, a més, han d'estar idènticament distribuïts (iid). En el context dels caràcters, això és tant com suposar que l'evolució és estocàstica, la qual cosa, si s'accepta el concepte d'historicitat, és certament falsa (Kluge i Wolf 1993, Brown 1994). D'altra banda, el *bootstrap* assumeix que l'estadístic a analitzar és una funció 'suau' de la distribució de la variable remostrejada, i això és cert per a estadístics que prenguin valors reals. Tanmateix, la presència o absència d'un clade és un valor binari i la teoria del *bootstrap* encara no ha estat desenvolupada en aquests supòsits (Brown 1994). Una altra assumpció del *bootstrap* és que el mostreig, elecció i codificació dels caràcters es realitza a l'atzar, és a dir, sense biaix per part de l'investigador. No obstant això, els sistemàtics tendeixen a recopilar els caràcters evolutivament més conservatius i més adients per al nivell de resolució d'interès (Jones i col. 1993, Kluge i Wolf 1993). Per aconseguir que la mostra recuperi la major part de les característiques de l'univers que representa, la seva mida ha de ser força gran. Malauradament però, la majoria d'investigadors es veuen obligats a treballar amb un nombre de dades molt limitat, si es té en compte l'enorme nombre de dades disponible en biologia comparativa (Jones i col. 1993, Kluge i Wolf 1993). Un altre problema és que, en calcular les proporcions de *bootstrap*, s'assumeix que han estat trobats els arbres més parsimoniosos per a cada pseudorèplica, és a dir, la millor estima de la filogènia que aquesta recolza. Tanmateix, quan el nombre de dades és gran, o ho és el nombre de pseudorèpliques que es pretén analitzar, es realitzen sempre cerques heurístiques que, com ja ha estat esmentat, no asseguren que es trobi l'arbre més parsimoniós (Sanderson 1995). Trueman (1993) ha argumentat que el *bootstrap,* com que en cada pseudorèplica elimina uns caràcters i multiplica uns altres, pot ser reinterpretat com una mena de ponderació dels caràcters a l'atzar, de

manera que les parts de l'arbre menys sensibles al pes assignat apareixeran més freqüentment en les pseudorèpliques. Cal encara veure si això és d'algún interès (Trueman 1993, Bremer 1994). D'altra banda, els valors de *bootstrap* són sensibles a la presència de caràcters autapomorfs, de manera que quan aquests apareixen, els valors de BP tendeixen a ser més baixos (Carpenter 1994). Tal i com va ser originalment proposat per Felsenstein (1985a), i com encara es pot veure en alguns estudis, els resultats finals del *bootstrap* es presenten en forma d'arbre de consens majoritari (veure apartat sobre arbres de consens) a on apareixen els clades amb uns BP superiors a un cert valor preestablert. Pot produir-se una situació en què en aquest arbre apareguin clades no presents en l'arbre més parsimoniós o, al contrari, que clades que hi eren presents en aquest desapareguin en el de consens. Això comporta el fet curiós, i epistemològicament injustificable, que la hipòtesi de filogènia escollida no sigui la més parsimoniosa (Miyamoto 1985). Malgrat aquest llistat de crítiques, que ni tan sols les inclou totes (per crítiques addicionals veure Carpenter 1992, Hillis i Bull 1993, Jones i col. 1993, Kluge i Wolf 1993, Trueman 1993, Brown 1994, Carpenter 1994), i que en paraules d'un dels proponents de la utilització del *bootstrap* '...la utilització del *bootstrap* en inferència filogenètica està mancada de qualsevol mena de justificació rigorosa' (Brown 1994), aquesta tècnica segueix sent la més intensament utilitzada en els estudis de reconstrucció filogenètica, principalment en aquells que utilitzen dades moleculars.

2. *Suport de Bremer (=Decay Index, índex de suport, suport de les branques).* Una aproximació radicalment diferent al problema de l'avaluació de l'estabilitat dels clades de l'arbre més parsimoniós ha estat proposada per Bremer (1988, 1994): un clade està tant més fortament recolzat per les dades com més gran sigui el nombre de pesos necessaris per col.lapsar-lo. Dit d'una altra manera, el suport d'un clade és la diferència en nombre de passos entre l'arbre més parsimoniós per les dades i l'arbre més parsimoniós que no contingui aquest clade. En principi, aquest valor és comparable entre matrius que continguin caràcters del mateix tipus, ja siguin morfològics o

moleculars (Bremer 1994). En el cas en què els caràcters han estat ponderats prèviament, es fa necessari corregir el valor de suport segons l'escala de ponderació aplicada. El valor de suport corregit resulta de dividir el valor de suport ponderat pel quocient entre el nombre de passos del cladograma amb les dades ponderades i el nombre de passos del mateix cladograma amb els caràcters uniformement ponderats i amb un pes igual a 1 (Gustafsson i Bremer 1995). Potser un dels avantatges més importants d'aquest mètode d'establir el suport de les branques és que es basa en les dades reals, no en pertorbacions d'aquestes (Bremer 1994). Com s'ha vist anteriorment, a partir del suport de Bremer s'ha derivat l'anomenat suport total, resultat de sumar els suports de Bremer de tots els clades de l'arbre, per mesurar l'estructura filogenètica de l'arbre (Källersjö i col. 1992). Tanmateix, Bremer (1994) proposa un índex, l'**índex de suport total**, per mesurar l'estabilitat global de l'arbre. L'índex de suport total (t_i) es defineix com el quocient entre el suport total (t) i el nombre de pesos del cladograma uniformement ponderat i amb un pes igual a 1 (p). Com que p és el nombre màxim que pot assolir t, el valor de t_i pot variar entre 0, en el cas en què no hi hagi cap suport per cap clade, és a dir, una politomia total, i 1, quan t i p siguin iguals, situació que es donarà només en absència total d'homoplàsia. La principal limitació del suport de Bremer és que es tracta d'un índex qualitatiu i que li manca un valor llindar a partir del qual considerar que un clade està ben recolzat. Faith (1991) va desenvolupar els mètodes de permutació per a utilitzar-los en el context del grau de suport de la monofília d'un grup. L'anomenat ***T-PTP*** (*topology-dependent permutation tail probability*) és un test de significació dissenyat per a comprovar si un clade determinat està significativament suportat. Amb tot, simulacions per ordenador del comportament d'aquest test han demostrat que en realitat no mesura exactament la monofília d'un grup (Huelsenbeck i col 1995, Swofford i col. 1996b).

3. *JMI (Jackknife Monophyly Index).* Siddall (1995) ha proposat de mesurar l'estabilitat de les branques del cladograma davant pertorbacions degudes

a la possible inclusió de nous taxons. Per tal de mesurar aquest efecte, es construeixen pseudorèpliques de la matriu original utilitzant jackknife de primer ordre (**fig. 20A**). A diferència dels altres mètodes d'aleatorització comentats, en aquest cas es remostregen els taxons, de manera que cada nova pseudorèplica és el resultat de l'eliminació d'un dels taxons. El pas següent és trobar l'arbre més parsimoniós de cada pseudorèplica i posteriorment es calcula la freqüència amb què cadascun dels clades de l'arbre original apareix en els arbres de cada pseudorèplica. Un clade es considera present en una pseudorèplica si apareix exactament igual que en l'arbre original o si només li falta el taxò que s'ha eliminat en la pseudorèplica considerada. El valor final de l'índex de cada clade es calcula sumant les proporcions amb què apareixen en cada pseudorèplica i dividint pel nombre total de taxons de la matriu, que és la mateixa manera com es calculen les proporcions de *bootstrap*. Es recomana que l'*outgroup* es mantingui al marge de les pertorbacions, ja que l'existència d'una arrel és indispensable per a la definició de monofília. Aquest índex sembla no ser sensible al nombre de sinapomorfies que recolzen els grups, ni a la presència de caràcters no informatius (Siddall 1995). Ha estat explícitament aconsellat no utilitzar el valor de l'índex com a mesura estadística, sinó simplement com a aproximació a la influència de l'homoplàsia sobre la relativa estabilitat dels clades en l'arbre més parsimoniós. S'ha observat que l'eliminació de certs taxons condueix a un augment molt important en el nombre d'arbres més parsimoniosos i que, per tant, la seva presència en l'anàlisi contribueix a estabilitzar els resultats. A aquests taxons se'ls anomena **crítics** Per contra, l'eliminació d'altres taxons té l'efecte contrari, i redueix el nombre inicial d'arbres. A aquests se'ls coneix com a **problemàtics**.

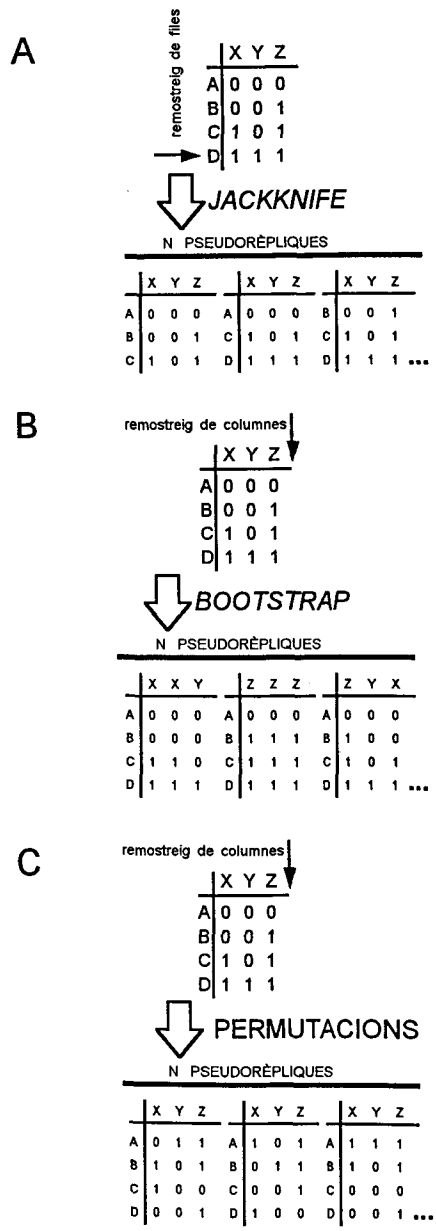


Figura 20.-A-C. Tipus d'aleatoritzacions aplicables sobre matrius.-**A.** *jackknife* dels taxons (files).-**B.** *bootstrap* dels caràcters (columnes).-**C.** permutacions dels caràcters (columnes).

3.3.4.6. Arbres de consens

Els arbres resultants de l'anàlisi filogenètica s'anomenen arbres **fonamentals** i són els que representen la informació continguda a les dades. Hi ha, però, un altre tipus d'arbres on el que es representa és la informació continguda en altres arbres i que reben el nom d'**arbres de consens**. L'objectiu principal dels mètodes de consens és representar gràficament els punts d'acord entre els diferents arbres fonamentals. Existeixen molts tipus diferents de metodologies per derivar arbres de consens (per descripció i resultats de l'aplicació d'alguns d'ells veure Kitching 1992, Quicke 1993, Wilkinson 1995b). Tanmateix, la gran majoria d'aquests mètodes produeixen arbres amb clades que només es troben en algun dels arbres fonamentals o, en el pitjor dels casos, clades que no es troben en cap dels arbres fonamentals. Per tant, la seva utilització com a sumari del conjunt d'arbres més parsimoniosos provinents d'una sola matriu és, com a mínim, poc recomanable (Nixon i Carpenter 1996a). Amb tot, la seva aplicació en el context de sintetitzar la informació d'arbres provinents de diferents matrius de dades, p. ex. en biogeografia, pot estar força justificada. El mètode de consens més àmpliament utilitzat és l'anomenat **consens estricte (*strict consensus*)**. Aquest és l'únic mètode que produeix un arbre a on només hi ha representats els clades que estan presents en tots els arbres fonamentals considerats (**fig. 21**). Un aspecte negatiu del consens estricte és que degut a la seva naturalesa tan conservativa, en moltes ocasions l'arbre resultant mostra un elevat nombre de politomies que comprometen seriosament el seu nivell de resolució. Tot i així, la pèrdua de resolució queda justificada per la inclusió en el consens de només aquells components que no són ambigus, és a dir, pels quals les dades són absolutament clares. Malauradament, en certes ocasions això pot no ser absolutament cert degut a l'existència en els arbres fonamentals de **branques de llargada zero (*zero-length***

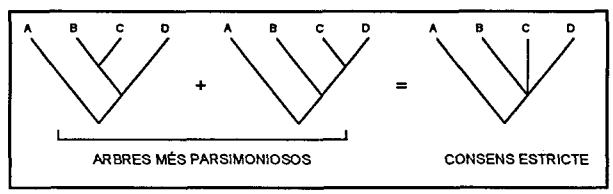


Figura 21.- Exemple de construcció d'un arbre de consens estricte.

branches) i de politomies. És ben conegut (Platnick i col. 1991b, Wilkinson 1995b), que en certes situacions alguns programes d'ordenador d'inferència filogenètica poden donar com a resultat cladogrames que no estan recolzats en cadascun dels seus nodes per canvis de caràcter. Aquest fenomen és el resultat de l'existència d'ambigüitat en l'optimització d'alguns caràcters. Així, un clade determinat pot estar recolzat per un canvi sota una certa optimització (p. ex DELTRAN), mentre un clade diferent està suportat per una optimització diferent (p. ex. ACCTRAN) del mateix caràcter. Ambdós clades no poden coexistir, ja que depenen d'optimitzacions alternatives del mateix caràcter. Tot i així, alguns programes ofereixen l'arbre que presenta aquests dos clades com a una de les solucions (Coddington i Scharff 1994). D'altra banda, no hi ha acord sobre quina és la millor manera de tractar els arbres amb branques de llargada zero. Goloboff (1996a,b,c,d) i Nixon i Carpenter (1996a) proposen col.lapsar tots aquells clades que no estiguin recolzats en cadascuna de les possibles optimitzacions dels caràcters, és a dir, que tinguin una llargada mínima de zero (**suport estricte**), el que comporta l'eliminació dels arbres que difereixen dels altres per l'existència d'un clade recolzat per una optimització particular d'un caràcter. Segons aquests autors, aquests arbres no representen hipòtesis filogenètiques dignes de ser considerades. Contràriament, Coddington i Scharff (1994) consideren que els arbres abans esmentats, és a dir, els que només difereixen per clades suportats per optimitzacions alternatives, són hipòtesis filogenètiques legítimes (**suport semiestricte**). Aquest autors suggereixen, per tant, considerar tots els arbres possibles en els quals els clades estiguin recolzats almenys sota alguna optimització i eliminar posteriorment els arbres a on apareixen branques de llargada zero manualment. Com ha estat esmentat amb anterioritat, en algunes ocasions alguns dels arbres més parsimoniosos poden mostrar una o vàries politomies degut a l'absència de canvis d'estat que donin suport a algun dels clades possibles dins d'aquesta politomia. Alhora, aquests arbres poden ser compatibles o incompatibles amb la resta d'arbres més parsimoniosos. Malgrat que, en general, aquestes politomies fonamentals s'interpreten com a 'suaus', els programes les interpreten com a 'dures'. Coddington i Scharff (1996) han argumentat que sota l'assumpció de les politomies com a 'suaus', qualsevol arbre

dins del conjunt del més parsimoniosos que presenti alguna politomia, compatible o no amb la resta, ha de ser eliminat de qualsevol anàlisi posterior. La justificació d'això és que els arbres més resolutius i més parsimoniosos maximitzen el nombre de clades, el contingut d'informació i el grau de falsificabilitat de les hipòtesis filogenètiques. Això suposa la inspecció ocular del conjunt d'arbres més parsimoniosos per a detectar i eliminar aquells que presentin politomies. Una aproximació diferent a aquestes situacions ha estat proposada per Nixon i Carpenter (1996a) que consideren que els arbres de consens (estricte) poden ser utilitzats com a font d'informació sobre el conflicte entre els caràcters dels arbres fonamentals. Aquest raonament es basa en la constatació de què el consens representa el límit superior del nombre de passos dels caràcters en els arbres fonamentals. A partir d'aquesta observació, es suggereix la utilització d'un nou índex, la **concordància de clades (*CC*)**, definit com:

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

A on, g_n és el nombre màxim de passos d'un caràcter en els arbres fonamentals, S_c és la llargada de l'arbre de consens i S_p és la llargada dels arbres fonamentals (= més parsimoniosos). El numerador $(\Sigma g_n - S_p)$ representa el grau de conflicte dels caràcter entre els cladogrames o, també, l'homoplàsia entre cladogrames. Aquest índex varia entre 0 i 1. L'índex és 0 quan el nombre de passos de tots i cadascun dels caràcters arriba al valor màxim almenys en un dels arbres fonamentals, que és alhora el mateix nombre de passos dels caràcters en el consens. L'índex tendeix cap a 1 a mesura que la diferència entre la llargada del consens i les llargades màximes dels caràcters entre els arbres fonamentals augmenta. A la pràctica, valors baixos de l'índex indiquen que les incongruències (=politomies) del consens són principalment degudes a conflictes entre els caràcters dels arbres fonamentals. Per contra, valors alts de l'índex suggereixen que no hi ha conflicte entre els caràcters, la qual cosa implica que encara hi ha informació per agrupar alguns dels taxons inclosos a les politomies. L'índex és u quan, o bé només hi ha un cladograma més parsimoniós, o bé els grups conflictius només estan recolzats per optimitzacions ambigües.

3.3.4.7. Combinació de dades de diferent naturalesa: anàlisi sobre particions o anàlisi simultània?

El gran desenvolupament i accessibilitat de moltes de les tècniques de la biologia molecular, han permès que en l'actualitat sigui relativament senzill pels sistemàtics d'obtenir diferents tipus de dades moleculars dels taxons objecte de la seva recerca. Paral.lelament, els últims anys han estat testimonis de la rehabilitació de les dades comportamentals, fisiològiques, etc... en el context de la reconstrucció filogenètica. La possibilitat d'incorporar conjunts de dades tan heterogènies com les provinents de la morfologia, el comportament, les seqüències nucleotídiques de diferents gens, els enzims, etc. planteja el problema de quina és la millor manera d'analitzar-les. S'han proposat tres aproximacions principals.

3.3.4.7.1. Anàlisi separada de les diferents particions de les dades (= taxonomic congruence, Kluge 1989*; = partitioned analysis,* Nixon i Carpenter 1996b*).*

Aquesta aproximació considera que la millor manera de tractar les diferents classes de dades és analitzar-les independentment, obtenir un o més arbres per cadascuna de les matrius i utilitzar un mètode de consens per representar gràficament la informació comuna (**fig. 22A**). Els suposats avantatges d'aquesta opció són que (Miyamoto i Fitch 1995): (1) aquest consens és una estima conservativa de la filogènia, (2) els caràcters provinents de diferents fonts són independents entre ells i, per tant, és menys probable que recolzin la mateixa filogènia si aquesta és errònia i (3) hi ha casos en què, per la naturalesa de les dades, és impossible aplicar el mateix mètode de reconstrucció a totes elles simultàniament (p. ex. caràcters morfològics amb dades d'hibridació de DNA-DNA). Tanmateix, la majoria d'aquests punts han estat rebatuts. Així, la utilització del consens no permet d'avaluar el suport relatiu dels grups conflictius entre els arbres originals (Nixon i Carpenter 1996b). D'altra banda, l'anàlisi separada i posterior consens comportarà

la ponderació diferencial dels caràcters de les diferents particions (Kluge i Wolf 1993), ja que si, com s'ha comentat, hom assumeix que hi ha més probabilitat de què els caràcters siguin independents entre diferents classes de dades, en analitzar cada classe per separat la dependència dels caràcters serà més gran. La presència de caràcters depenents té el mateix efecte que l'augment de pes d'un caràcter que suporti les mateixes agrupacions que aquests. L'única manera d'evitar aquest efecte és combinar les matrius per a augmentar el nombre de caràcters independents (Nixon i Carpenter 1996b). Finalment, tots el arbres a partir dels quals es deriva el consens hi contribueixen de la mateixa manera, malgrat que alguns d'ells s'hagin obtingut de particions amb un major senyal filogenètic que d'altres.

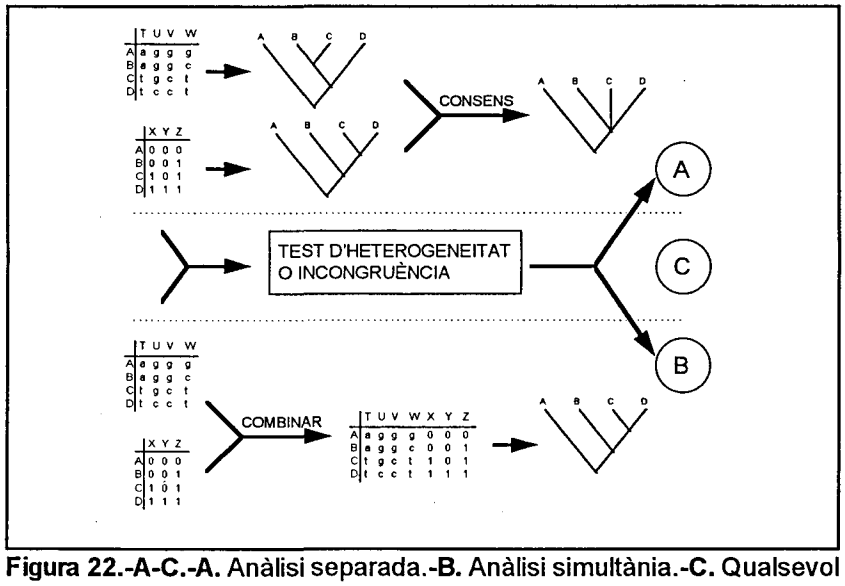


Figura 22.-A-C.-A. Anàlisi separada.-**B.** Anàlisi simultània.-**C.** Qualsevol dels anteriors, depenent de l'homogenïtat de les dades.

3.3.4.7.2. Anàlisi conjunta de les dades (= character congruence o total evidence, Kluge 1989*; = simultaneous analysis* Nixon i Carpenter 1996b*).*

Aquesta aproximació proposa combinar les dades en una sola matriu i analitzar-les conjuntament (**fig. 22B**). Els avantatges d'aquest tractament són què: (1) les hipòtesis obtingudes estan basades en la màxima evidència possible (Kluge i Wolf 1993), és a dir, el conflicte entre els caràcters és resolt en base a tota l'evidència

disponible, (2) en certes situacions diferents conjunts de dades ofereixen informació a distints nivells filogenètics, (3) un conjunt de dades amb un senyal filogenètic dèbil degut a la presència de molta homoplàsia pot veure incrementada la seva senyal per congruència amb altres caràcters de conjunts diferents i (4) els caràcters combinats en una sola matriu poden ser reestructurats durant l'anàlisi i recolzar clades que no es trobarien en l'anàlisi separada. Un problema força freqüent quan es combinen matrius de dades diferents és el fet que alguns dels taxons poden no estar representats en totes les matrius. En aquests casos l'única opció possible és omplir amb interrogants els caràcters de les particions on el taxó és absent o, simplement, eliminar aquests taxons. Alguns autors han argumentat en contra de la combinació de dades en els casos en què es demostri que les dades per separat recolzen filogènies irreconciliables (Bull i col. 1993, De Queiroz 1993, Huelsenbeck i col. 1994). El raonament és que aquesta situació és conseqüència de què les diferents dades estan sotmeses a diferents processos evolutius que no poden ser copsats per l'aplicació d'una sola metodologia d'inferència filogenètica comuna a totes elles.

3.3.4.7.3. Anàlisi conjunta condicional (=prior agreement approach **Chippindale i Wiens 1994*).***

Aquesta aproximació proposa fonamentar l'elecció d'un dels dos tipus esmentats d'anàlisi en la comprovació prèvia de l'existència de congruència o homogeneïtat entre les diferents dades (**fig. 22C**). Evidentment, l'aplicació d'aquesta estratègia fa necessària la definició d'un test de detecció d'incongruència o heterogeneïtat i han estat descrit un cert nombre de tests estadístics per investigar la incongruència. Recentment, Cunningham (1997b) ha estudiat el comportament d'alguns dels tests d'incongruència (Templeton 1983, Rodrigo i col 1993, Farris i col. 1994) desenvolupats per a parsimònia i ha arribat a la conclusió de què l'anomenat **test *ILD*** (***Incongruence Length Difference***, Farris i col. 1994) és el més útil. Aquest test de significació es basa en un índex originalment descrit per Mickevich i Farris

(1981) per a distingir l'homoplàsia resultant de la combinació de matrius diferents de la deguda a la incongruència dels caràcters dins de cada matriu. Aquest índex es defineix com:

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

A on S_{a+b} és el nombre de passos del cladograma resultant de la matriu combinada i S_a, S_b els passos dels arbres de cada matriu per separat. La distribució de l'estadístic *ILD* es construeix calculant el seu valor original i els valors resultants de l'aleatorització de les particions. Aquestes particions aleatoritzades tenen la mateixa mida que les originals, però estan formades per caràcters seleccionats a l'atzar entre les matrius combinades, és a dir, barrejant caràcters de les dues particions originals. L'hipòtesi nul.la és que hi ha congruència entre les particions i és rebutjada quan la proporció de vegades en què el valor d'incongruència aleatoritzat és igual o més gran que l'original, és menor que un cert nivell de significació (p.ex. 0,05[2]). Aquest test pot ser aplicat en presència de ponderació diferencial de les dades i, també, simultàniament a més de dues particions de les dades. D'altra banda, es recomana d'eliminar els caràcters invariables en el cas que les diferents particions difereixin en el nombre d'aquests (Cunningham 1997b). Amb tot, alguns autors consideren que l'índex *ILD* és útil per detectar els caràcters discordants i els patrons de congruència entre les diferents particions, però que aquestes han de ser igualment combinades i analitzades simultàniament fins i tot si el test resulta significatiu (Nixon i Carpenter 1996b).

3.4. ESTUDI DELS PROCESSOS

Durant molt de temps, diferents branques de la biologia comparada, especialment l'ecologia, l'etologia i la biologia de poblacions, han estat dominades per explicacions basades en processos observables a l'actualitat (Wanntorp 1990). Tanmateix, als últims anys aquestes ciències han viscut una petita revolució amb

[2] Cunningham (1997b, i referències allí) considera aquest valor massa conservatiu.

l'adopció de la filogènia com a eina per a revelar els patrons i processos evolutius (Miller i Wenzel 1995). Una perspectiva històrica en biologia condueix a un coneixement més profund de la natura.

Els beneficis d'adreçar les qüestions evolutives en un context filogenètic són diversos (Brooks i col. 1995, Miller i Wenzel 1995, Desutter-Grandcolas 1997). D'una banda, la filogènia proveeix informació sobre la polaritat dels caràcters, és a dir, sobre la seqüència de canvis que descriuen la transformació del caràcter al llarg del temps evolutiu. D'una altra, permet testar hipòtesi sobre els processos responsables d'una determinada configuració. Generalment, existeixen diversos escenari ecològics, geològics, etc., que permeten explicar una determinada observació biològica, els quals poden ser perfectament lògics però mútuament excloents. La filogènia permet contrastar aquests escenari mitjançant un cert patró de l'origen i distribució del caràcters. Finalment, la filogènia permet abordar l'estudi d'associacions (correl.lacions) entre trets (coevolució entre caràcters) o entre clades (coespeciació), oferint informació sobre els diferents nivells de dependència entre les comparacions realitzades (Harvey i Pagel 1991).

Existeixen dues aproximacions diferents a l'estudi de processos biològics sota una perspectiva filogenètica, una fonamentalment qualitativa i un altre d'estadística. Sota l'aproximàcio qualitativa, la filogènia s'utilitza per obtenir la polaritat dels caràcters i la distribució de les seves transformacions. En el cas de que el caràcter objecte d'interès hagi estat inclòs a la matriu analitzada, s'obtè la polaritat del mateix directament de l'arbre. Si el caràcter no ha estat inclòs, llavors s'ha de traçar (= mapar) sobre el cladograma i optimitzar-lo. La qüestió de si el caràcter que s'estudia ha de ser inclòs o no a l'anàlisi filogenètica ha estat subjecte de debat. Els partidaris de no incloure'l argumenten que d'aquesta manera s'elimina la circularitat del raonament, ja que el caràcter es contrastat envers d'una filogènia construïda amb dades independents (Coddington 1988, Brooks i McLennan 1991). Els partidaris d'incloure'l consideren que la història evolutiva dels caràcters ha de ser jutjada amb referència al cladograma que millor encaixa amb el conjunt de l'evidència disponible, de la qual forma part també el caràcter investigat (Kluge i Wolf 1993).

Les aproximacions estadístiques es centren principalment a l'estudi de les correl.lacions entre trets o organismes. El problema és que aquestes correl.lacions es basen en comparacions que no són independents, ja que tenen un lligam filogenètic, la qual cosa es tradueix en una disminució dels graus de llibertat dels estadístics de correl.lació i un augment del nivell de significació. Si es té un cladograma, es poden contar els successos evolutius separats i millorar el càlculs. S'han desenvolupat tot un seguit de tests estadístics de correl.lacions entre variables utilitzant la informació sobre les relacions genealògiques dels organismes (Harvey i Pagel 1991). Aquesta aproximació ha estat criticada argumentant que els valors de significació que es deriven amb aquestes tècniques es basen en assumpcions falses i estan mancats d'un sentit biològic (Miller i Wenzel 1995).

CAPÍTOL 4. RESULTATS I DISCUSSIÓ

4.1. TAXONOMIA: DEFINICIÓ I DELIMITACIÓ DE LES UNITATS D'ESTUDI

4.1.1 Les illes occidentals: La Gomera, La Palma i El Hierro

Arnedo, M. A., Oromí, P. & Ribera, C. 1996. Radiation of the genus *Dysdera* (Araneae, Haplogynae, Dysderidae) in the Canary Islands: The western islands.- *Zoologica Scripta 25*: 241-274.