K.U.Leuven
Faculteit der Wetenschappen
Instituut voor Plantkunde
Laboratorium voor Systematiek

A reconsideration of three-item analysis, the use of implied weights in cladistics, and a practical application in Gentianaceae

Jan De Laet

Proefschrift voorgedragen tot het behalen van de graad van
Doctor in de Wetenschappen

Promotor: Prof. Dr. E. Smets

This pdf version has subtle differences in page layout (but not content) compared to the original hardcopy

K.U.Leuven
Faculteit der Wetenschappen
Instituut voor Plantkunde
Laboratorium voor Systematiek

A reconsideration of three-item analysis, the use of implied weights in cladistics, and a practical application in Gentianaceae

Jan De Laet

INCLUDING THE
1998 ERRATUM
FOR APPENDIX B

Proefschrift voorgedragen tot het behalen van de graad van
Doctor in de Wetenschappen
27 mei 1997

Leden van de jury:

Promotor, Prof. Dr. E. Smets	K.U.Leuven
Prof. Dr. F. Symons	K.U.Leuven
Prof. Dr. J. Van Assche	K.U.Leuven
Prof. Dr. E. Robbrecht	Nationale Plantentuin van België, U.I.A.
Dr. J. S. Farris	Naturhistoriska Riksmuseet, Stockholm

LIST OF ABBREVIATIONS

Apart from K and $\bar{S}$ (Goloboff 1993a, c) and C (chapter 3), all abbreviations and definitions are from Farris (1969, 1989) and Kluge & Farris (1969).

C concavity constant of the logarithmic fit function $-\ln((1+h)/C)$

ci the consistency index of a given character on a given cladogram
$ci = m/s$

CI the ensemble consistency index of a data set on a cladogram
$CI = M/S$

g the length of a character on a completely unresolved cladogram

G the length of a data set on a completely unresolved cladogram
$G = \Sigma_{i=1..\text{number of characters}}\, g_i$

h the homoplasy of a given character on a given cladogram
$h = s-m$

H the total homoplasy of a given data set on a given cladogram
$H = \Sigma_{i=1..\text{number of characters}}\, h_i = S-M$

K concavity constant of the hyperbolic fit function $K/(K+h)$

m the observed variation in a given character: the minimum number of steps the character can have on any cladogram

M the total observed variation of a given data set
$M = \Sigma_{i=1..\text{number of characters}}\, m_i = S-H$

rc the rescaled consistency index of a character on a given cladogram
$rc = ci*ri$

RC the ensemble rescaled consistency index of a data set on a given cladogram
$RC = CI*RI$

ri the retention index of a given character on a given cladogram
$rc = (g-s)/(g-m)$

RI the ensemble retention index of a data set on a given cladogram
$RC = (G-S)/(G-M)$

s the minimal length or number of steps of a character on a cladogram
$s = m+h$

S the total minimal length or number of steps of a data set on a cladogram
$S = \Sigma_{i=1..\text{number of characters}}\, s_i = M+H$

$\bar{S}$ the mean total minimal length of a data set on all possible dichotomous cladograms

There is a lesson in all this.

We do not know in advance
what are the right questions to ask,
and we often do not find out until we are close to an answer.

(Weinberg 1997: 215)

Vooraf ...

... wil ik een woordje van dank richten

aan Prof. Dr. E. Smets, promotor van dit proefschrift, voor zijn steun en begeleiding, vooral tijdens de laatste maanden. Met grote snelheid las hij de geschreven stukken door en gaf hij waar nodig zijn advies.

aan alle labcollega's van vroeger en nu voor de toffe werksfeer en samenwerking; in het bijzonder aan Peter voor het maken van figuren 1.1 en 1.5.

aan Jos voor de steeds bereidwillige hulp bij het bibliotheekwerk (en het wisselen van vijffrankstukken voor de koffiemachine...).

aan alle studenten die de voorbije jaren mijn pad gekruist hebben; ik hoop dat ze zoveel van mij geleerd hebben als ik van hen.

aan Sandra, broers en zussen en alle vrienden die me de voorbije maanden een welgemeend duwtje in de rug gegeven hebben; de laatste loodjes wogen inderdaad zwaar.

Tenslotte wil ik mijn ouders vermelden. Zonder hen zou dit werk er niet gekomen zijn, en dat is meer dan een triviale biologische vaststelling.

Jan

Contents

Introduction

The central theme of this thesis is cladistics. This approach to phylogenetic analysis has its roots in Willi Hennig's theoretical work of the fifties (see chapter 1), and after a modest take-off in the sixties and a period of exponential growth in the seventies and the eighties, cladistics has now become a basic tool in systematic research. Its merits are that it has stimulated the development of a conceptual framework that enables us to think and talk in a clear way about phylogenetic relationships, and that it provides a set of powerful methods to analyze systematic data in order to discover the underlying phylogenetic relationships. In the first chapter, a basic survey of the main concepts and terms of cladistics is given. It is presented in Dutch because to date no general introductions to cladistics exist in Dutch.

At present, there is a set of generally accepted methods in cladistics. However, this does not mean that the theoretical work has come to an end. To the contrary, old ideas are constantly being refined and new ideas keep popping up. Two of those are treated in the second and the third chapter.

The first one is three-item analysis, a method that was introduced some years ago as a novel approach to parsimony analysis in both biogeography (Nelson & Ladiges 1991a, b) and systematics (Nelson & Platnick 1991). The name three-item analysis refers to the fact that each statement about relationships between more than three items (areas in biogeography, homologous features in systematics) is decomposed into a series of basic statements, each of which involves only three items. Such a basic statement simply says which two of the three items are thought to be related more closely to each other than either is related to the third. Following its introduction, three-item analysis has been severely criticized because of three basic defects: (1) it is flawed because it presupposes that character evolution is irreversible; (2) it is flawed because basic statements that are not logically independent are treated as if they are; (3) it is flawed because some of the three-item statements that are considered as independent support for a given tree may be mutually exclusive on that tree. In the second chapter it is shown that these criticisms only relate to the particular way that the approach was implemented by Nelson & Platnick (1991), and an alternative implementation that solves each of the three basic problems is derived. However, the resulting method is not an improvement over standard parsimony analysis: it is identical to the standard approach but for one small constraint, which is a highly unnatural restriction on the maximum amount of homoplasy that may be concentrated in a single character state. As this restriction follows directly from the decomposition of character state distributions into basic statements, it is concluded that any approach that is based on such decompositions will be defective.

The second one is about character weighting. Some years ago, Goloboff (1993a) proposed a non-iterative homoplasy-based weighting method in which the weight or fit of a character on a cladogram is defined as a hyperbolic decreasing function of its homoplasy. The best trees are those that have the highest total fit over all characters of a data set. Goloboff considered his approach to be in direct agreement with cladistic ideas, but most parsimonious trees are those trees that imply the lowest amount of weighted homoplasy (Farris 1983), and these are not necessarily the trees that imply that the characters have the highest total fit, as is shown in chapter three. Several implications of this observation are discussed, and an alternative way of weighting characters is proposed. A computer program in which this approach is available is discussed in appendix A, and the approach is illustrated by using an indecisive data set (see chapter six) and the morphological Gentianaceae data set that is presented in chapter five.

Several cladistic analyses based on various types of data indicate that the Gentianaceae, a cosmoplolitan family of medium size, is one of the principal families of a monophyletic order Gentianales. Recent developments concerning the order Gentianales are reviewed against a historical background in chapter four. While a consensus is emerging about the monophyly of the Gentianales, much work remains to be done concerning the interfamilial and intrafamilial relationships within the order.

The most recent worldwide monograph of the Gentianaceae is over a century old (Gilg 1895). The 21 genera that are selected for the current analysis represent all Gilg’s tribes and subtribes except Leiphaimeae, Rusbyantheae and Voyrieae. Standard parsimony analyses and analyses using Goloboff’s approach of maximum fit give congruent results as far as the global relationships are concerned. The best supported clade contains *Eustoma* (Tachiinae) and all included Gentianinae, Erythraeinae and Chironiinae. The basal parts of the cladograms, involving the woody tropical representatives and *Exacum*, are poorly resolved.

This thesis is concluded with a short chapter on indecisive data sets. Goloboff (1991a, b) defined the cladistic decisiveness of a data set as the degree to which all possible resolved trees for the data set differ in length. He proposed a measure of the decisiveness of data sets, the DD statistic, and discussed some properties of indecisive data sets, a special type of data set for which every possible cladogram has the same length. His discussion of indecisive data sets was restricted to characters that have no missing entries. In this chapter I will first show how indecisive data sets

can be constructed when missing entries are present. Without missing entries, there is essentially only a single indecisive data set for a given number of taxa, but by allowing missing entries a wide variety of different indecisive data sets with a wide range of ensemble consistency and retention indices can be constructed (easy-to-calculate formulas for the length of an indecisive data set on a dichotomous tree and on an unresolved bush are derived in Appendix C). Such data sets are useful in the construction of hypothetical examples that illustrate the elusive nature of data decisiveness. It is concluded that simple measures such as Goloboff’s DD statistic are unable to capture the various aspects of the concept.

1. INLEIDING TOT HET CLADISME[1]

1.1 Inleiding

In deze bijdrage wordt aan de hand van een aantal voorbeelden een algemeen beeld geschetst van de cladistische methode, waarbij de nadruk niet zozeer op technische details of op practische aspecten ligt, maar op de algemene redeneringen van de cladistische filosofie. Een degelijk inzicht in deze basisprincipes is immers onontbeerlijk om de waarde van het cladisme realistisch te kunnen inschatten.

Cladisme ("*cladism*", "*cladistics*") of **spaarzaamheidsanalyse** ("*parsimony analysis*") is één van de drie grote groepen numerieke methodes die zich gedurende de voorbije decennia ontwikkeld hebben (en die nog steeds verder evolueren) om de fylogenetische geschiedenis van de levende wereld te reconstrueren. De twee andere groepen, enerzijds de **afstandsmethoden** ("*distance methods*"; gebaseerd op matrices met paarsgewijze afstanden tussen de betrokken taxa, bijvoorbeeld UPGMA) en anderzijds methoden gebaseerd op het statistisch principe van grootste aannemelijkheid ("*maximum likelihood methods*"), komen verder niet aan bod; voor verdere informatie en onderlinge verbanden wordt verwezen naar Darlu & Tassy (1993), Felsenstein (1988a, 1988b), en Swofford et al. (1996).

Binnen het cladisme ligt de nadruk in de eerste plaats op het vertakkingspatroon of het **cladistisch** aspect van de evolutie (κλαδοσ: tak). Dit patroon tracht men te reconstrueren door een analyse van de verspreiding van kenmerktoestanden over de bestudeerde taxa (bijvoorbeeld de toestanden geel, rood en blauw van het kenmerk bloemkleur). Grafisch wordt dit voorgesteld door middel van een **cladogram**. Dergelijke cladogrammen vormen op hun beurt een belangrijk interpretatiekader voor de **patristische**, **chronistische** en **fenetische** aspecten van de evolutie (fig. 1.1; sommige van deze termen hebben wel al naargelang de geraadpleegde auteurs verschillende en soms erg uiteenlopende definities).

Het **cladistische** aspect heeft te maken met het vertakkingspatroon van de evolutionaire lijnen; zo zijn taxa B en C in fig. 1.1 van elkaar gescheiden door slechts één vertakkingspunt, taxa C en D daarentegen door twee; C is cladistisch dus nauwer verwant met B dan met D. Het **patristische** aspect heeft te maken met de divergentie

[1] Gebaseerd op De Laet & Smets (1994a).

van kenmerken binnen evolutionaire lijnen (dit wordt soms ook **anagenetische** of **fyletische** evolutie genoemd; deze kunnen al dan niet gepaard gaan met speciatie binnen de lijn); in fig. 1.1 zijn B en C ontstaan uit eenzelfde meest recente gemeenschappelijke voorouder, maar B wijkt veel meer af van deze voorouder dan C; dit betekent dat de divergentie in de lijn naar B veel groter is dan in de lijn naar C.

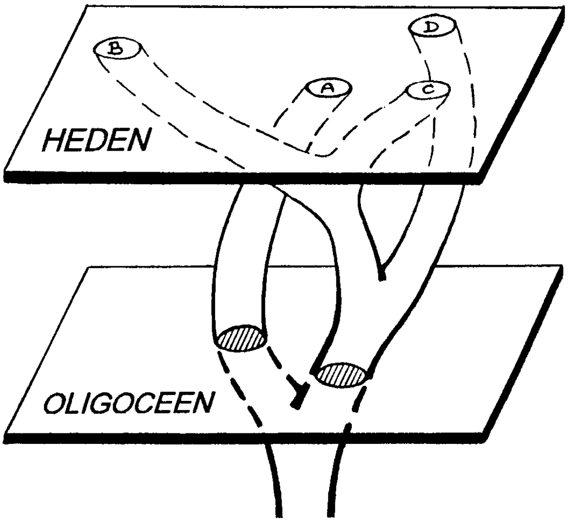


Fig. 1.1. Een schematische voorstelling van de evolutie van vier taxa, A, B, C en D. Zie tekst voor verdere uitleg.

Het **chronistische** aspect heeft betrekking op de datering van evolutionaire gebeurtenissen; zo leefde de meest recente gemeenschappelijke voorouder van A, B, C en D net voor het oligoceen. Het **fenetische** aspect heeft betrekking op het globaal verschil in kenmerken tussen taxa in een bepaalde tijdsdoorsnede (merk op dat dit zowel fenotypische als genotypische kenmerken kunnen zijn); in de huidige tijdsdoorsnede bijvoorbeeld vertonen A, C en D een grote gelijkenis, terwijl B er sterk van afwijkt.

Het cladisme zoals we het nu kennen vindt zijn oorsprong in de werken van Willi Hennig (1913-1976) een Duits entomoloog die aan *Diptera* werkte (het theoretisch werk van Hennig ontstond uiteraard niet in een vacuum; voor de historische context verwijzen we naar Bowler 1996). Hennig schreef in 1950 een eerste theoretisch werk, getiteld “*Grundzüge einer Theorie der phylogenetischen Systematik*”, dat echter weinig succes kende. “*Phylogenetic Systematics*” (1966), zijn tweede boek, kende veel meer bijval en ligt aan de basis van de doorbraak die het cladisme vanaf de jaren '70 kende, eerst in zoölogische kringen en enkele jaren later ook in de plantensystematiek. Hiernaast heeft ook de moleculaire systematiek reeds van in de vroege jaren ‘60 een belangrijke bijdrage geleverd aan het cladisme, voornamelijk op het vlak van de ontwikkeling van algoritmes die toelieten om de complexe analyses door computers te laten uitvoeren (Felsenstein 1988a).

W.H. Wagner, een Amerikaans pteridoloog, ontwikkelde in de jaren vijftig de **grondplandivergentie-methode** (“*grounplan divergence method*”; zie Wagner 1961). Deze methode heeft als doel om grafisch voor te stellen hoe de nakomelingen van een gemeenschappelijke voorouder in de loop van de evolutie van elkaar divergeren. De achterliggende theoretische overwegingen sluiten nauw aan bij Hennig's ideeën, maar werden onafhankelijk hiervan ontwikkeld. Wagner's werk bleef, net zoals Hennig's eerste boek, praktisch onopgemerkt, waardoor het een minder belangrijke rol gespeeld heeft in de ontwikkeling van het cladisme.

In het verleden werd de cladistische literatuur vaak gekenmerkt door heftige discussies en extreme standpunten. Met de jaren zijn de inzichten echter gerijpt en momenteel bestaat er over de belangrijkste punten vrijwel eensgezindheid. Een recent overzicht, voornamelijk vanuit het standpunt van de moleculaire systematiek, wordt gegeven door Stewart (1993). De logische basis van het cladisme werd uitgebreid behandeld door Farris (1983). Cronquist (1987) formuleerde een aantal kritieken op het cladisme zoals die bij veel botanici leefden. Deze kritieken, die vaak berustten op een onvolledige kennis van de cladistische theorie, werden door Humphries & Chappill (1988) en door Donoghue & Cantino (1988) vanuit een puur cladistische filosofie bekeken en weerlegd.

Recent heeft de veralgemeende doorbraak van moleculaire systematiek een nieuwe impuls gegeven aan de ontwikkeling van fylogenetische methoden. Getuige daarvan is bijvoorbeeld het tijdschrift “*Molecular Phylogenetics and Evolution*” dat in 1992 opgericht werd. Eén van de doelstellingen van dit tijdschrift is om de samenwerking en de dialoog tussen klassieke en moleculaire systematici te bevorderen (zie in dit verband eveneens Doyle 1993). Ook in de klassieke systematische tijdschriften krijgen moleculaire studies meer en meer hun plaats. Een indrukwekkend voorbeeld hiervan is het derde nummer van “*Annals of the Missouri Botanical Garden*” uit 1993, dat volledig gewijd is aan de cladistische analyse van de nucleotidesequentie van het *rbc*L-gen bij de zaadplanten, voornamelijk angiospermen. Het hoofdartikel van dit nummer (Chase et al., 1993; met meer dan 40 coauteurs uit meer dan 20 laboratoria; zie Baum 1994 voor enkele beschouwingen) is gewijd aan een analyse van 500 verschillende sequenties, verspreid over alle zaadplanten.

Eén van de eerste wetenschappers die het belang van het theoretisch werk van Hennig ingezien heeft, was de Gentse zoöloog Kiriakoff. Zo is zijn handboek “Beginselen der dierkundige systematiek voor hoogstudenten en biologen” (1956) reeds helemaal geschreven vanuit de cladistische filosofie. De ideeën van Hennig waren toen nog vrij onbekend, en het is dan ook niet verwonderlijk dat Kiriafoff in 1960 (Kiriakoff 1960: 15) vaststelde dat “de literatuur in onze taal erover betrekkelijk arm is”. Meer dan 30 jaar later en na de sterke ontwikkeling die het cladisme doorgemaakt heeft, is dat echter nog steeds zo. Voor zover we konden nagaan, bestaat er buiten de summiere inleiding van Schockaert (1992) geen andere recente en ruim beschikbare Nederlandse literatuur over dit onderwerp. De meningen over het cladisme zijn nog steeds al te vaak sterk gepolariseerd. Enerzijds is er bij sommige klassieke systematici een afwijzende houding die voornamelijk gebaseerd is op de foutieve veronderstelling dat er binnen de cladistische filosofie geen plaats zou zijn voor het traditioneel systematisch werk, of voor de inzichten die daaruit voortvloeien. Anderzijds vindt men vooral in kringen van moleculaire systematici soms een te blind en ongenuanceerd vertrouwen in de evolutionaire waarheid van alle mogelijke dendrogrammen die op een computerscherm te voorschijn kunnen gehaald worden. We hopen dat deze tekst kan bijdragen tot het tot stand komen van een meer realistische visie op het cladisme en van een dieper inzicht in de systematische methode in het algemeen.

1.2 Basisprincipes

In de loop van de evolutie kan een bepaalde evolutionaire lijn kenmerken ontwikkelen die haar van alle andere lijnen onderscheidt. Deze waarneembare en overerfbare eigenschappen worden evolutionaire nieuwigheden of “**apomorfieën**” genoemd. Wanneer een lijn, gekenmerkt door een aantal apomorfieën, zich later zal splitsen, zullen de dochterlijnen deze apomorfieën overerven. Apomorfieën zijn dus de sleutel om cladistische verwantschappen tussen lijnen op te sporen: een apomorfie die voorkomt in twee of meer groepen, is met grote waarschijnlijkheid ontstaan in de meest recente gemeenschappelijke voorouder van deze groepen. Tegenover apomorfie of afgeleide toestand staat “**plesiomorfie**”, wat op een primitieve toestand slaat. In tegenstelling tot apomorfieën zijn plesiomorfieën niet informatief met betrekking tot cladistische relaties, wat in de volgende voorbeelden geïllustreerd wordt. Deze voorbeelden benadrukken eveneens aan dat apomorfie en plesiomorfie **relatieve begrippen** zijn, die pas een betekenis krijgen wanneer duidelijk gemaakt wordt op welk hiërarchisch niveau ze betrekking hebben.

Voorbeeld 1: zeefvaten en zeefcellen.

Het transport van organische stoffen gebeurt bij de vaatplanten in een gespecialiseerd weefsel, het floëem. Meer specifiek gebeurt dit transport bij de varenachtigen en de gymnospermen via zeefcellen en bij de angiospermen via zeefvaten. Deze laatste zijn opgebouwd uit zeefvatelementen (met de bijbehorende begeleidende cellen). Een zeefvatelement wordt algemeen beschouwd als een sterk gespecialiseerde en afgeleide zeefcel. Men neemt dus aan dat zeefvatelementen (met hun begeleidende cellen) zich in de loop van de evolutie ontwikkeld hebben uit zeefcelachtige cellen. Zeefvatelementen onderscheiden zich van “gewone” zeefcellen ondermeer door het voorkomen van zeefplaten. Verder in de tekst wordt een onderscheid gemaakt tussen zeefcellen in de **enge zin** (met uitsluiting van zeefvaten) en zeefcellen in de **brede zin** (met inbegrip van zeefvaten).

Wanneer we de verwantschap tussen alle families van vaatplanten willen bestuderen, zijn de zeefvaten van de angiospermen apomorf of afgeleid ten opzichte van de zeefcellen die voorkomen bij de verschillende groepen varenachtigen en gymnospermen. Het voorkomen van zeefvaten is dus een bruikbaar argument voor de conclusie dat alle angiospermenfamilies onderling nauwer met elkaar verwant zijn dan met eender welke familie van varenachtigen of gymnospermen. De aanwezigheid van zeefcellen (in de enge zin) daarentegen is op het niveau van de vaatplanten een plesiomorfie en onbruikbaar voor fylogenetische conclusies. Op basis hiervan kan namelijk niet besloten worden dat de varenachtigen en de gymnospermen nauwer met elkaar verwant zijn dan één van hen met de angiospermen.

We kunnen vervolgens het bereik van de analyse verkleinen en ons beperken tot de verwantschappen binnen de monocotylen. Vermits alle angiospermen zeefvaten bezitten, is het voorkomen van zeefvaten op dit niveau plesiomorf en dus onbruikbaar om cladistische verwantschappen te achterhalen.

De situatie is weer verschillend wanneer de oorspronkelijke analyse uitgebreid wordt met de bladmossen, de levermossen en de groenwieren. Op dit niveau (alle groene planten) beschouwen we het zeefvatelement van de angiospermen als een gespecialiseerd type zeefcel en is de aanwezigheid van zeefcellen in de brede zin apomorf ten opzichte van hun afwezigheid. Het voorkomen van zeefcellen is nu dus wel fylogenetisch informatief en kan gebruikt worden om te argumenteren dat alle vaatplanten nauwer met elkaar verwant zijn dan met eender welk mos of groenwier. De afwezigheid van zeefcellen bij bladmossen, levermossen en groenwieren is plesiomorf en kan niet gebruikt worden om bijvoorbeeld te argumenteren dat de levermossen en de bladmossen nauwer met elkaar verwant zijn dan met de vaatplanten.

Zonder aan te geven over welk hiërarchisch niveau men spreekt of ten opzichte waarvan een kenmerktoestand afgeleid of primitief is, hebben uitspraken over al dan niet afgeleid zijn dus duidelijk geen zin. De uitspraak “de aanwezigheid van zeefvaten is afgeleid” is even onvolledig als de uitspraak “Jan is ouder dan”.

Wanneer we er rekening mee houden dat kenmerken convergente evolutie en/of evolutionaire omkeringen (reversies) kunnen vertonen, wordt de situatie complexer. Dit verandert echter niets aan de essentie van de redenering, zoals het volgende voorbeeld aantoont.

<u>Voorbeeld 2</u>: secundair triploïd endosperm

De vorming van secundair triploïd endosperm in de zaden van angiospermen is een karakteristiek die niet aanwezig is bij andere zaadplanten. Op het niveau van alle zaadplanten is de vorming van endosperm dus een apomorfie of een afgeleid kenmerk dat geïnterpreteerd wordt als ontstaan in de gemeenschappelijke voorouder van alle angiospermen (zie echter Friedman 1992). Deze apomorfie kan gebruikt worden om te argumenteren dat de angiospermen onderling nauwer met elkaar verwant zijn dan met eender welke andere groep van zaadplanten. De afwezigheid van secundair endosperm is op dit niveau de plesiomorfe toestand.

De situatie is anders wanneer we bijvoorbeeld de verwantschap tussen de genera van de Orchidaceae zouden willen bestuderen. In sommige genera van deze familie wordt secundair endosperm aangelegd, terwijl in heel wat andere genera geen secundair endosperm gevormd wordt (of zelfs geen dubbele bevruchting plaatsvindt). Op basis van andere kenmerken is er echter geen twijfel mogelijk dat de Orchidaceae monocotylen zijn en dus tot de angiospermen behoren. De afwezigheid van secundair endosperm in sommige geslachten van deze familie wordt daarom beschouwd als een evolutionaire omkering. Op dit hiërarchisch niveau is de aanwezigheid van secundair endosperm dus de plesiomorfe toestand en niet bruikbaar om te argumenteren dat de genera met secundair endosperm nauwer met elkaar verwant zijn dan met de genera zonder. Omgekeerd is de afwezigheid van secundair endosperm op dit niveau apomorf en dus wel fylogenetisch informatief.

Een apomorfie die door een aantal taxa gedeeld wordt, wordt een “**synapomorfie**” van deze groep genoemd; al de taxa die de synapomorfie bezitten vormen een **monofyletische groep**. In het bovenstaande voorbeeld vormen varens, gymnospermen en angiospermen een monofyletische groep en is de aanwezigheid van zeefcellen in de brede zin een synapomorfie die deze groep kenmerkt. Een monofyletische groep wordt beschouwd als de verzameling van alle nakomelingen van de voorouder die de apomorfie ontwikkelde. Monofyletische groepen van verschillende niveaus zijn hiërarchisch gerangschikt (binnen de hogervermelde groep vormen bijvoorbeeld de angiospermen op hun beurt een monofyletische groep van een lager niveau, gekenmerkt door de aanwezigheid van zeefvaten). Zulk een hiërarchie van monofyletische groepen wordt grafisch voorgesteld in een **cladogram**. Een cladogram geeft dus de historische volgorde van de splitsingen tussen groepen weer. Elk vertakkingspunt in een cladogram staat voor de splitsing van een stamgroep in twee dochtergroepen.

Twee monofyletische groepen die rechtstreeks ontstaan zijn uit eenzelfde stamgroep noemt men **zustergroepen** ("*sister groups*"). Zo vormen de groepen (B, C, D) en (E, F) in fig. 1.2 twee zustergroepen; (B) en (C, D) zijn eveneens twee zustergroepen. Uit de definitie van zustergroepen volgt dat elke monofyletische groep met haar zustergroep een monofyletische groep vormt van hogere rang. Wat men in een cladistische analyse wil ontdekken zijn precies deze **zustergroeprelaties**.

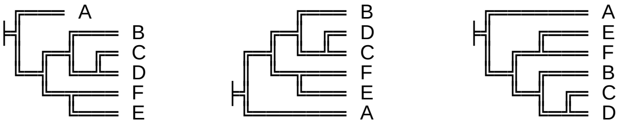
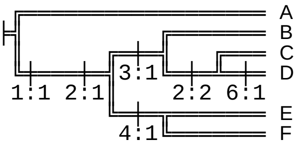


Fig. 1.2. Drie mogelijke voorstellingen van eenzelfde cladogram: de volgorde waarin de zustergroepen geplaatst worden speelt geen enkele rol.

De groep van al de taxa waartussen men de cladistische verwantschappen wil ophelderen wordt de **binnengroep** ("*ingroup*") genoemd. De zustergroep van de binnengroep noemt men de **buitengroep** ("*outgroup*"). Wanneer de gebruikte kenmerken gepolariseerd worden door middel van buitengroepvergelijking (zie 1.4.1), zullen ook de taxa van de buitengroep in de analyse betrokken worden. De term buitengroep wordt dan vaak wat losser gebruikt, namelijk voor eender welk taxon dat nauw verwant is met de binnengroep. De individuele taxa die in de gegevensmatrix opgenomen zijn vormen de **evolutionaire units** of **EU**'s van de analyse (cf. de operationele taxonomische units of OTU's uit de numerieke taxonomie). Omdat deze EU's verschijnen aan de uiteinden van de takken van het cladogram, worden ze vaak de **terminale taxa** genoemd. In een cladistische studie probeert men de zustergroeprelaties binnen de binnengroep op te helderen door de verspreiding van kenmerktoestanden over de bestudeerde taxa te analyseren.

De **lengte van een kenmerk** op een bepaald cladogram is het minimaal aantal overgangen tussen kenmerktoestanden dat volgens het cladogram vereist is om de toestandsverspreiding over de EU's te verklaren. In fig. 1.3 worden enkele voorbeelden gegeven. Rechtsboven staat de verspreiding van zes kenmerken (1-6; de codes 0, 1, en 2 in de matrix staan voor verschillende kenmerktoestanden) over zes taxa (A-F) weergegeven. Op het cladogram linksboven hebben kenmerken 1, 3, 4 en 6 lengte 1, kenmerk 2 heeft lengte 2 en kenmerk 5 lengte 0. De **lengte van een cladogram** behorende bij een bepaalde gegevensmatrix is de som van de lengte van alle kenmerken uit die matrix op dat cladogram. Het cladogram van fig. 1.3 heeft dus lengte 6.



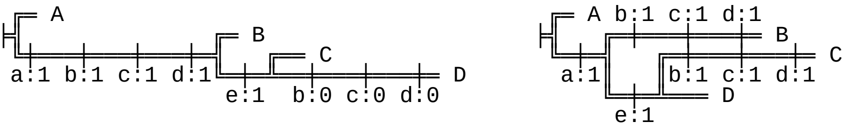
	kenmerk	123456
taxon	A	000000
	B	111000
	C	121000
	D	121001
	E	110100
	F	110100

monofyletische groepen	**synapomorfieën**
BCDEF	1:1; 2:1
BCD	3:1
CD	2:2
EF	4:1

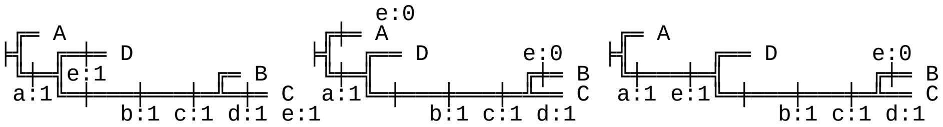
Fig. 1.3. Een voorbeeld van een cladogram voor 6 taxa, gebaseerd op een gegevensmatrix die de verspreiding geeft van de toestanden van 6 kenmerken over deze taxa; A is de buitengroep, B, C, D, E en F vormen samen de binnengroep. De cladistische structuur van de binnengroep volgt uit de matrix met de kenmerkverdelingen (zie tekst); “a:b” staat voor het ontstaan van kenmerktoestand b van kenmerk a.

Voor een bepaald aantal taxa bestaat een groot aantal verschillende cladogrammen. Het cladogram dat in fig. 1.3 getoond wordt, is het kortst mogelijke of **meest spaarzame cladogram** voor de gegeven matrix. In een cladistische analyse van een bepaalde gegevensmatrix aanvaardt men het meest spaarzame cladogram als de meest waarschijnlijke evolutionaire hypothese die gemaakt kan worden aan de hand van de informatie uit die matrix. Andere hypothesen zijn mogelijk, maar deze vereisen meer stappen en zijn dus minder waarschijnlijk (zie ook 1.3). Deze keuze op basis van de minimale lengte is een toepassing van het zogenaamde **parsimonie- of spaarzaamheidscriterium** (“*parsimony criterion*”). Op basis van het dit criterium kiezen we voor het cladogram waarop de globale **congruentie tussen de kenmerken** maximaal is. Dit impliceert eveneens dat de globale **homoplasie** (convergente evolutie + evolutionaire omkeringen) minimaal is. Op het cladogram van fig. 1.3 vertoont geen enkel kenmerk homoplasie en is de congruentie tussen de kenmerken dus 100%.

	a	b	c	d	e	
A	0	0	0	0	0	(buitengroep)
B	1	1	1	1	0	
C	1	1	1	1	1	
D	1	0	0	0	1	



totale lengte: 8; kenmerken b, c en d vertonen elk 1 stap homoplasie op dit cladogram; voor deze kenmerken zijn reversie (links) en convergente evolutie (rechts) even plausibel (een derde mogelijkheid bestaat erin dat toestand 0 van kenmerken b, c en d convergent ontstaan is in taxa A en D)



totale lengte: 6; enkel kenmerk e vertoont homoplasie (1 stap) op dit cladogram; de verdeling van de toestanden van het kenmerk kan op verschillende manieren verklaard worden

Fig. 1.4. Lengte van twee cladogrammen (onder) voor de gegevensmatrix (boven) voor vier taxa. A fungeert als buitengroep (zie verder); het onderste cladogram is het meest spaarzame.

Volledige congruentie tussen verschillende kenmerken treedt in de praktijk zelden op. Een realistischer voorbeeld wordt gegeven in fig. 1.4. In deze figuur werden de kenmerken uit de gegevensmatrix (boven) uitgezet op twee verschillende cladogrammen. Het onderst cladogram is spaarzamer dan het bovenste doordat op dit cladogram slechts één kenmerk (kenmerk e) een toestandsverdeling heeft die niet congruent is met de groepen zoals ze door het cladogram gespecifieerd worden. Het bovenste cladogram heeft 3 zulke kenmerken (b, c, en d). Het voorbeeld toont eveneens aan dat het herkennen van homoplasie op zich niet altijd volstaat om de oorzaak van deze homoplasie te achterhalen. Zo kan de verdeling van de toestanden van kenmerken b, c, en d op het bovenste cladogram net zo goed verklaard worden door reversie als door convergente evolutie. Ook wat betreft de evolutie van kenmerk e op het meest spaarzame cladogram zijn er meerdere mogelijkheden. De eerste mogelijkheid (links) impliceert dat taxa A en B toestand 0 overgeërfd hebben van een gemeenshappelijk voorouder, en bijgevolg dat toestand 1 convergent ontstaan is in taxa C en D; de tweede mogelijkheid (midden) impliceert dat taxa C en D toestand 1 overgeërfd hebben van een gemeenschappelijk voorouder, en bijgevolg dat toestand 0 convergent ontstaan is in taxa A en B. In het derde is toestand 1 ontstaan in de gemeenschappelijke voorouder van taxa B, C en D, en nadien is er in taxon B reversie opgetreden. Uit deze voorbeelden blijkt dat de gegevens uit een matrix op zich niet steeds volstaan om voor elk kenmerk ondubbelzinnig te bepalen welke toestand plesiomorf in de binnengroep.

De graad van homoplasie van een kenmerk op een bepaald cladogram wordt vaak uitgedrukt door de **consistentie-index ci** of door **de retentie-index ri** (Kluge & Farris 1969, Farris 1989; zie ook Archie 1996). De consistentie-index is de verhouding van de lengte van het kenmerk indien het geen homoplasie zou vertonen (m) en de werkelijke lengte op het cladogram (s); ci bereikt zijn maximale waarde 1 wanneer het kenmerk geen homoplasie vertoont en wordt kleiner naarmate de homoplasie toeneemt. In de retentie-index wordt ook rekening gehouden met de maximale homoplasie die een kenmerk op eender welk cladogram kan vertonen. Deze maximale waarde is (g-m), met g de maximale lengte die het kenmerk op eender welk cladogram kan vertonen (zo is bijvoorbeeld g voor kenmerk 3 van fig. 1.3 gelijk aan 3: toestand 1 van taxa B, C, en D kan in het slechtste geval immers drie keer onafhankelijk ontstaan zijn). De retentie-index is dan gedefinieerd als (g-s)/(g-m). Ook ri bereikt zijn maximale waarde 1 wanneer er geen homoplasie aanwezig is. De homoplasie van een volledige gegevensmatrix op een cladogram wordt vaak uitgedrukt door de **globale consistentie-index CI** (*ensemble consistency index*; CI = $\sum m / \sum s$; de sommatie gebeurt over alle kenmerken) of door de **globale retentie-index RI** (*ensemble retention index*; RI = $\sum(g-s)/\sum(g-m)$).

Een plesiomorfie die door een aantal taxa gedeeld wordt, wordt een **symplesiomorfie** van deze groep genoemd. Kenmerktoestand 5:0 in fig. 1.3 is een voorbeeld van een symplesiomorfie. Deze toestand was reeds aanwezig in de gemeenschappelijke voorouder van A, B, C, D, E en F en is bijgevolg niet informatief op het niveau van de onderlinge relaties tussen deze taxa. Een symplesiomorfie is eigenlijk een synapomorfie die op een te laag taxonomisch niveau beschouwd wordt. Enkele symplesiomorfieën van de angiospermen zijn bijvoorbeeld het bezit van mitochondria, het voorkomen van tracheïden, de aanwezigheid van chlorofyl en het voorkomen van zaden.

Een **autoapomorfie** (**autapomorfie**) duidt op de unieke aanwezigheid van een apomorfie in slechts één enkel taxon uit de analyse (bijvoorbeeld kenmerktoestand 6:1 voor taxon D in fig. 1.3). Autoapomorfieën zijn net als symplesiomorfieën niet bruikbaar voor het achterhalen van verwantschappen. Om informatief te zijn moet een kenmerktoestand immers in minstens twee EU's voorkomen. Dit betekent echter niet dat autoapomorfieën totaal onbelangrijk zijn. Autoapomorfieën zijn immers aanwijzingen dat deze EU's elk op zich monofyletisch zijn, en dit is een belangrijke randvoorwaarde, zoals het volgende voorbeeld aantoont. Men zou de families van de angiospermen op basis van de eerste letter van hun naam kunnen indelen in 26 suprafamiliale taxa. Louter technisch is er geen probleem om de “verwantschappen” tussen deze alfabetische groepen cladistisch te analyseren, maar aangezien de EU’s in dit geval hoogst artificiële groepen zijn, is een dergelijke analyse absurd. De vraag of de EU’s al dan niet monofyletisch zijn, is dan ook één van de eerste vragen die men zich moet stellen wanneer men een gegevensmatrix wil opstellen. Strikt genomen staat de monofylie van de EU’s tijdens de eigenlijke analyse niet meer ter discussie.

Ook de vraag of de groep van alle EU’s samen een monofyletische groep vormt, valt buiten de analyse in de strikte zin. Zo is het bijvoorbeeld perfect mogelijk om een cladistische analyse uit te voeren van een gegevensmatrix voor alle angiospermenfamilies waarvan de naam met een A begint. Het feit dat het technisch mogelijk is om hiervoor cladogrammen op te stellen, betekent echter niet dat men mag besluiten dat deze families een monofyletische groep zouden vormen binnen de angiospermen. Dit zou men enkel kunnen concluderen op basis van synapomorfieën die deze groep families onderscheidt van alle andere families.

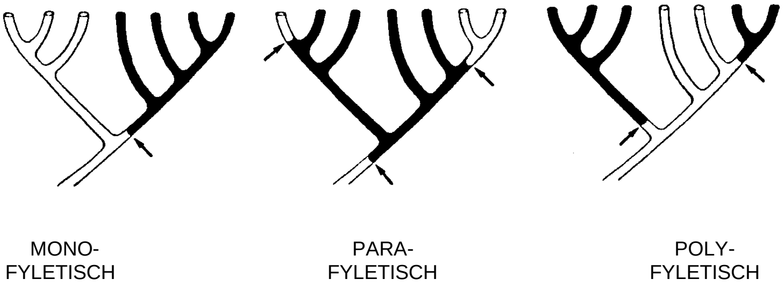


Fig. 1.5. Een voorbeeld van een mono-, een para- en een polyfyletische groep (naar Dahlgren 1983b).

Naast monofyletische groepen, kan men ook parafyletische en polyfyletische groepen onderscheiden (fig. 1.5). Een **parafyletische groep** is een groep die gekarakteriseerd wordt door plesiomorfe kenmerktoestanden (Farris 1991). Zo kan men bijvoorbeeld binnen de monofyletische vaatplanten een parafyletische groep “varenachtigen + progymnospermen” (zie fig. 1.6) afbakenen op basis van het ontbreken van de kenmerktoestand zaadvorming. Deze kenmerktoestand, een levenscyclus zonder zaden, is binnen de vaatplanten plesiomorf ten opzichte van een levenscyclus met zaden. Een parafyletische groep is dus eigenlijk een monofyletische groep waaruit een aantal takken die de apomorfe toestand van een kenmerk ontwikkeld hebben, verwijderd zijn. Andere bekende voorbeelden zijn de groenwieren en de sporeplanten. **Polyfyletische groepen** zijn groepen die gekarakteriseerd worden door convergente kenmerktoestanden (Farris 1991). Mogelijke voorbeelden hiervan op het niveau van de angiospermen zijn de groepering van al de parasitaire bloemplanten, of alle soorten met blauwe bloemen, al de waterplanten enz.

Ook de begrippen "*grade*" en "*clade*" staan hiermee in verband. "*Grades*" zijn groepen die gekarakteriseerd worden door een bepaald ontwikkelingsniveau, dat vaak begrepen kan worden als een aanpassing aan een specifiek milieu. Het betreft dan ook vaak artificiële, niet-monofyletische groepen, zoals bijvoorbeeld de mossen (parafyletisch) of alle parasitaire planten (polyfyletisch). "*Clades*" daarentegen zijn per definitie monofyletische groepen.

De term "monofylie" werd al lang voor het cladisme zich ontwikkeld had, gebruikt om groepen aan te duiden die volgens de bovenstaande definities ofwel mono- ofwel parafyletisch zijn. Monofyletische groepen werden dan aangeduid als holofyletisch; voor parafyletische groepen bestond geen aparte term. Dit terminologisch verschil heeft in de jaren '70 vaak tot misverstanden en steriele discussies geleid tussen voor- en tegenstanders van het cladisme. Sporadisch wordt in de literatuur ook over convexe groepen gesproken. Dit zijn eveneens groepen die volgens bovenstaande definities ofwel mono- ofwel parafyletisch zijn.

Ter illustratie volgen nu enkele concrete voorbeelden. Een eerste is een geschematiseerde cladistische voorstelling van de klassieke visie op de relaties binnen de landplanten (fig. 1.6; naar Crane 1985).

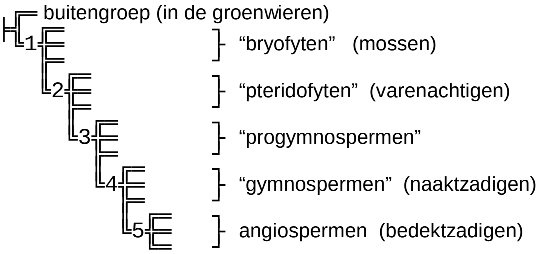

buitengroep (in de groenwieren)
1
"bryofyten" (mossen)
2
"pteridofyten" (varenachtigen)
3
"progymnospermen"
4
"gymnospermen" (naaktzadigen)
5
angiospermen (bedektzadigen)

Fig. 1.6. Cladogram van de embryofyten (naar Crane 1985); de parafyletische groepen staan tussen aanhalingstekens.

De volgende groepen zijn hierin monofyletisch:

(1) Embryofyten

synapomorfie: de zygote produceert een multicellulair embryo dat zijn vroege ontwikkeling reeds doormaakt in het archegonium of in de embryozak;

(2) Tracheofyten

synapomorfie: de aanwezigheid van tracheïden met secundaire celwandverdikkingen;

(3) deze groep bezit nog geen naam

synapomorfie: de aanwezigheid van secundair xyleem en floëem;

(4) Spermatofyten

synapomorfie: vorming van zaden;

(5) Angiospermen

synapomorfie: dubbele bevruchting met vorming van secundair endosperm.

Bryofyten, pteridofyten, progymnospermen en gymnospermen zijn parafyletisch.

Het tweede voorbeeld (Weynants 1993) behandelt de cladistische verwantschappen van de Primulanae (sensu Smets 1988a) op familieniveau. Deze superorde van de angiospermen omvat twee ordes, de Primulales en de Ebenales. Naast de vijf families van de Primulales (Theophrastaceae, Myrsinaceae, Aegicerataceae, Primulaceae en Coridaceae) en de vier families van de Ebenales (Sapotaceae, Ebenaceae, Lissocarpaceae en Styracaceae) werden ook de Symplocaceae in de analyse betrokken. Deze familie (in de Cornales sensu Smets 1988) wordt door sommige auteurs in de Ebenales geplaatst en fungeert hier als buitengroep. De gegevensmatrix werd opgesteld aan de hand van gegevens uit de literatuur en omvat 66 morfologische, anatomische en chemische kenmerken.

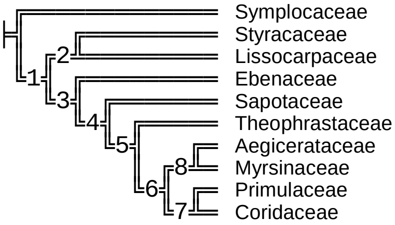


Fig. 1.7. Het meest spaarzame cladogram voor de Primulanae (lengte 115; Weynants 1993).

Het meest spaarzame cladogram voor deze matrix (fig. 1.7) heeft lengte 115. Op basis van dit cladogram kunnen we besluiten dat de Primulales een monofyletische orde vormen, terwijl de Ebenales parafyletisch zijn. In dezelfde analyse werd ook nagegaan hoeveel cladogrammen er waren met één stap meer (dus lengte 116). Dit waren er drie, die grotendeels overeenkwamen met het kortste cladogram. In fig. 1.8 wordt het **strikte consensuscladogram** van deze vier cladogrammen gegeven. In een consensuscladogram worden de overeenkomsten tussen een aantal verschillende cladogrammen weergegeven (zie Page 1993 voor de verschillende manieren waarop dit kan gebeuren).

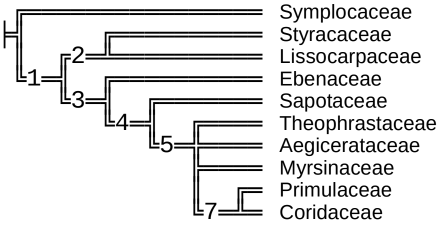


Fig. 1.8. Strikt consensusdiagram van alle bomen met lengte 115 en 116 (Weynants 1993).

Het hiërarchisch niveau waarop een bepaalde analyse plaatsvindt, beperkt de mogelijke vragen waarop de analyse een antwoord kan vinden. In de bovenstaande analyse van de Primulanae zijn bijvoorbeeld de Coridaceae een monotypische familie (*Coris* als enige geslacht), die door meerdere auteurs opgenomen wordt in de Primulaceae sensu lato (hetzelfde geldt voor de monotypische Aegicerataceae ten opzichte van de Myrsinaceae). De bovenstaande analyse kan echter geen antwoord geven op de vraag of *Coris* al dan niet de zustergroep is van de Primulaceae sensu stricto en bijgevolg eventueel als familie mag worden erkend. Door de Primulaceae sensu stricto in deze analyse als EU te aanvaarden, maken we immers impliciet de hypothese dat de Primulaceae sensu stricto monofyletisch zijn. De mogelijkheid bestaat echter dat dit niet zo is en dat de Primulaceae zonder *Coris* een parafyletische groep vormen. Of dit effectief zo is kan enkel uitgemaakt worden door een cladistische analyse op een lager taxonomisch niveau uit te voeren. Een mogelijkheid is bijvoorbeeld om de verschillende genera van de Primulaceae samen met *Coris* te analyseren. Op basis van de bovenstaande analyse kunnen we als buitengroep(en) enkele genera van de Theophrastaceae en van de Myrsinaceae sensu lato toevoegen (zie verder: buitengroepvergelijking). Op die wijze spelen analyses op verschillende hiërarchische niveaus op elkaar in.

1.3 Cladisme in een breder kader: het homologieconcept

Het homologieconcept is een centraal concept in systematisch onderzoek. De introductie van de term “homologie” wordt gewoonlijk toegeschreven aan de Britse anatoom en paleozoöloog Richard Owen (1804-1892), maar de wortels van het concept achter de term kunnen zonder veel moeite gevolgd worden tot in de achttiende eeuw. Owen definieerde een homologie ("*homologue*") als "*the same organ in different animals under every variety of form and function*". Oorspronkelijk was het een zuiver morfologisch concept, maar na de doorbraak van het evolutionair denken in de tweede helft van de vorige eeuw was het snel duidelijk dat homologie alles te maken had met gemeenschappelijk afkomst. Sindsdien is er bijna ononderbroken een controverse geweest rond de vraag of homologie een morfologisch (in de breedste zin van het woord) dan wel een evolutionair concept is of zou moeten zijn. Al naargelang de aspecten die men het belangrijkst vindt of wil benadrukken, bestaan er meerdere mogelijkheden om deze vraag te benaderen (zie Donoghue 1992 en Hall 1994 voor een overzicht).

Een mogelijk antwoord op deze vraag vertrekt van het standpunt dat het homologieconcept beide aspecten in zich verenigt. Elke vergelijkende biologische studie is immers een procedure die uit twee onafhankelijke en complementaire stappen bestaat. In de eerste fase staat het **genereren** van homologiehypothesen centraal. Deze primaire hypothesen worden in de tweede fase **getest** op hun algemeenheid (Rieppel 1988, de Pinna, 1991).

In de eerste fase van een vergelijkend onderzoek gaat men op zoek naar gelijkenissen waarvan men op basis van vergelijkend morfologisch-anatomisch, ontogenetisch en/of ander onderzoek kan vermoeden dat ze op gemeenschappelijke afkomst wijzen. Zulke gelijkenissen worden **primaire homologieën** genoemd (de Pinna 1991). Twee regelmatig gebruikte synoniemen zijn **similariteit** (bijvoorbeeld Reeck et al., 1987) of **topografische correspondentie** (Rieppel 1988).

De criteria die in de eerste fase gehanteerd worden, worden in de literatuur vaak de **homologiecriteria** genoemd. Deze zijn niet absoluut, maar fungeren als algemene richtlijnen die helpen om goed gekarakteriseerde van minder goed gekarakteriseerde kenmerken te onderscheiden. Hieronder vallen bijvoorbeeld de drie klassieke hoofdcriteria van Remane (1952). Het eerste hiervan is het **positiecriterium** ("*das Kriterium der Lage*"). Dit criterium zoekt naar gelijkenissen op basis van de positie van een structuur binnen een gepaald grondplan, zoals bijvoorbeeld de positie van een nektarklier in een pentamere tetracyclische bloem, of de positie van een bepaald nucleotide in de sequentie van het *rbc*L-gen. Hogerop hebben we zeefvaten gekarakteriseerd als cellen waardoor transport van organische stoffen plaatsvindt en die bij volledige differentiatie zeefplaten bezitten. Deze karakterisatie kan gezien worden als een toepassing van Remane's tweede criterium, het criterium van **speciale eigenschappen** ("*das Kriterium der speziellen Qualität*"). Remane's derde criterium, dat van de **overgangsvormen** ("das *Stetigkeitskriterium*"), kan meestal herleid worden tot één van de twee vorige criteria.

Soms wordt ook het **conjunctiecriterium** gebruikt. Dit criterium stelt dat een structuur a die voorkomt in een bepaald taxon niet homoloog kan zijn met een structuur b in een ander taxon indien er taxa bestaan die zowel structuur a als structuur b bezitten.

Bij het tot stand komen van primaire homologiehypothesen (het definiëren van goede kenmerken) mogen de mogelijke evolutionaire verwantschappen tussen de bestudeerde organismen geen rol spelen. Het is immers precies de functie van primaire homologieën om, gebaseerd op intrinsieke eigenschappen van de bestudeerde structuren, dergelijke evolutionaire hypothesen te genereren. Zo is in ons voorbeeld met betrekking tot de zeefvaten de primaire evolutionaire hypothese dat alle planten die zeefvaten bezitten nakomelingen zijn van eenzelfde meest recente voorouder. Binnen de vaatplanten bezitten enkel de angiospermen zeefvaten, maar als we de analyse uitbreiden tot alle planten, dan blijkt dat ook bij heel wat bruinwieren de organische stoffen getransporteerd worden via cellen die zeefplaten bezitten. Indien we de bovenstaande karakterisatie aanvaarden, moeten we dit eveneens zeefvaten noemen. De primaire homologiehypothese is dan dat deze bruinwieren samen met alle angiospermen afstammen van eenzelfde meest recente gemeenschappelijke voorouder.

Of een dergelijke primaire hypothese al dan niet correct is, wordt getest in de tweede fase van vergelijkend onderzoek. Dit testen gebeurt niet door kenmerken individueel te gaan bekijken, zoals in de eerste fase, maar door een groot aantal primaire homologieën tegelijkertijd te vergelijken. Dit is precies wat er gebeurt in een cladistische analyse die als doel heeft om de **congruentie** tussen zoveel mogelijk primaire homologieën te maximaliseren. In een dergelijke analyse zal blijken dat er veel meer kenmerken zijn die de angiospermen met de overige landplanten verenigen dan met de bruinwieren. Op basis van deze **congruentietest** besluiten we dan dat zeefvaten in de loop van de evolutie van de planten tweemaal ontstaan zijn: éénmaal binnen de bruinwieren, en een tweede maal in de gemeenschappelijke voorouder van de angiospermen. Dit is een voorbeeld van convergente evolutie.

Merk op dat de congruentietest een primaire homologiehypothese nooit volledig verwerpt: de congruentietest zal ofwel de primaire homologie op het oorspronkelijk niveau van algemeenheid bevestigen (wanneer het kenmerk geen homoplasie vertoont op het cladogram), ofwel dat niveau vervangen door twee of meer lagere niveaus van algemeenheid (het kenmerk vertoont wel homoplasie). In het voorbeeld van de zeefvaten is de oorspronkelijk primaire hypothese omgezet in twee *secundaire hypothesen*: (1) binnen de landplanten zijn alle zeefvaten homoloog en (2) binnen de bruinwieren zijn alle “zeefvaten” homoloog. Het feit dat het oorspronkelijke niveau van algemeenheid (alle planten) verworpen wordt, betekent evenmin dat het kenmerk al zijn waarde verliest. Binnen de landplanten blijft de aanwezigheid van zeefvaten in de angiospermen immers een synapomorfie die deze groep onderscheidt van alle andere landplanten. In deze benadering van het homologieconcept worden **secundaire homologieën** dus gelijkgesteld met synapomorfieën.

Tussen het onderzoek van individuele kenmerken (de eerste fase) en de analyse van de congruentie tussen verschillende kenmerken (de tweede fase) bestaat een constante wisselwerking. Deze wisselwerking werd door Hennig **wederzijdse opheldering** ("*reciprocal illumination*" of "*reciprocal clarification*") genoemd. Zo kan men na het uitvoeren van de congruentietest de "zeefvaten" van de bruinwieren en de zeefvaten van de angiospermen aan een nieuw morfologisch onderzoek onderwerpen. Indien deze studie tot nieuwe structurele informatie leidt, kan een hernieuwde en verfijnde karakterisering voorgesteld worden die dan op haar beurt kan gebruikt worden in nieuwe cladistische analyses.

Een cladogram is in essentie gewoon een hiërarchische samenvatting van de verspreiding van kenmerktoestanden; het spaarzaamheidscriterium zorgt hierbij voor maximale congruentie tussen verschillende kenmerken. Dit resulteert in de efficiëntste hiërarchische samenvatting van de gegevens, in de zin dat het meest spaarzame cladogram ervoor zorgt dat de grootste hoeveelheid primaire homologieën de congruentietest doorstaat. Deze hiërarchie wordt vervolgens evolutionair geïnterpreteerd, waarbij we aannemen dat evolutie heeft plaatsgevonden en tot een hiërarchisch patroon leidt (merk op dat dit niet altijd opgaat; bij planten komt bijvoorbeeld frequent hybridisatie voor, en door hybridisatie ontstaat er veeleer een netwerk in plaats van een hiërarchie; zie hiervoor bijvoorbeeld Funk 1985). Het cladisme is dus een procedure om de hiërarchische structuur van de levende wereld te **ontdekken** en te **beschrijven**, terwijl de evolutietheorie dit patroon **verklaart**.

Meermaals werd naar de geschiedenis van de systematiek verwezen om deze visie te onderstrepen (bijvoorbeeld Brady 1985, 1994): reeds in de achttiende eeuw had men proefondervindelijk ontdekt dat de belangrijkste patronen in de levende wereld hiërarchische patronen waren. Deze hiërarchische structuur was een eigenschap van de levende wereld die men voor 1859, het jaar waarin Darwin's "*Origin of species*" verscheen, moeilijk kon verklaren. Een hiërarchische structuur volgt daarentegen op een heel natuurlijke wijze uit Darwin's theorie van "*descent with modification*". Darwin beschouwde dit trouwens als één van de sterkste argumenten die hij voor zijn ideeën naar voren kon brengen. Dat men de hiërarchische structuur van de levende wereld kon ontdekken los van de verklarende evolutietheorie, hoeft niet te verwonderen. Een classificatie van een bepaalde groep organismen is immers niet iets dat volgt uit de kennis van de evolutie van die groep, maar uit een studie van de kenmerken van die groep. Zo zijn bijvoorbeeld convergenties en evolutionaire omkeringen systematische conclusies (die kunnen afgelezen worden van een cladogram), veeleer dan "feiten" waarop systematische conclusies gebaseerd zouden moeten zijn.

1.4 Polariseren en ordenen van kenmerken

Het maken van dit onderscheid tussen apomorfe en plesiomorfe kenmerktoestanden noemt men het **polariseren** van een kenmerk. Het hoeft weinig betoog dat dit een essentieel punt is: de cladistische analyse is er immers op gericht om groepen te ontdekken die door synapomorfieën gekenmerkt worden. In de loop van de jaren werden heel wat polarisatiecriteria voorgesteld en besproken. De meeste hiervan zijn echter onhoudbaar gebleken of te herleiden tot enkele basiscriteria (zie bijvoorbeeld de Jongh 1980 of Stevens 1980 voor een overzicht).

Momenteel erkent men twee fundamentele criteria: **buitengroepvergelijking** ("*outgroup comparison*"; zie Nixon & Carpenter 1993 voor een recente discussie) en het **ontogenetisch criterium** (zie Weston 1988). Van deze twee wordt in de praktijk voornamelijk buitengroepvergelijking toegepast. Vooraleer we deze criteria bespreken, gaan we even in op twee andere polarisatiemethodes die vaak opduiken in de literatuur.

De eerste, gebaseerd op **fossielen**, kan mits de nodige omzichtigheid in sommige omstandigheden toch bruikbaar zijn. Een mogelijke formulering is als volgt: "wanneer er twee fossielen bestaan die elk een andere toestand van een kenmerk bezitten en wanneer beide fossielen sterk in ouderdom verschillen, is de toestand in het oudste fossiel de plesiomorfe toestand". De volgende variante maakt een vergelijking tussen fossielen en nu levende organismen: "wanneer een fossiel en een nu levend organisme een verschillende toestand van een kenmerk bezitten, is de toestand in het fossiel plesiomorf." Op het eerste gezicht zijn deze regels heel plausibel, maar er zijn meerdere factoren die de juistheid van de redenering kunnen beïnvloeden. In de eerste plaats is men nooit zeker of de vroegste vertegenwoordigers met een bepaalde kenmerktoestand ook gefossiliseerd zijn. En zelfs indien dit zo is, moeten deze fossielen ook nog gevonden worden. Verder is het eveneens mogelijk dat een bepaald fossiel weliswaar jong is, maar afkomstig is van een soort die vrij veel plesiomorfe kenmerken bewaard heeft (cf. "levende fossielen"); anderzijds kunnen heel oude fossielen vertegenwoordigers zijn van sterk gespecialiseerde evolutionaire lijnen. Deze overwegingen weerspiegelen gewoon dat elke evolutionaire lijn op elk moment een mengeling van afgeleide en primitieve kenmerken bezit. Dit verschijnsel wordt soms aangeduid met de term **heterobatmie**

Een tweede criterium, dat vooral in de vroege cladistische literatuur gehanteerd werd, is het "***common is primitive***" principe. Dit principe stelt dat de kenmerktoestand die het vaakst voorkomt in de binnengroep de plesiomorfe toestand is (mogelijke varianten van dit principe bekijken het voorkomen in buitengroep + binnengroep of uitsluitend in de buitengroep). Dit principe, in al zijn varianten, is totaal verschillend van buitengroepvergelijking, waarmee het vroeger vaak verward werd. Het is vrij gemakkelijk om aan te tonen dat een strikte toepassing van "*common is primitive*" tot foute resultaten leidt. Stel dat je bijvoorbeeld de verwantschap tussen drie taxa bestudeert. De enige kenmerken die voor een dergelijk drietaxonprobleem relevant zijn, zijn synapomorfieën die in twee van de drie soorten voorkomen. Maar volgens het "*common is primitive*" principe zijn alle toestanden die in twee van de drie taxa voorkomen plesiomorf, zodat je zou moeten besluiten dat er enkel autoapomorfieën voorkomen en elk drietaxonprobleem in principe onoplosbaar is. Meer concreet zou toepassing van dit principe binnen de angiospermen bijvoorbeeld kunnen leiden tot de hypothese dat cyclisch ingeplante bloemorganen primitief zijn en spiralig ingeplante afgeleid; binnen de zaadplanten zou dubbele bevruchting primitief zijn; toegepast op alle levende wezens zou de eukaryote celstructuur primitiever zijn dan de prokaryote, ...

1.4.1 Buitengroepvergelijking

In een vereenvoudigde versie kan dit criterium als volgt geformuleerd worden: "de kenmerktoestand die plesiomorf is voor een gegeven binnengroep is die toestand die ook voorkomt in de zustergroep van die binnengroep (= de buitengroep)". Deze vereenvoudigde formulering gaat echter niet steeds op, zoals hogerop reeds geïllustreerd werd (kenmerk e op het onderste cladogram van fig. 1.4). Practisch gezien komt het uitvoeren van buitengroepvergelijking erop neer dat aan de gegevensmatrix voor de binnengroep ook gegevens toegevoegd worden van één of meerdere buitengroepen. De uitgebreide matrix wordt dan onderworpen aan een parsimonieanalyse waarbij de globale congruentie over binnen- en buitengroep samen gemaximaliseerd wordt. Het resulterende cladogram wordt dan zo afgebeeld dat de tak die binnen- en buitengroep verbindt aan de basis ligt van de binnengroep. Nadat het cladogram op die wijze geöriënteerd is, kan het gebruikt worden om erop af te lezen welke kenmerktoestanden in de binnengroep apomorf of plesiomorf zijn.

Buitengroepvergelijking is theoretisch heel goed gefundeerd, maar heeft als practisch nadeel dat de buitengroep bekend moet zijn, waarvoor een cladistische analyse op een hoger taxonomisch niveau vereist is (waar dan weer de buitengroep voor dat hoger niveau bekend moet zijn...). In de praktijk zal men vaak analyses uitvoeren zonder dat het “echt” vaststaat wat de buitengroep is (het blijft immers steeds een hypothese). Er kunnen bijvoorbeeld meerdere kandidaat-buitengroepen zijn, wat meer regel dan uitzondering is.

Wanneer kandidaat-buitengroepen moeilijk te bepalen zijn, gebruikt men vaak een **hypothetische voorouder** (“*hypothetical ancestor*”). Dit is een fictief taxon dat wordt toegevoegd aan de matrix met de gegevens van de binnengroep. De kenmerktoestanden die toegekend worden aan dit taxon zijn die toestanden waarvan men op basis van het voorkomen in de mogelijke buitengroepen vermoedt dat ze plesiomorf zijn voor de binnengroep. Dit hoeft niet noodzakelijk voor alle kenmerken te gebeuren. De hypothetische voorouder wordt soms ook de **synthetische buitengroep** (“*synthetic outgroup*”) genoemd.

1.4.2 Het ontogenetisch criterium

Dit criterium legt een verband tussen ontogenie en fylogenie. Een mogelijke formulering is als volgt: “wanneer er gedurende de ontogenie een overgang waargenomen kan worden tussen twee toestanden van een kenmerk, dan is de minst algemeen voorkomende toestand apomorf, de meest algemeen voorkomende plesiomorf”. Dit criterium kan zowel toegepast worden op de ontogenie van een volledig organisme (bijvoorbeeld zaad - kiemplant - vegetatieve groei - generatieve groei - afsterven), als op de ontogenie van individuele organen (de organogenese).

Het volgende voorbeeld (uit Weston 1988) toont aan hoe dit criterium correct gebruikt kan worden. Bij sommige soorten van het geslacht *Acacia* (Fabaceae) worden er gedurende heel de levensloop samengestelde bladeren gevormd. Bij de andere soorten komen enkel in het kiemplantstadium samengestelde bladeren voor, terwijl de volwassen planten enkelvoudige fylloden dragen. Vermits (1) alle soorten die in volwassen toestand fylloden dragen samengestelde bladeren bezitten in hun kiemplantstadium en (2) er soorten zijn die uitsluitend samengestelde bladeren hebben, is het bezit van samengestelde bladeren binnen *Acacia* meer algemeen voorkomend dan het bezit van fylloden. Samengestelde bladeren zijn bijgevolg plesiomorf en fylloden apomorf. Merk op dat deze polarisatie onafhankelijk is van het aantal soorten met fylloden: ook al zouden 99% van alle *Acacia*-soorten fylloden dragen in volwassen toestand, het bezit van fylloden blijft een apomorfie volgens dit criterium (terwijl het volgens “*common is primitive*” dan een plesiomorfie zou zijn).

Zoals het voorbeeld aantoont, is één van de voordelen van het ontogenetisch criterium dat de buitengroep niet bekend hoeft te zijn: de vergelijking gebeurt volledig binnen de binnengroep. Daarom wordt dit ook een **direct** criterium genoemd. Buitengroepvergelijking is daarentegen een **indirecte** methode.

Hoewel ontogenetisch onderzoek in plantensystematisch onderzoek een belangrijke rol speelt om hypothesen te ontwikkelen over primaire homologieën tussen bepaalde structuren, wordt ontogenie als polarisatiecriterium maar heel weinig gebruikt. De reden is heel eenvoudig: gevallen zoals het bovenstaande voorbeeld waar het criterium zonder al te grote problemen toegepast kan worden, zijn heel schaars.

1.4.3 Polariseren versus ordenen

Tot nu toe hebben we nagenoeg uitsluitend **tweetoestandskenmerken** of **binaire kenmerken** beschouwd. Hiernaast bestaan ook **veeltoestandskenmerken** ("*multistate characters*") of kenmerken met meer dan twee toestanden. De polarisatie verloopt hier net zoals bij binaire kenmerken, maar er is een bijkomend probleem: wat is de relatie tussen de overige toestanden? Meestal worden deze op basis van bijvoorbeeld morfologische informatie in een **toestandsboom** ("*character state tree*") geplaatst (de termen **morfocline, transformatieserie** en **semofyletische reeks** zijn min of meer synoniemen hiervoor). Het opstellen van zulk een toestandsboom noemt men het **ordenen** van een kenmerk. Hierdoor wordt vastgelegd welke directe overgangen tussen de kenmerktoestanden aanvaard worden en welke verworpen.

Een concreet voorbeeld is het kenmerk "transport van organische stoffen" in de landplanten (fig. 1.9a). Wanneer we veronderstellen dat de zustergroep van de landplanten groenwieren zijn, vinden we met buitengroepvergelijking dat de plesiomorfe toestand in de landplanten "transport via relatief ongespecialiseerde cellen" is. Indien we nu op basis van bijvoorbeeld morfologische, anatomische of ontogenetische informatie kunnen aantonen dat zeefvaten opgebouwd zijn uit een gespecialiseerd soort zeefcellen, kunnen we de toestanden ordenen. Dit betekent dat we de hypothese maken dat zeefvaten in de loop van de evolutie nooit rechtstreeks uit relatief ongespecialiseerde cellen ontstaan zijn, maar altijd uit zeefcellen. Deze zeefcellen hebben zich in de loop van de evolutie op hun beurt uit relatief weinig gespecialiseerde cellen ontwikkeld. Dit impliceert dus dat een overgang van toestand 0 naar toestand 2 op een cladogram niet 1 maar 2 stappen zal bijdragen tot de lengte van het cladogram. Zo worden de resultaten van het vergelijkend morfologisch-anatomisch onderzoek geïntegreerd in de cladistische analyse. Geordende veeltoestandskenmerken kunnen altijd voorgesteld worden als een reeks van binaire kenmerken (figs. 1.9b en 1.9c). Dit wordt **binair additief coderen** ("*binary additive coding*") genoemd.

(0) via relatief ongespecialiseerde cellen
(1) via zeefcellen in de enge zin
(2) via zeefvaten

1.9.a. De veeltoestandsvorm: 1 kenmerk met 3 toestanden

(0) via ongespecialiseerde cellen
(1) via zeefcellen in de brede zin

(0) niet via zeefvaten
(1) via zeefvaten

1.9.b. Binair additief: twee binaire kenmerken

a	b
0	00
1	10
2	11

De binaire combinatie 01 wordt op basis van het kenmerken-onderzoek uitgesloten (zie tekst).

1.9.c. Correspondentie tussen beide voorstellingswijzen

Fig. 1.9. Twee evenwaardige voorstellingen van hetzelfde geordend veeltoestandskenmerk “transport van organische stoffen.

In het theoretisch voorbeeld van fig. 1.10 worden de zestien verschillende mogelijkheden opgesomd waarop een willekeurig kenmerk met vier toestanden kan geordend worden (merk op dat toestandsbomen ook vertakt kunnen zijn). Vermits elk van deze ordeningen op vier verschillende mogelijkheden gepolariseerd kan worden, zijn er in totaal 64 verschillende combinaties van ordening en polarisatie. Elk van deze combinaties weerspiegelt een verschillende visie op de evolutie van het kenmerk.

```
B   C  A   C  A   B  A   B
 \ /    \ /    \ /    \ /
  A      B      C      D
  |      |      |      |
  D      D      D      C
```

A-B-C-D	A-B-D-C	A-D-B-C
A-C-B-D	A-C-D-B	A-D-C-B
B-A-C-D	B-A-D-C	B-C-A-D
C-A-B-D	C-A-D-B	C-B-A-D

1.10.a. Alle mogelijke ordeningen van een kenmerk met vier toestanden.

```
A   B  A   B  A   B  A   B
 \ /    \ /    \ /    \ /
  C      C      C      C
  |      |      |      |
  D      D      D      D

A-B-D-C    A-B-D-C    A-B-D-C    A-B-D-C
```

1.10.b Alle mogelijke polarisaties voor twee van de bovenstaande ordeningen
(de onderlijnde toestand is plesiomorf)

Fig. 1.10. Ordenen en polariseren van kenmerken.

Wanneer men anderzijds een veeltoestandskenmerk heeft waarvoor nog geen uitgebreid kenmerkenonderzoek plaats heeft gevonden, kan men beslissen in de analyse geen beperkingen op te leggen aan de mogelijke toestandsveranderingen binnen dat kenmerk. Het kenmerk wordt dan als **ongeordend** of **niet-additief** beschouwd, wat inhoudt dat elke overgang tussen twee willekeurige toestanden van een kenmerk steeds maar één enkele stap vereist. Andere kenmerken lenen zich door hun aard niet tot ordenen van toestanden. Een voorbeeld hiervan vindt men bijvoorbeeld in DNA-sequenties, waarbij elke positie in de sequentie 1 kenmerk levert, met als vier mogelijke toestanden de basen A, C, G, en T.

Een ongeordende analyse van veeltoestandskenmerken wordt ook **spaarzaamheid volgens Fitch** ("*Fitch parsimony*") genoemd, naar Fitch die als eerste een practisch bruikbaar algoritme ontwierp dat een ongeordende analyse van kenmerken toelaat (Fitch 1971). De geordende analyse wordt dan **spaarzaamheid volgens Wagner** ("*Wagner parsimony*") genoemd, dit omdat de eerste algoritmes die hiervoor gebruikt werden, gebaseerd waren op het werk van Wagner (zie 1.1).

1.5 De cladistische classificatie

Het aanwenden van een cladistische analyse om inzicht te krijgen in de evolutie van een bepaalde groep resulteert niet zonder meer in een classificatie van die groep. In dit punt gaan we even in op het classificatieaspect: hoe kan de informatie uit een cladogram in een classificatie weerspiegeld worden? Een strikt cladistische classificatie - dit is een classificatie van waaruit het vertakkingspatroon van het onderliggende cladogram ondubbelzinnig kan worden gereconstrueerd - vereist (1) dat enkel monofyletische groepen worden erkend en (2) dat zustergroepen dezelfde rang krijgen. Binnen deze twee regels blijven meerdere mogelijkheden open: er wordt immers enkel vereist dat zustergroepen dezelfde rang krijgen, welke rang dat moet zijn volgt niet zonder meer uit het cladogram.

Een strikt cladistische classificatie van de Primulanae volgens het cladogram van fig. 1.7 zou er kunnen uitzien zoals in fig. 1.11 (rechts). Uit deze figuur blijkt onmiddelijk dat een strikt cladistische classificatie heel wat rangen vereist. Om deze proliferatie van rangen enigszins in te dijken, bestaan een aantal conventies (Wiley 1979), waarvan de sequentieregel de belangrijkste is. Deze regel stelt dat een opeenvolging van een aantal taxa van dezelfde rang in een classificatie impliceert dat het eerste taxon van deze sequentie de zustergroep is van alle andere taxa uit die reeks en zo verder. Met behulp van deze regel kan bijvoorbeeld de linkerclassificatie van fig. 1.11 voorgesteld worden (ook hier blijven meerdere mogelijkheden open).

```
Classis (1)
 Subclassis (2)
            Familia Styracaceae
     Familia Lissocarpaceae
 Subclassis (3)
  Superordo
     Familia Ebenaceae
  Superordo (4)
   Ordo
     Familia Sapotaceae
   Ordo (5)
    Subordo
     Familia Theophrastaceae
    Subordo (6)
     Superfamilia (8)
            Familia Aegicerataceae
     Familia Myrsinaceae
     Superfamilia (7)
     Familia Primulaceae
     Familia Coridaceae
```

```
Superordo (1) Primulanae
  Ordo (2)
         Familia Styracaceae
    Familia Lissocarpaceae
  Ordo (3)
    Familia Ebenaceae
  Ordo (4)
    Familia Sapotaceae
  Ordo (5)
   Subordo
    Familia Theophrastaceae
   Subordo (8)
    Familia Aegicerataceae
    Familia Myrsinaceae
   Subordo (7)
         Familia Primulaceae
    Familia Coridaceae
```

Fig. 1.11. Links: een strikt cladistische classificatie van de Primulanae voor het cladogram van fig. 1.7. Rechts: met gebruik van de sequentieregel (Wiley 1979) kan het aantal rangen gereduceerd worden. In beide gevallen verwijzen de getallen naar de vertakkingspunten op het cladogram van fig. 1.7.

Vaak zijn bepaalde takken van een cladogram maar erg weinig ondersteund. In zulke gevallen kan het wenselijk zijn om toch parafyletische groepen te aanvaarden. Op die manier kan men soms bestaande classificaties voorlopig behouden. Het zou immers al te voorbarig zijn om bestaande classificaties op basis van weinig ondersteunde takken van een cladogram te gaan wijzigen. Zo kan men zich in dit voorbeeld aansluiten bij de ideeën van Takhtajan (vide Brummitt 1992) en diens indeling van de Primulanae voorlopig aanvaarden (fig. 1.12). Volgens de uitgevoerde analyse zijn de Ebenales uit deze classificatie dan wel parafyletisch.

Merk op dat de twee principes van cladistische classificatie losstaan van de logica achter het gebruik van cladistiek als fylogenetische reconstructiemethode. Het verwerpen van een strikt cladistische benadering van classificatie vormt dus op zich geen kritiek voor het cladisme als fylogenetische reconstructiemethode.

```
Superordo (1) Primulanae
   Ordo Ebenales
             Familia Styracaceae
      Familia Lissocarpaceae
      Familia Ebenaceae
   Ordo Sapotales
      Familia Sapotaceae
```

```
    Ordo (5) Primulales
Familia Theophrastaceae
       Familia Aegicerataceae
       Familia Myrsinaceae
       Familia Primulaceae
       Familia Coridaceae
```

Fig. 1.12. Een classificatie (met inbegrip van de parafyletische Ebenales) van de Primulanae voor het cladogram van fig. 1.7 (de getallen verwijzen naar de vertakkingspunten op dat cladogram).

1.6 Samenvatting

De belangrijkste theoretische principes van het cladisme kunnen als volgt samengevat worden (naar Scotland 1992):

1. de hiërarchie in de natuur kan worden weergegeven door een vertakkend diagram, een cladogram;
2. de status van een kenmerktoestand verandert al naargelang het beschouwde hiërarchisch niveau; zo zijn bijvoorbeeld kenmerktoestanden die in alle leden van een bestudeerde groep aanwezig zijn niet bruikbaar om verwantschappen binnen deze groep te bestuderen;
3. congruentie van kenmerken is van doorslaggevend belang om homologie van niet-homologie te onderscheiden;
4. congruentie van kenmerken kan geoptimaliseerd worden door het spaarzaamheidsprincipe.

Practisch gezien zullen bij elke cladistische analyse de volgende stappen aan bod komen (naar Stuessy 1990):

1. de keuze van de taxa die gezamenlijk de binnengroep zullen uitmaken; deze taxa moeten elk op zich monofyletisch zijn (dus autoapomorfieën bezitten) en het geheel van de binnengroep moet eveneens een monofyletische groep vormen;
2. de keuze van goede kenmerken die bovendien voldoende variatie vertonen binnen de binnengroep (uit de literatuur en/of uit eigen onderzoek); een goed morfologisch kenmerk kan niet gedefinieerd worden zonder grondig vergelijkend onderzoek in de bestudeerde groep;
3. het polariseren van de kenmerken (bij buitengroepvergelijking gebeurt dit door één of meerdere buitengroepen in de analyse te betrekken);
4. het opstellen van de gegevensmatrix;
5. het genereren van de cladogrammen met behulp van het passende spaarzaamheidsalgoritme;
6. het voorstellen van een classificatie die gebaseerd is op de gevonden cladogrammen.

In de praktijk zullen de stappen twee en drie de meeste problemen opleveren omdat goed gedefinieerde kenmerken en kenmerktoestanden schaars zijn, zeker op een hoger taxonomisch niveau.

2. THREE-ITEM ANALYSIS[1]

2.1 Introduction

Some years ago, three-item analysis was introduced as a novel approach to parsimony analysis in both biogeography (Nelson & Ladiges 1991a, 1991b) and systematics (Nelson & Platnick 1991). The name three-item analysis refers to the fact that each statement about relationships between more than three items (areas in biogeography, homologous features in systematics) is decomposed into a series of basic statements, each of which involves only three items. Such a basic statement says which two of the three items are thought to be related more closely to each other than either is related to the third. It was hoped (Nelson & Ladiges 1991a, 1991b, Nelson & Platnick 1991) that three-item analysis might increase the precision of parsimony, i.e. its sensivity to differences in the fit of data to alternative cladograms.

The three-item approach has been further explained and clarified by Nelson (1992, 1993, 1994, 1996), Nelson & Ladiges (1992, 1993), and Platnick (1993). Practical applications in systematics, using a set of computer programs written by Nelson & Ladiges (1995), can be found in Nelson & Ladiges (1994), Patterson & Johnson (1995) and Udovicic et al. (1995). The use of three-item analysis in biogeography has been further explored by e.g. Ladiges et al. (1992), Morrone & Carpenter (1994) and Nelson & Ladiges (1996). As the justification for using three-item statements in systematics on the one hand and biogeography on the other may be different (Nelson 1992), it should be noted that the argumentation in this chapter is limited to systematics.

After its introduction in systematics, the three-item approach has been criticized by Harvey (1992), Kluge (1993, 1994), Wilkinson (1994b), De Laet & Smets (1995) and Farris et al. (1995). At first sight, the points of criticism are numerous and seem to involve many different aspects of the method. However, closer inspection reveals that many criticisms are related, and the following three basic problems emerge: (1) three-item analysis is flawed because it presupposes that character evolution is irreversible; (2) three-item analysis is flawed because basic statements that are not logically independent are treated as if they are; (3) three-item analysis is

[1] Part of this chapter (2.3.1-2.3.2) was presented at the XIVth meeting of the Willi Hennig Society (July 30 - August 3, 1995, College Station, Texas; see De Laet & Smets 1995).

flawed because some of the three-item statements that are considered as independent support for a given tree may be mutually exclusive on that tree. As will be shown, none of these basic criticisms has been adequately answered by Nelson (1992, 1993, 1994, 1996; see also Nelson & Ladiges 1992, 1993, 1994) or Platnick (1993). De Laet & Smets (1995) proposed four-item analysis as a solution to the first problem, but this modification of three-item analysis still suffers from the two other problems (Farris, pers. comm.). In this chapter, I will examine if four-item analysis can be further refined in order to remove these problems. I will first shortly describe three-item analysis as it was originally proposed by Nelson & Platnick (1991) and then discuss the three basic problems.

2.2 Three-item analysis

2.2.1 Theory

In the standard approach, character state distributions are typically given in the form of a matrix with the rows representing taxa and the columns characters. An example, showing the character state distributions of two binary characters over five taxa, is given in the left part of fig. 2.1. During standard parsimony analysis, these character state distributions are fitted in their entirety onto cladograms, and in this way the congruency between characters and cladograms is maximized. The two ground intuitions behind three-item analysis (Nelson & Platnick 1991) seem to be (1) the idea that the state distribution of a single character is a compound statement that might somehow be further decomposed into more basic, atomic statements, and (2) the idea that a parsimony analysis on the level of these basic statements might provide a better measure of congruence between data and cladograms. The nature of such basic statements then becomes a central question.

Because the smallest possible statement that is still informative about relative cladistic branching necessarily involves three taxa (e.g. ‘taxa A and B are related more closely to each other than either is to C’), Nelson & Platnick (1991) proposed that basic statements are statements about three taxa, two of which have a character state that is derived with respect to the character state that is present in the third taxon. Such a statement hypothesizes that, on the basis of the character involved, the two taxa with the derived state are more closely related to each other than either is to the third; i.e. they belong to a monophyletic group from which the third is excluded. Following this line of reasoning, the first step of three-item analysis is the decomposition of the character state distribution of each character under study into the series of such statements that are implied by the distribution. This decomposition is mostly called a “transformation”, sometimes with the negative connotation that the information content of the data is being distorted (e.g. Harvey 1992, Kluge 1993; note that each standard matrix has a unique three-item representation, and that the reverse is not true).

	STANDARD APPROACH			THREE-ITEM APPROACH		
characters		a b			a	b
				O	000	000000
taxa	A	0 0		A	0??	000???
	B	0 0	**"TRANSFORMATION"**	B	?0?	???000
	C	0 1	→	C	??0	11?11?
	D	1 1		D	111	1?11?1
	E	1 1		E	111	?11?11

Fig. 2.1. The representation of the character state distributions of two characters, a and b, over five taxa, A-E. Left: representation in the standard approach. Right: representation in the three-item approach; each column stands for one three-item statement; the character state distribution of character a implies three three-item statements, character b implies six three-item statements; an outgroup (O) is added to indicate that 0 represents the plesiomorphic state, ‘?’ is used as a placeholder to indicate taxa that are not part of a particular statement; see text for further discussion.

In order to perform the transformation or decomposition (fig. 2.1, right part; each column is a single basic three-item statement), Nelson & Platnick (1991) make the assumption that for each character the plesiomorphic state has been determined a priori. These plesiomorphic states are by convention coded as ‘0’, and are assigned to a hypothetical outgroup taxon, O, that is added to the data set. Because 0 is assumed to be plesiomorphic, only 0-1-1 three-item statements (one taxon having state 0, two taxa having state 1) have to be considered; 0-0-1 statements (two taxa having state 0, one taxon having state 1) are not informative with respect to cladistic branching: they merely indicate that one out of three taxa has a derived character state. It follows that the complete transformed representation of the character state distribution of a single character consists of the set of all possible 0-1-1 three-item statements that can be derived from the standard representation of the character. In a single three-item statement, the three taxa involved are indicated using their character state (0 or 1); question marks, mostly used for missing entries (Platnick et al. 1991), serve as placeholders for the remaining taxa. It is clear that the number of informative three-item statements that are implied by a character state distribution, NTIS_{ch}, depends not only on the total number of taxa, N, but also on the numbers of taxa having the apomorphic state (ot_{ch}) and plesiomorphic state (zt_{ch}) for that character: $\mathrm{NTIS}_{ch} = \mathrm{zt}_{ch}*\mathrm{ot}_{ch}*(\mathrm{ot}_{ch}-1)/2$. When the character has no missing entries, $\mathrm{zt}_{ch} = N - \mathrm{ot}_{ch}$.

In the final step of three-item analysis, the matrix consisting of the three-item statements is subjected to standard parsimony analysis, using any of the programs available (e.g. Farris 1988, Swofford 1993, Goloboff 1993b; the all-zero hypothetical outgroup is included as a technical necessity to force 0 effectively into the plesiomorphic role). An individual three-item statement either fits a cladogram or not, and therefore the resulting cladograms will be those that maximize the number of three-item statements that can be accommodated. Each of these accommodated statements can be interpreted as a valid indicator of monophyly.

2.2.2 Example

Nelson & Platnick (1991) introduced the method as a means to increase the precision of parsimony by breaking up full character state distributions into the smallest possible statements that remain informative. However, they were far from precise in explaining what exactly was meant by this increased precision. In this respect, a discussion of a small hypothetical data set (from Nelson 1996) may be more illuminating than a reiteration of Nelson and Platnick's often problematic theoretical comments on this point.

A	000
B	110
C	101
D	011

O	00 00 00
A	0? 0? 0?
B	11 11 ?0
C	11 ?0 11
D	?0 11 11

Fig. 2.2. Hypothetical data set in standard (left) and three-item (right) representation.

Consider a data set (fig. 2.2) containing the state distribution of three characters over four taxa. Each of the three characters resolves BCD differently. A standard analysis of this matrix (fig. 2.3) yields six most parsimonious trees of length 5: three trees with a basal trichotomy, AB(CD), AC(BD), AD(BC), and three fully resolved trees, A(B(C D))), A(C(B D)), A(D(B C)). On each of these trees, one character is free of homoplasy, while the two other both require one extra step. The strict consensus tree of the most parsimonious trees is the uninformative bush ABCD.

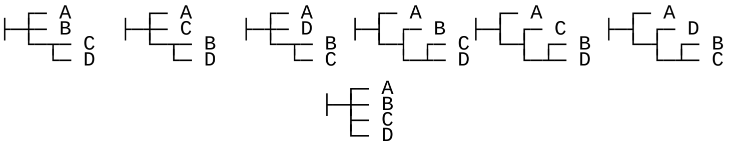


Fig. 2.3. The six most parsimonious trees (upper row) and their strict consensus tree (lower row) for the hypothetical data set of fig. 2.2 in the standard representation.

Replicating the original trio of characters does not alter the results beyond increasing tree length by 5 with every replication; e.g. the matrix shown in fig. 2.4 yields the same six trees, each with length 50. Nelson (1996) then raises the question if, when confronted with a matrix as in fig. 2.4, we might not eventually judge that taxa BCD are related more closely to each other than any of them is related to A. Indeed, even if the matrix contains three conflicting suites of characters (each resolving BCD differently), none of the characters contradicts A(BCD), and for each taxon of BCD there is plenty evidence that it is related more closely to the other members of BCD than to A.

A	000 000 000 000 000 000 000 000 000 000
B	110 110 110 110 110 110 110 110 110 110
C	101 101 101 101 101 101 101 101 101 101
D	011 011 011 011 011 011 011 011 011 011

Fig. 2.4. Standard representation of a hypothetical data set in which every character of the data set from fig. 2.2 occurs ten times.

A closer relationship between members of BCD with respect to A is precisely what is obtained by analyzing the three-item representation of the matrix in fig. 2.3, or of any matrix in which this series of three-item statements is replicated any time: there are three different most parsimonious trees (fig. 2.5), on which four out of the six three-item statements can be accommodated: A(B(C D)), A(C(B D)), and A(D(B C)). The strict consensus of these is A(BCD). The BCD component in the strict consensus reflects the judgement that the matrix as a whole contains evidence for a (BCD) group, while lack of resolution within BCD follows from the conflicting resolutions of the group in the fundamental cladograms, reflecting the ambiguity of the data with respect to its inner relationships.

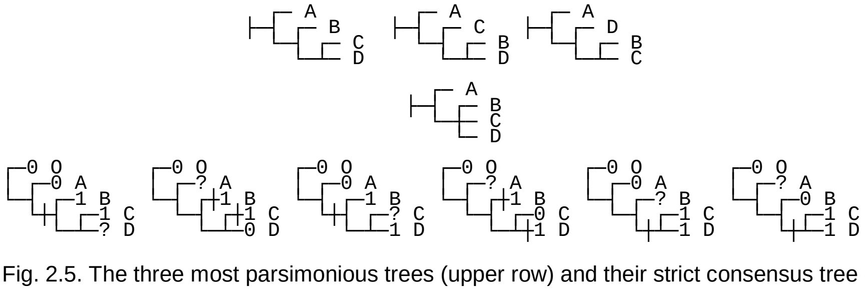


Fig. 2.5. The three most parsimonious trees (upper row) and their strict consensus tree (middle row) for the six three-item statements of the hypothetical data set of fig. 2.2. In the lower row, the six three-item statements are shown on one of the three most parsimonious trees, with the hypothetical outgroup included; four of the three-item statements require only one step, two of them require an extra step (steps are indicated by vertical bars; other optimizations than the ones shown are possible).

Of course it remains open to discussion whether this kind of evidence for an A(BCD) grouping should be incorporated in a cladistic analysis, and, if yes, whether three-item analysis is the right way for doing so (the loss of the BCD group in the strict consensus of the standard analysis might as well follow from the way zero-length branches are treated; cf. Coddington & Scharff 1994). The example nevertheless gives an intuitive grasp of some of the considerations that are involved when pursuing alternative approaches such as three-item analysis.

2.3 The problem of irreversibility

2.3.1 Introduction

In the standard approach, most parsimonious trees can be constructed under the strong assumption that all character evolution is forward: once a derived character state has evolved it will never revert to the plesiomorphic state. Consequently, all homoplasy is explained in terms of convergence, and reversals are not allowed. Coupled with the assumptions that character state order (for multistate characters) and polarity can be determined prior to the parsimony analysis, this has been called Camin-Sokal parsimony (e.g. Swofford et al. 1996; cf. Camin & Sokal 1965: 312). From the discussion in the previous section, it may appear at first sight that the only of these assumptions involved in three-item analysis is a priori polarization. However, the decision not to include 0-0-1 three-item statements in the three-item matrix implies the much stronger assumption of irreversibility. Indeed, 0-0-1 statements were left out of the matrix because they were uninformative, but they are only so if it is assumed that state 1 can never be plesiomorphic with respect to a reverted state 0. If three-item analysis is indeed an alternative to standard parsimony analysis, it is only so under the very restrictive assumptions of Camin-Sokal parsimony (De Laet & Smets 1995; see also Kluge 1993: 251).

Contrary to the situation in Camin-Sokal parsimony, standard Wagner parsimony (Kluge & Farris 1969) and standard Fitch parsimony (Fitch 1971) do not make the assumption of irreversibility of character evolution. Under these conditions, it is no longer possible to determine a priori whether a given 0-1-1 or 0-0-1 three-item statement is either informative or uninformative. All depends on the state that is plesiomorphic relative to the three taxa under consideration.

Nelson & Platnick (1991: 362-363; see also Platnick 1993: 268) considered the possibility that reversals could be problematic in three-item analysis. Nevertheless, by providing a hypothetical data set they showed by example that three-item analysis can identify clades that are supported by reversals only[2]. Kluge (1993, 1994) rightly pointed out that such hypothetical examples do not solve the basic problem: a three-item matrix is constructed in a way that congruence can no longer be used to test putative symplesiomorphies as evolutionary reversals.

Platnick (1993: 268) asserted that there is no real problem: in principle any individual character polarity could be altered in any possible combination of polarities of the other characters, and these alternative a priori polarities could be compared to achieve maximum congruence. However, rather than solving the problem, this suggestion merely reverses it for any individual character: either one or the other state is assumed a priori to be plesiomorphic throughout the complete tree. For the same reason, it would not help to substitute the hypothetical outgroup for a real one.

2.3.2 Solution

If the basic intuition of three-item analysis is that character state distributions should be broken up into the smallest possible statements that are still informative with respect to cladistic relationships, a generalization that does not assume irreversibility or a priori polarization suggests itself. Consider a 0-0-1-1 four-item statement, i.e. a statement about four taxa, two of which have state 0, and two of which have state 1. Such a statement is always informative with respect to cladistic branching. Indeed, independent of the state that is plesiomorphic, a 0-0-1-1 four-item statement will either be accommodated on a particular tree (only one step required) or not (two steps required). The other possible types of four-items statements (0-0-0-0, 0-0-0-1, 0-1-1-1, and 1-1-1-1) are all uninformative because they require the same number of steps on any tree (no steps for 0-0-0-0 and 1-1-1-1; one step for 0-1-1-1 and 0-0-0-1). In order to denote a particular four-item statement, all taxa that have the same state will be put between square brackets. E.g. [ABC][D] means either that taxa A, B, and C have state zero, and that taxon D has state one, or vice versa; this statement is uninformative. [AB][DE], on the other hand, is informative: there are two pairs of taxa that have a different state. Square brackets are used to avoid confusion between four-item statements on the one hand and resolutions of the statement on particular cladograms on the other. As an example, the informative four-item statement [AB][CD] is accommodated on cladograms that resolve the relationships between taxa A, B, C, and D as e.g. (AB)(CD) or (D(C(A B))) but not on cladograms that resolve the relationships as e.g. (A(C(BD))) or (BC)(AD).

[2] Incidentally, their example contains an error: if the data set is analyzed as it is presented, no such clade is identified; the example only works under differential weighting, e.g. with weight 3 for characters 1-3 and weight 2 for characters 4-5.

In a similar way as the 0-1-1 three-item matrix is derived from the standard representation of the character state distributions, a 0-0-1-1 four-item matrix can be derived. Such a matrix should include all possible 0-0-1-1 four-item statements that are implied by the standard representation of the character state distributions. An example, using the same taxa and characters as in fig. 2.1 is shown in fig. 2.6.

A	0 0	A	00?	000
B	0 0	B	0?0	000
C	0 1	C	?00	11?
D	1 1	D	111	1?1
E	1 1	E	111	?11

Fig. 2.6. The hypothetical data set of fig. 2.2 in standard (left) and four-item (right) representation.

As was the case with three-item statements, the number of informative four-item statements that are implied by a character state distribution, $NFIS_{ch}$, depends not only on the total number of taxa, N, but also on the numbers of taxa having the apomorphic state (ot_{ch}) and plesiomorphic state (zt_{ch}): $NFIS_{ch} = (zt_{ch}*(zt_{ch}-1)/2)*(ot_{ch}*(ot_{ch}-1)/2)$. Mostly the number of implied four-item statements greatly exceeds the number of implied three-item statements. Only when $zt_{ch} = 2$ or $zt_{ch} = 1$, there are more three- than four-item statements (twice as much three-item statements for $zt_{ch}=2$; no implied four-item statements at all for $zt_{ch}=1$); when $zt_{ch} = 0$, $ot_{ch} = 0$ or $ot_{ch} = 1$ (no implied three- or four-item statements) or when $zt_{ch} = 3$, the numbers of three- and four-item statements are equal.

Because the individual four-item statements of a four-item matrix do not imply assumptions about polarity, parsimony analysis will yield undirected topologies rather than cladograms. As long as the cladogram is not directed, the meaning of an accommodated four-item statement [AB][CD] remains equivocal: either A and B are more closely related to each other than either is to C or D, or C and D are more closely related to each other than either is to A or B; all that can be said is that at least one of both interpretations must be correct. In order to obtain cladograms from which hypotheses of polarity can be read, these topologies have to be directed.This is completely analogous to the situation in standard Fitch or Wagner parsimony, and the same basic possibilities to direct the topologies exist (Nixon & Carpenter 1993): either good hypotheses about possible outgroup taxa are present or not. In the first case, the four-item statements that are considered should include the outgroup taxa. After the most parsimonious cladograms for the resulting four-item matrix are obtained, they can be rooted between ingroup and outgroups if at least the ingroup is monophyletic.

If outgroup taxa should appear within the ingroup, the initial assumption of ingroup monophyly is not supported by the data. When good hypotheses for outgroup taxa are lacking, one can still fall back on a kind of hypothetical outgroup that reflects a priori assumptions of plesiomorphy, avoiding, however, assumptions of irreversibility. Therefore, only four-item statements about the ingroup taxa should be included in the four-item matrix, and only after the most parsimonious topologies are obtained, the hypothetical ancestor is used to determine the position of the root by inserting it in the most parsimonious position. This way of using hypothetical ancestors is completely analogous to the situation in the standard approach (Lundberg 1972, Nixon & Carpenter 1993).

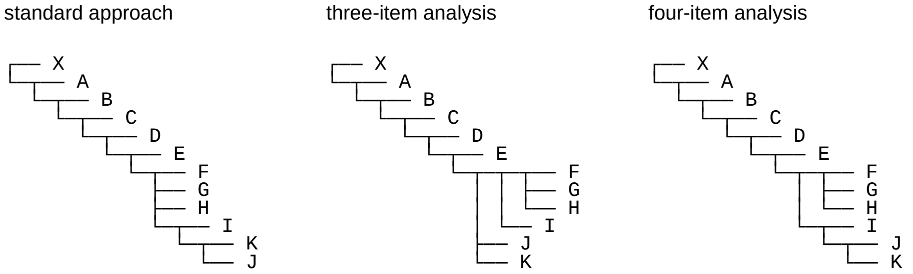
The fact that the use of three-item statements goes along with stronger a priori assumptions about evolutionary processes than does the use of four-item statements has an anology in phylogenetic distance methods (see e.g. Swofford et al. 1996). Ultrametric distance methods are methods that assume that mutation rates are equal among lineages, i.e. that there exists a universal evolutionary clock such that all lineages are equally diverged; this strong assumption can be tested a priori by examining if each possible triplet (i.e. a group of three taxa) in the distance matrix satisfies the so-called three-point condition. Additive distance methods, on the other hand, require only that the sum of all branch-lengths between two terminal taxa equals the observed pairwise distance between these taxa. This assumption is less restrictive than the assumption of a universal clock, and can be tested a priori by examining if each possible quartet of taxa satisfies the so-called four-point condition. Curiously enough there exist distance methods for data that are approximately additive that, from the point of view presented in this chapter, may be said to apply the four-item approach to pairwise distance data (e.g. Sattath & Tversky 1977, Fitch 1981): for each group of four taxa, the observed distances are used to derive a basic unpolarized statement concerning the relationships between these four taxa, and in a following step the trees on which the largest number of these basic statements is accommodated are identified.

2.3.3 Examples

Kluge (1994: 408-410) presented two hypothetical data sets to illustrate that three-item analysis does have problems in finding clades that are supported by reversals only. The rationale for using four-item statements in stead of three-item statements was that the a priori assumption of irreversibility should be avoided (2.3.2). If four-item analysis as described in the previous sections is indeed effective in removing that assumption, it should have no problem in identifying such clades.

Therefore, Kluge's (1994) examples are reproduced here (figs. 2.7 and 2.8), complemented with the results of a four-item analysis (using the computer program ViTA2; cf. Appendix 1).

X	000000000
A	100000000
B	110000000
C	111000000
D	111100000
E	111110000
F	111111111
G	111111111
H	111111111
I	011111111
J	001111111
K	001111111


standard approach

three-item analysis

four-item analysis

Fig. 2.7. A hypothetical data set (Kluge 1993: 408); taxon X is a hypothetical outgroup; reversals in the two first characters specify a clade (I(J K)) that is nested within clade F-K. Left: single most parsimonious tree in standard parsimony analysis; middle: single most parsimonious tree in three-item analysis; right: strict consensus of nine most parsimonious trees in four-item analysis.

Looking at Kluge's (1994) first hypothetical matrix (fig. 2.7, top), it is clear that these data imply a (FGH(I(J K))) clade in which the nested monophyletic groups (I(J K)) are supported by reversals only. This is confirmed in the single most parsimonious tree found by standard parsimony analysis (fig. 2.7, lower left). The analysis of the three-item representation of the same characters yields nine most parsimonious trees, each accomodating 589 out of 708 informative three-item statements. From their strict consensus (fig. 2.7, lower middle) it is clear that three-item analysis results in a paraphyletic taxon IJK. For four-item analysis, there are 1014 informative four-item statements for eleven taxa (the hypothetical ancestor is excluded). The single most parsimonious tree accommodates 964 of these. Adding the hypothetical outgroup in the most parsimonious way results in the cladogram shown in fig. 2.7 (lower right; the same cladogram is obtained when X is considered as a real outgroup taxon; in this case, there are 1722 four-item statements for twelve taxa, 1547 of which are accommodated). As expected, (I(J K)) is present in this cladogram. The single difference between the standard and the four-item analysis lies in the fact that four-item analysis identifies a FGH clade that is apparently completely unsupported by the data.

X	000000
A	100000
B	110000
C	111000
D	111100
E	111110
F	111111
G	001111

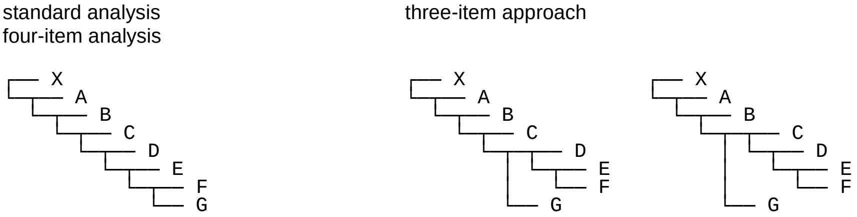


Fig. 2.8. A hypothetical data set (Kluge 1993: 408); taxon X is a hypothetical outgroup; G and F are highly derived sister taxa, with G exhibiting a reversal in the two first characters. This is confirmed by standard analysis as well as four-item analysis (left). Three-item analysis (middle and right) identifies two most parsimonious trees, in both of which G and F are far removed from each other.

In Kluge's (1993) second hypothetical data set (fig. 2.8, top), there are two highly derived taxa, F and G, one of which, G, shows reversal in two characters. In this case, the standard approach and four-item analysis give exactly the same result: there is one most parsimonious tree (fig. 2.8, lower left), in which G and F are sister groups. This tree accommodates 56 out of 66 four-item statements (or 121 out of 156 when taxon X is considered as a real outgroup taxon). Three-item analysis identifies two most parsimonious trees (fig. 2.8, lower middle and right), accomodating 67 out of 90 three-item statements. In both trees, G is far removed from F.

2.4 Algorithms

2.4.1 Introduction

As discussed above, the best trees according to the three-item approach are those that accommodate the largest number of basic statements that are implied by the characters at hand. Given this optimality criterion, one could devise many different procedures to arrive at these best trees. The procedure proposed by Nelson & Platnick (1991) consists of two steps: first transform the data into a matrix of three-item statements, and then find the shortest trees for this transformed matrix. Because an accommodated statement takes a single step, and a statement that is not accommodated takes two steps, the shortest trees for the transformed matrix are indeed those that maximize the number of accommodated statements. An obvious and major advantage of this two-step procedure is that - except for the transformation, which is readily automated - no new algorithms or computer programs are required. Indeed, once the data are transformed, any of the existing computer programs for standard parsimony analysis can be used. However, this two-step approach has also some serious drawbacks.

The first one is practical: the three-item matrix may become very large as the number of taxa increases. E.g. for twenty taxa, a single character with ten 0-entries and ten 1-entries implies 10*(10*9)/2 = 450 different three-item statements; for the double number of taxa, a character with twenty 0-entries and twenty 1-entries would yield already 20*(20*19)/2 = 3800 three-item statements. The problem is even worse in four-item analysis: for the two same characters, the numbers of implied four-item statements are (10*9/2)*(10*9/2) = 2025 and (20*19/2)*(20*19/2) = 36100 respectively.

A more fundamental drawback is that the matrix of basic statements does not contain the information which series of statements have been derived from single characters. This may seem unproblematic because the standard algorithms for evaluating the length of a character on a tree do not require that kind of information anyhow. However, as will be clear from the following, it is exactly this point that causes the problems of dependency (see 2.4), and to deal with this problem one has to know exactly which basic statements belong together. This information could easily be added to the transformed matrix, but then the standard algorithms for evaluating the number of steps are no longer appropriate, and special algorithms to calculate the number of accommodated basic statements become unavoidable.

If non-standard algorithms can no longer be avoided, Nelson & Platnick's (1991) two-step procedure might as well be reconsidered completely. In stead of complicating things by extending the first step to include the supplementary information, one could use algorithms that operate directly on the standard representation of character state distributions. In this way, the transformation step can be eliminated. As far as removal of a priori irreversibility assumptions is concerned, Nelson & Platnick's (1991) two-step procedure is still sound, and therefore it is not neccesary to present such alternative algorithms at this point. It is useful to do so, however, in order to have a starting point for further refinements as the problems of dependency and mutually exclusive optimizations will be treated.

The discussion is limited to binary characters without polymorphisms and to trees that are strictly dichotomous. Because the direction that is imposed on a tree does not influence the number of accommodated four-item statements, it may be assumed without loss of generality that the trees are directed by selecting arbitrarily one taxon as an outgroup. This makes it easier to distinguish between the three different branches that are incident on any inner node. The inner node that has the outgroup as one of its incident branches is called the basal node.

2.4.2 Calculating the number of accommodated four-item statements

Algorithms for calculating the length of a character on a tree under standard Fitch or Wagner parsimony require only a single pass over the nodes of the tree. The algorithms start by visiting the terminal nodes first and then proceed towards the basal node, thereby visiting only internal nodes whose descendants have already been visited previously (a post order traversal of the tree). The following algorithm for calculating the number of accomodated four-item statements also requires only a single post order traversal of the tree, and at each visited node it is calculated how many new (i.e. not counted already at previously visited nodes) accommodated four-item statements have their 00-part or 11-part above the node and the corresponding part below. The restriction to new accomodated statements ensures that ultimately these values will sum to the total number of accommodated four-item statements on the tree.

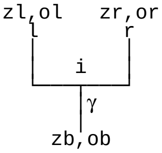


Fig. 2.9. An isolated inner node i with nodes l and r as its left and right descendants, and with branch γ leading downwards to i’s ancestor (or to the outgroup if i is the basal node); zl, ol, zr, or, zb, and ob are the numbers of taxa having state zero or one in the left descendant, in the right descendant, or in the part of the tree below node i.

Assume a character for which the total number of taxa having state zero is **zt**, and the total number of taxa having state one is **ot**. Furthermore, for each internal node, the numbers zl, zr, ol, and or are defined as follows (fig. 2.9):

- **zl**: the number of taxa having state zero in the left descendant
- **zr**: the number of taxa having state zero in the right descendant

- **ol**: the number of taxa having state one in the left descendant
- **or**: the number of taxa having state one in the right descendant.

These numbers are easily determined by a post order traversal of the tree. For convenience, two more numbers are defined for each internal node:

- **zb** = zt-(zl+zr): the number of taxa having state zero below the node
- **ob** = ot-(ol+or): the number of taxa having state one below the node

During the traversal of the tree in order to calculate the number of accommodated statements, the variables **ACC**, **ZZPV**, and **OOPV** are increased as new nodes are visited: **ACC** sums the number of accommodated four-item statements for all nodes already visited, **ZZPV** accumulates the number of 00-pairs of taxa that are present above nodes that have already been visited (including the current node), and **OOPV** accumulates the number of 11-pairs of taxa that are present above nodes that have already been visited. ZZPV and OOPV will be used to avoid double counting of accommodated statements. The calculation of the total number of accommodated four-item statements then proceeds as presented in fig. 2.10. For the sake of presentation it is assumed that the values of zl, zr, zb, ol, or, and ob for each internal node have already been determined during a previous postorder traversal of the tree, but they might as well be calculated along with the calculation of ACC, ZZPV, and OOPV. In this way, a single post order traversal of the tree suffices to obtain the required result.

```
• Initialize ACC, ZZPV, and OOPV as zero.
• Visit all internal nodes in post order and for each node do the following:
    1. add zl*zr to ZZPV;
    2. add ol*or to OOPV;
    3. add (zl*zr)*(ob*(ob-1)/2-(OOPV-(ol+or)*(ol+or-1)/2)) to ACC;
    4. add (ol*or)*(zb*(zb-1)/2-(ZZPV-(zl+zr)*(zl+zr-1)/2)) to ACC;
• The total number of accommodated four-item statements is the current value of ACC
```

Fig. 2.10. An algorithm for calculating the number of accommodated four-item statements for a given binary character on a given dichotomous tree. See text for explanation.

In the first step of the main part of the algorithm (fig. 2.10), ZZPV is increased with the number of 00-pairs above the current node that have not already been encountered above any previously visited node. As it is a post order traversal, the left and right daughter nodes of the current node have already been visited, and so all 00-pairs above the left daughter and all 00-pairs above the right daughter have already been considered. The remaining 00-pairs above the current node are those with one 0-taxon in left daughter and the other 0-taxon in the right daughter. The total number of these is zl*zr. In the second step, OOPV is increased in a similar way with the number of 11-pairs above the current node that have not already been encountered.

In the third step, ACC is increased with the number of not previously encountered accommodated four-item statements having their 00-part above the current node and their 11-part below the current node. In the fourth step, the same is done for the not previously encountered accommodated four-item statements having their 11-part above the current node and their 00-part below the current node. In this way, all new accommodated statements are accounted for. The expression that yields the number of new accommodated four-item statements having their 00-part above the current node and their 11-part below the current node (step 3) is obtained as follows: each of the zl*zr new 00-pairs above the current node yields an accommodated four-item statement in combination with any of the ob*(ob-1)/2 11-pairs below the current node; however, some of these 11-pairs below may already have been considered at previous nodes; this double counted number equals the total number of 11-pairs already encountered above visited nodes (including the current node) minus those that are present above the current node: OOPV - ((ol+or)*(ol+or-1)/2)). The expression in step 4 is obtained in a similar way.

2.4.3 Calculating the number of unaccommodated four-item statements

Once the final value for ACC is known, the total number of unaccommodated statements is easily determined as (zt*(zt-1)/2)*(ot*(ot-1)/2)-ACC, or the total number of statements minus those that are accommodated. However, this number may also be calculated directly (fig. 2.11), using an algorithm that is similar to the one presented above (fig. 2.10). In this case, the accumulating parameter **ZOPV** is used to accumulate the numbers of 01-pairs of taxa that are present above any of the nodes that have already been visited, including the current node. The number of unaccommodated statements is summed in **UNACC**.

```
• Initialize UNACC and ZOPV as zero.
• Visit all internal nodes in post order and for each node do the following:
    1. add zl*or+ol*zr to ZOPV
    2. add (zl*or + ol*zr)*(zb*ob-(ZOPV-(zl+zr)*(ol+or)) to ACC
• The total number of unaccommodated four-item statements is the current value of UNACC
```

Fig. 2.11. An algorithm for calculating the number of unaccommodated four-item statements for a given binary character on a given dichotomous tree. See text for explanation.

In the first step of the loop (fig. 2.11), ZOPV is increased with the number of 01-pairs above the current node that have not already been encountered above any previously visited node. Analogously to the situation above, these are the 01-pairs that have either a 0-taxon in the left and a 1-taxon in the right daughter, or a 1-taxon in the left and a 0-taxon in the right daughter. This number amounts to zl*or+ol*zr. In the second step, the number of new unaccommodated statements is added to UNACC. Each of the zl*or+ol*zr new 01-pairs above the current node yields an unaccommodated four-item statement in combination with any of the ob*zb 01-pairs below the current node; however, some of these 01-pairs below may already have been considered at previous nodes, and this double counted number equals the total number of 01-pairs already encountered minus those that are present above the current node: ZOPV-(ol+or)*(zl+zr).

2.4.4 An example

Both algorithms are illustrated by means of a hypothetical tree and character state distribution (fig. 2.12) for twelve taxa A-L. The cladogram is directed by using taxon A as an outgroup. At each inner node, the values of zl, zr, zb, ol, or, and ob that correspond to the given character state distribution are specified. The sequence of inner nodes a-b-c-d-e-f-g-h-i-j is one of the possible post order traversals of this cladogram. The values of the the accumulating parameters ZZPV, OOPV, ZOPV, ACC, and UNACC that are obtained for the internal nodes when they are visited in this order are shown in fig. 2.13. The final number of ACC, 64, is the number of accommodated statements; the final number of UNACC, 161, is the number of unaccommodated statements. The complete list of all 64 accommodated four-item statements, with the relevant nodes specified for each statement, is given in fig. 2.14.

Inner node a, the node that specifies a sister group relationship between taxa B and C, is the first node to be visited. Taxa B and C both have state one, so there is one 11-pair present above the node. This pair, added to OOPV in step 2, yields one accommodated four-item statement for each 00-pair of taxa below the node. As all six taxa having state 0 are present below node a, there are (6*5)/2=15 such pairs, and the fifteen resulting statements are added to ACC (step 4).

The next node, b, unites taxa D and E. Both taxa have state zero, so there is one 00-pair present above the node. This pair, added to ZZPV in step 2, yields one accommodated four-item statement for each 11-pair of taxa below the node. As all six taxa having state 1 are present below node a, there are (6*5)/2=15 such pairs. However, the statement that results from the 11-pair BC has already been counted at node a, so it is subtracted, and only the remaining fourteen statements are added to ACC (step 3), yielding a total of 29.

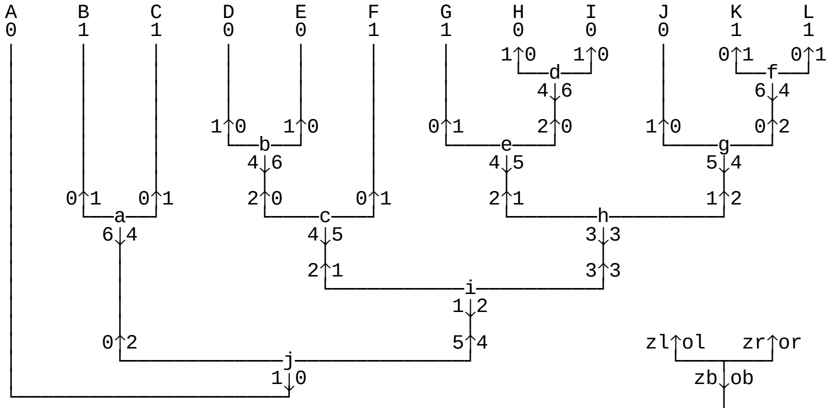


Fig. 2.12. An example cladogram for twelve taxa A-L, showing the state distribution of a single character and the corresponding numbers zl, ol, zr, or, zb, and ob for each inner node a-j.

At node c, no new 00- or 11-pairs are encountered above the node, so no new accommodated statements will be found and nothing happens to the parameters. There are, however, for the first time 01-pairs present above the node: DF and EF. These are added to ZOPV in the first step of algorithm of fig. 2.11. Each of these forms an unaccommodated statement with each 01-pair below the node. As there are four 0-taxa and five 1-taxa below the node, there are 4*5=20 such 01-pairs. The product of the 01-pairs above and below, 2*20 is added to UNACC in step 2.

Node d is completely analogous to node b, yielding an increase of 1 for ZZPV, and an increase of fourteen for ACC.

Node e does not yield new 00- or 11-pairs above the node, but there are 01-pairs present: HG and HI. These are added to ZOPV in the first step of the algorithm of fig. 2.11. Both HG and HI form an unaccommodated statement with each 01-pair below the node. As there are four 0-taxa and five 1-taxa below the node, there are 4*5=20 such 01-pairs. However, two of these, DF and EF, have already been encountered at node c, leaving only eighteen pairs below. The product of the new 01-pairs above and below, 2*18 is added to UNACC in step 2.

Inner node f, specifying a sister group relationship between taxa K and L, is visited next. Taxa K and L both have state one, so there is one 11-pair present above the node. This pair, added to OOPV in step 2, yields one accommodated four-item statement for each 00-pair of taxa below the node. As all six taxa having state 0 are present below node a, there are (6*5)/2=15 such pairs. However, two of these have been encountered before: DE at node b, and HI at node d. The thirteen statements that result from the remaining 00-pairs below are added to ACC in step 4. Note that the algorithm does not know at which node DE and HI have been encountered before, or not even the identity of these two pairs. All that must be known is the number of 00-pairs previously encountered, which is the current value of ZZPV (possibly diminished with the number present above the current node, see node h for an example). The treatment of node g is similar to the treatment of node e. The difference is that now four of the twenty 01-pairs below have already been encountered (DF and EF at node c; HG and IG at node d), which results in 2*(20-4)=32 new unaccommodated statements.

	algorithm of fig.2.10				algorithm of fig. 2.11	
inner node	ZZPV step1	OOPV step2	ACC step3	ACC step4	ZOPV step1	UNACC step2
a	0	1	0	15	0	0
b	1	1	29	29	0	0
c	1	1	29	29	2	40
d	2	1	43	43	2	40
e	2	1	43	43	4	76
f	2	2	43	56	4	76
g	2	2	56	56	6	108
h	4	4	60	64	11	143
i	10	7	64	64	20	161
j	10	15	64	**64**	30	**161**

Fig. 2.13. The values of the accumulating parameters (algorithms of figs. 2.10 and 2.11) for the character and the cladogram of fig. 2.12, corresponding to postorder traversal a-b-c-d-e-f-g-h-i-j.

The next node, node h, is an example of the most complex case the algorithms have to deal with. There are three taxa having state 0 above the node: two in the left daughter (H and I), and one in the right daughter (J). Of the three possible 00-pairs (HI, HJ, and IJ), only those that combine a 0-taxon from the left daughter with a 0-taxon from the right daughter are new: HJ and IJ (the number of new 00-pairs is calculated directly as zr*zl, and ZZPV is increased by two in step 1; similarly, OOPV is increased by two in step 2 for the two new 11-pairs). Both form an accommodated statement in combination with any 11-pair below. As there are three taxa having state 1 below the node (B, C, and F), there are 3 such pairs (BC, BF, and CF; the number is calculated as zb*(zb-1)/2). One of these (BC) has already been encountered before, at node a. However, at this point of the algorithm, the number of 11-pairs already encountered, OOPV, does not equal 1, but 4. Indeed, three out of the four 11-pairs already encountered are above the current node: KL was encountered at node f, h’s right daughter, and GK and GL are encountered at the current node itself. So, in order to obtain the number of already encountered 11-pairs below the current node, these three have to be subtracted from OOPV (the number to be subtracted is calculated as

[AD][BC]	**a_4**	
[AD][KL]	**f_4**	h_4
[AD][GK]	**h_4**	
[AD][GL]	h_4	
[AE][BC]	**a_4**	
[AE][KL]	f_4	h_4
[AE][GK]	h_4	
[AE][GL]	h_4	
[AH][BC]	**a_4**	
[AH][KL]	**f_4**	
[AI][BC]	**a_4**	
[AI][KL]	f_4	
[AJ][BC]	**a_4**	
[AJ][KL]	**f_4**	
[DE][BC]	a_4	b_3 i_3
[DE][BF]	**b_3**	
[DE][BG]	**b_3**	
[DE][BK]	**b_3**	
[DE][BL]	**b_3**	
[DE][CF]	b_3	
[DE][CG]	b_3	
[DE][CK]	b_3	
[DE][CL]	b_3	
[DE][FG]	b_3	
[DE][FK]	b_3	
[DE][FL]	b_3	
[DE][GK]	b_3	h_4
[DE][GL]	b_3	h_4
[DE][KL]	b_3	f_4 h_4
[DH][BC]	a_4	i_3
[DH][KL]	f_4	
[DI][BC]	a_4	i_3
[DI][KL]	f_4	
[DJ][BC]	a_4	i_3
[DJ][KL]	f_4	
[EH][BC]	a_4	i_3
[EH][KL]	f_4	
[EI][BC]	a_4	i_3
[EI][KL]	f_4	
[EJ][BC]	a_4	i_3
[EJ][KL]	f_4	
[HI][BC]	a_4	d_3 h_3 i_3
[HI][BF]	**d_3**	h_3
[HI][BG]	**d_3**	
[HI][BK]	**d_3**	
[HI][BL]	**d_3**	
[HI][CF]	d_3	h_3
[HI][CG]	d_3	
[HI][CK]	d_3	
[HI][CL]	d_3	
[HI][FG]	d_3	
[HI][FK]	d_3	
[HI][FL]	d_3	
[HI][GK]	d_3	
[HI][GL]	d_3	
[HI][KL]	d_3	f_4
[HJ][BF]	**h_3**	
[HJ][CF]	h_3	
[HJ][BC]	a_4	h_3 i_3
[HJ][KL]	f_4	
[IJ][CF]	h_3	
[IJ][BF]	h_3	
[IJ][BC]	a_4	h_3 i_3
[IJ][KL]	f_4	

Fig. 2.14. All 64 accommodated informative four-item statements for the character and the cladogram of fig. 2.12; the first letter following each node specifies the inner node at which the statement is first encountered using the algorithm of fig. 2.10 and the post order traversal a-b-c-d-e-f-g-h-i-j; any following letters specify nodes at which double counting has been avoided; the subscripts refer to steps 3 and 4 of the algorithm. For the meaning of bold script, see 2.5.3.

(ol+or)*(ol+or)/2: the total number of 11-pairs present above the current node). So ultimately there are 2*(3-(4-3)) new accommodated statements that are found during step 3 of the algoritm: [HI][BF], [HI][CF], [HJ][BF], and [HJ][CF]. Similar reasoning yields four more accommodated statements in step 4 of the algorithm ([AD][GK], [AD][GL], [AE][GK], and [AE][GL]), and thirtyfive new unaccommodated statements in step two of the algorithm of fig. 2.11. Finally, the results for node i and j are obtained in the same way.

2.5 The problem of dependency

2.5.1 Introduction

Nelson & Platnick (1991: 363) noted that it might be problematic for the three-item approach that not all three-item statements implied by a character are logically independent. E.g. a character having states 0111 for taxa A-D implies a total of three three-item statements: A[BC], A[BD], and A[CD]. However, because these three statements are derived from the same character, only two of the three are logically independent: whatever two are selected, the third one follows deductively, leaving a ratio of 2/3 of independent to total number of statements[4]. The independency ratio equals 2/3 in this particular case, but it may have different values in other cases. As an example, a character having states 0011 for the same taxa A-D implies only two three-item statements, A[CD] and B[CD], and both are independent, which gives an independency ratio of 2/2=1 in stead of 2/3. In general, the ratio of independent to total number of statements equals 2/ot (Nelson & Ladiges 1992; with zt the number of taxa having state 0 and ot the number of taxa having state 1, there are zt*(ot*(ot-1)/2) three-item statements in total, and only zt*(ot-1) of them are independent; the ratio of both is ot/2).

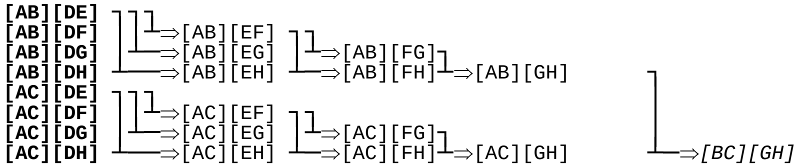
Nelson & Platnick (1991: 363) suggested to compensate for this phenomenon of different ratios of independent statements by reducing the weight of individual statements so that the total weight of all statements that are derived from a single character is equal to the number of independent statements for that character. This is accomplished by downweighting all statements by the independency ratio of their character (Nelson & Ladiges 1992; see also Nelson & Ladiges 1994). Because the ratio is by definition a fraction and because it is applied as a weight, the procedure is called fractional weighting, and the ratio a fractional weight (Nelson & Ladiges 1992).

At other places both Nelson and Platnick seem to hold the opinion that it does not pose a problem that some statements are logically implied by others. E.g. Platnick (1993) discusses a possible theoretical justification of three-item analysis without even mentioning the problem. Nelson, on the other hand, noted that dependency between statements does not alter the data (Nelson 1992: 356), and even denied that ‘three-item analysis produces non-independent characters’ (see Kluge 1994 and Farris et al. 1995 for comments). In the following, it is accepted that dependence is a problem, and the question is if fractional weighting provides a solution.

A similar procedure of fractional weighting might be devised for four-item analysis. In this case, the total number of four-item statements equals (zt*(zt-1)/2)*(ot*(ot-1)/2), and only (zt-1)*(ot-1) of them are independent (see fig. 2.15 for some examples). The ratio of both yields a fractional weight of zt*ot/4. However, the procedure of fractional weighting as proposed by Nelson & Ladiges (1992) does not properly solve the problems that are caused by dependency between basic statements. This is explained below, taking three-item analysis as an example. The same problems would arise in four-item analysis.

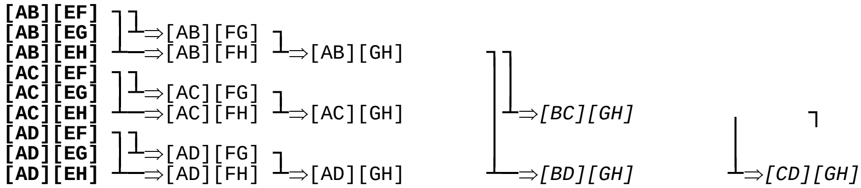
[4] Note that similar dependency problems exist in the methods of Sattath & Tversky (1977) and Fitch (1981) (cf. 2.3.2). These must be added to the inherent dependency problems of distance methods.

Character 11100000 for taxa ABCDEFGH: 8 independent statements on a total of 30 (zt=5, ot=3)



The ten statements of the form [BC][XY] result from combining [AB][XY] and [AC][XY]. This is only shown for X=G and Y=H (italic)

Character 11110000 for ABCDEFGH: 9 independent statements on a total of 36 (zt=4, ot=4)



The 18 statements of the form [BC][XY], [BD][XY], and [CD][XY] are obtained as shown for X=G and Y=H (italic).

Fig. 2.15. For any character having zt 0-taxa and ot 1-taxa, all (zt*(zt-1)/2)*(ot*(ot-1)/2) different implied four-item statements can be deduced from any set of (zt-1)*(ot-1) independent statements. Two examples are shown, and each time a possible independent set is indicated in bold.

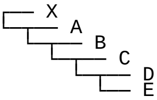
2.5.2 Fractional weighting

A simple example of fractional weighting is presented in fig. 2.16. Character a produces 6 three-item statements, only three of which are independent. A possible choice of independent statements might be A[BC], A[BD], and A[BE]: they collectively imply A[CD], A[CE], and A[DE] and are independent among themselves (in general, any choice of three statements will do as long as they are independent among themselves; an alternative choice of independent statements might be A[BC], A[CD], and A[DE]). Character b also yields six statements, but here four out of the six are independent (e.g. A[CD], A[CE], B[CD], and B[CE] are independent and imply A[DE] and B[DE]). Character c yields three statements, A[DE], B[DE], and C[DE], all three of which are independent.

On the single most parsimonious tree for these data, all three-item statements of all three characters are accommodated. However, when all three-item statements are equally weighted, the relative importance of the characters with a low indepency ratio is overrated because many of the accommodated statements are not independent. Applying the fractional weights rightly reduces the relative importance of the characters to their number of independent statements.

	abc
A	000
B	100
C	110
D	111
E	111

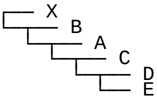
	a	b	c
X	**000**000	**00**0000	**000**
A	**000**000	**00**0???	**0??**
B	**111**???	**??**?000	**?0?**
C	**1??**11?	**11**?11?	**??0**
D	**?1?**1?1	**1?**11?1	**111**
E	**??1**?11	**?1**1?11	**111**
fractional weights:	3/6=**1/2**	2/6=**1/3**	3/3=**1**



accommodated statements	a	b	c	a+b+c
independent	3	2	3	8
total, unweighted	6	6	3	15
total, applying fractional weights	6*1/2=3	6*1/3=2	3*1=3	8

Fig. 2.16. Hypothetical data set in standard (top left) and three-item (top right) representation; for each character a possible choice of independent statements in the three-item representation is indicated in bold; the single most parsimonious tree has no homoplasy; the total number of weighted accommodated statements equals the total number of independent accommodated statements.

In the above case, the procedure of fractional weighting works correctly because there is no homoplasy in the data set. Indeed, the fractional weights as defined by Nelson and Ladiges (1992) reflect the ratio between independent and total number of accommodated three-item statements only in the absence of homoplasy. This is illustrated by inspecting a second tree for the above data (fig. 2.17). On this tree, all three-item statements of characters b and c are still accommodated, but in character a there is homoplasy: A[BC], A[BD], and A[BE] are not accommodated, leaving only A[CD], A[CE], and A[DE] accommodated. From these three, two are independent. Therefore, on this particular tree, the independency ratio of the accomodated statements for character a is 2/3 in stead of 1/2. The fractional weight as defined by Nelson & Ladiges (1992), however, remains fixed to 1/2. As a result, the total number of weighted accommodated statements for this character (3*1/2=1.5) underestimates the total number of independent accommodated statements (2).



accommodated statements	a	b	c	total
independent	2	2	3	7
total, unweighted	3	6	3	12
total, applying fractional weights	3*1/2=1.5	6*1/3=2	3*1=3	6.5

Fig. 2.17. On this tree, fractional weighting underestimates the total number of independent accommodated statements for character a (see fig. 2.16 for the character state distribution).

When the results for the two trees presented in figs. 2.16 and 2.17 are compared, the tree of fig. 2.17 is selected as the best tree by both unweighted and fractional weighted three-item analysis, and this tree also happens to be tree that accommodates the highest number of accommodated independent three-item-statements. However, because the unweighted approach takes into account many dependent statements and because fractional weighting may correct for this in the wrong way, neither the unweighted approach nor the fractional weighting necessarily find the trees that accommodate the highest number of accommodated independent statements. An example where both unweighted and fractional weighted three-item analysis do not select the tree with the highest number of independent statements is presented in figs. 2.18-2.20.

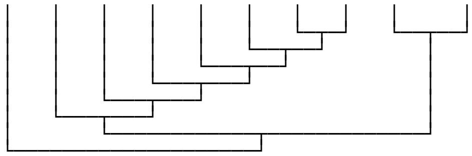
	012345 67
A	100000 00
B	110000 00
C	111000 00
D	111100 00
E	111110 00
F	111110 10
G	000001 00
H	000001 01
I	?????? 11

Fig. 2.18. A hypothetical data set specifying eight character state distributions (0-7) for 9 taxa (A-I).

First consider the data set shown in fig. 2.18 and assume that characters 0-5 each have an a priori weight that is higher than the total number of three-item statements implied by characters 6 and 7. Because of these weights and because characters 0-5 are fully congruent among themselves, the relative positions of taxa A-H in the best trees according to the three-item approach will be as specified by characters 0-5. Within these relationships, characters 6 and 7 specify the position of taxon I (fig. 2.19; see below for characters 8 and 9): in TREE1 I is the sister group of taxon F (character 6), in TREE2 it is the sister group of taxon H (character 7). Characters 6 and 7 have the same number of 0-taxa and 1-taxa, and therefore the same number of implied three-item statements. Moreover, none of the three-item statements of character 7 is accommodated on TREE1, and none of the three-item statements of character 6 is accommodated on TREE2. Therefore, both trees accommodate the same number of three-item statements. Any other position of taxon I in TREE1 beyond that specified in TREE2 decreases the number of accommodated statements for character 6 more than it increases the number of accommodated

TREE 1

	X	A	B	C	D	E	F	I	G	H
c8	0	0	1	1	1	1	1	1	0	1
c9	0	1	1	1	1	1	0	1	1	1



Accommodated three-item statements for character 8:

A[BC] A[BD] A[BE] A[BF] A[BI] A[CD] A[CE] A[CF] A[CI] A[DE] A[DF] A[DI] A[EF] A[EI] A[FI]

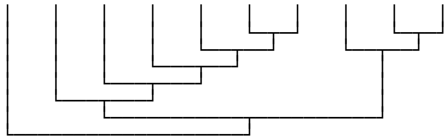
G[BC] G[BD] G[BE] G[BF] G[BI] G[CD] G[CE] G[CF] G[CI] G[DE] G[DF] G[DI] G[EF] G[EI] G[FI]

Accommodated three-item statements for character 9:

F[GH]

TREE 2

	X	A	B	C	D	E	F	G	H	I
c8	0	0	1	1	1	1	1	0	1	1
c9	0	1	1	1	1	1	0	1	1	1



Accommodated three-item statements for character 8:

A[BC] A[BD] A[BE] A[BF] A[CD] A[CE] A[CF] A[DE] A[DF] A[EF]

G[BC] G[BD] G[BE] G[BF] G[CD] G[CE] G[CF] G[DE] G[DF] G[EF]

A[HI]

G[HI]

Accommodated three-item statements for character 9:

F[GH] F[GI] F[HI]

Fig. 2.19. Two trees for ten taxa (A-I, X is the hypothetical outgroup that is added for three-item analysis), and the accommodated three-item statements for characters c8 and c9; possible sets of independent accommodated statements are underlined.

statements for character 7; conversely, any other position of taxon I in TREE2 beyond that specified in TREE1 decreases the number of accommodated statements for character 7 more than it increases the number off accommodated statements for character 6. Therefore, TREE1 and TREE 2 are the best three-item trees. Because characters 6 and 7 have the same number of 1-taxa, they have the same fractional weight, and both trees are the best trees under fractional weighting as well. Next consider two more characters, c8 and c9 (character state distribution shown in fig. 2.19), and assume that characters 0-7 each have an a priori weight that is higher than the total number of three-item statements implied by characters 8 and 9. Under these conditions, the best tree for the enlarged data set for ten characters (0-9) will be either TREE1 or TREE2, or both, depending solely on the numbers of accommodated statements for characters c8 and c9. The accommodated statements for characters c8 and c9 are listed in fig. 2.19, together with the indication of a possible set of independent statements.

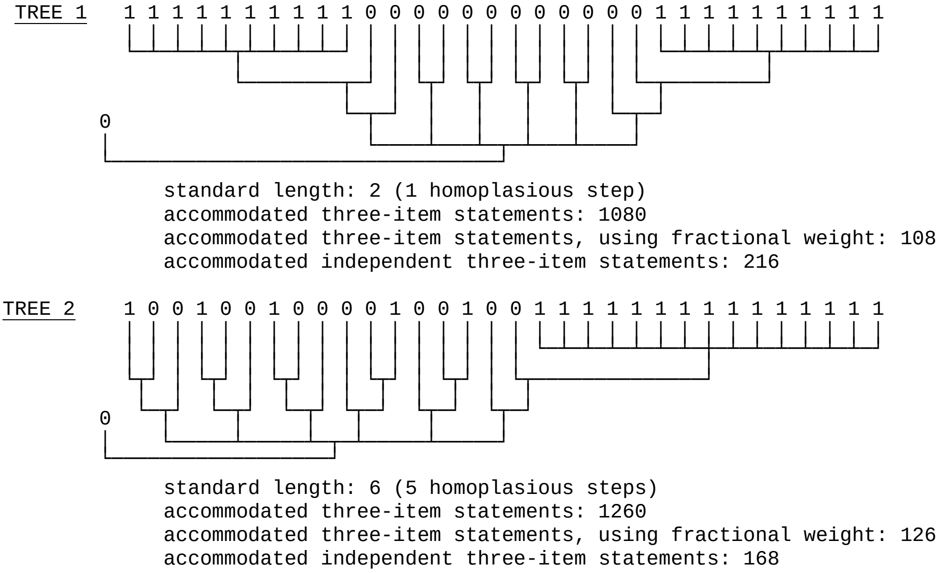
	TREE 1				TREE 2		
	t	t*f	i		t	t*f	i
c8	30	8.57	10	c8	22	6.29	10
c9	1	0.25	1	c9	3	0.75	2
Σ	**31**	**8.82**	**11**	**Σ**	**25**	**7.04**	**12**

Fig. 2.20. Summary of the three-item analysis of TREE1 and TREE2 (fig. 2.19); t: total number of accommodated three-item statements; f: fractional weight (2/7 for character c8, 2/8 for c9); i: total number of independent accommodated three-item statements.

From the summary of these results (fig. 2.20), it is clear that both unweighted and fractional weighted three-item analysis select TREE1 as the best tree, even though TREE2 accommodates one more independent three-item statement than does TREE1. In the unweighted analysis, TREE1 is preferred mainly because of the high number of dependent statements in the character 8 (20 out of 30). Using fractional weights, the number of independent accommodated statements is underestimated in both cases, but more so in TREE2 than in TREE1, leading once again to a preference for TREE1.

Fractional weighting was introduced as a means to eliminate the distortion produced by different independency ratios for different characters (Nelson & Ladiges 1992; see also Nelson & Ladiges 1994). Based on the above examples, it can be stated as a general conclusion that it works correctly only in the absence of homoplasy, i.e. when the correct topology is obtained anyhow. Whenever homoplasy is present in a character, the true number of its independent accommodated statements is underestimated, and the degree of underestimation is not related to the number of accommodated independent statements.

This phenomenon is also illustrated in an example presented by Farris et al. (1995: 213), reproduced here, slightly elaborated, as fig. 2.21. The example deals with a single character that has twenty 1-entries for 32 taxa (+ hypothetical outgroup). In fig. 2.21, the distribution of the states on two different trees is shown. Standard parsimony analysis clearly prefers the first tree, as this one only requires a single step of homoplasy, compared to five extra steps for the second tree. The same preference is expressed by the numbers of independent accommodated three-item statements (216 vs. 168). However, when the total number of accommodated three-item statements is taken into account (1080 vs. 1260), the second tree is preferred. As fractional weighting gives the same weight to all statements that are derived from a single character (1/10 in this case), it is obvious that the use of these weights will not change this preference (108 vs. 126).



standard length: 2 (1 homoplasious step)
accommodated three-item statements: 1080
accommodated three-item statements, using fractional weight: 108
accommodated independent three-item statements: 216

standard length: 6 (5 homoplasious steps)
accommodated three-item statements: 1260
accommodated three-item statements, using fractional weight: 126
accommodated independent three-item statements: 168

Fig. 2.21. A hypothetical example, showing counterintuitive results when all accommodated three-item statements are taken into account (Farris et al. 1995: 213).

So fractional weighting does not properly solve the problems caused by dependency because it assumes wrongly that on any given tree the ratio between dependent and total number of three-item statements is fixed at 2/ot for any character, irrespective of the amount of homoplasy that is present on the tree under consideration. As a solution, one might imagine a variation of fractional weighting so that the character's fractional weight is adjusted for each individual tree. However, in order to calculate such dynamic fractional weights, the number of independent statements on each tree has to be known in the first place, and the whole procedure of fractional weighting becomes redundant, as shown in the next section.

2.5.3 Calculating the number of independent accommodated four-item statements

In 2.4.2 (fig. 2.10), an algorithm to calculate the total number of accommodated four-item statements for a given character on a given tree was presented. This algorithm can be easily modified to obtain directly the number of independent accommodated four-item statements for the character. In this way, the problem of dependency is sidestepped.

First consider the distribution of independent 00-pairs and independent 11-pairs of taxa on a tree. As an example, consider a character with zt 0-taxa (zt>0). Such a character has zt*(zt-1)/2 00-pairs, only zt-1 of which are independent. In fig. 2.22 it is shown how these are distributed on the tree at an inner node where 0-taxa are present at right, at left, and below. In that case, the zl*(zl-1)/2 00-pairs at left can be deduced from a set IL of zl-1 independent 00-pairs that have their 0-taxa at the left of the node (note that the essential point is the number of independent pairs, and not their exact identity: any set of zl-1independent pairs will suffice to deduce all remaining ones). Similarly, all zr*(zr-1)/2 00-pairs at right can be deduced from a set IR of zr-1 independent 00-pairs that have their 0-taxa at the right of the node, and all zb*(zb-1)/2 00-pairs below can be deduced from a set IB of zb-1 independent 00-pairs that have their 0-taxa below the node. Any 00-pair that has it two taxa at the left (right, below) is independent from any 00-pair that has its two taxa at the right (left, right) or its two taxa below (below, left). Therefore, all zl+zr+zb-3 = zt-3 00-pairs in the union of the sets IL, IR, and IB are independent. However, in order to deduce all zt*(zt-1)/2 00, zt-1 independent pairs are required, so two more are necessary. These are obtained by selecting two independent pairs that each have their two taxa across the node. At inner nodes where one of the three subtrees (left, right, or below) does not have 0-taxa, the sum of the independent pairs of the two other subtrees is zt-2, and only a single pair with its 0-taxa across the node has to be added. When two of the subtrees have no 0-taxa, all zt-1 independent pairs are present in the third subtree.

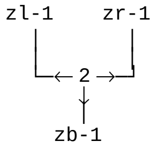


Fig. 2.22. The distribution of the zt-1 independent 00-pairs at an inner node where zl, zr, and zb are greater than zero. See text for explanation.

The logic of the new algorithm (fig. 2.23) is similar to the logic used in the algorithm of fig. 2.10. In that algorithm, it was calculated at any node how many new accommodated four-item statements were encountered. The new algorithm goes one step further and calculates at any node how many new independent statements must be added to the set of accommodated statements that have already been encountered so that the resulting set is sufficient to deduce these new accommodated statements. In the algorithm of fig. 2.10, the parameters ZZPV and OOPV accumulate the total number of 00-pairs and 11-pairs of taxa that are present above nodes that have already been visited (including the current node). In this algorithm, the parameters **IZZPV** and **IOOPV** are used. They accumulate the numbers of independent 00-pairs and 11-pairs that are present above previously visited nodes (including the current node). As ZZPV and OOPV in the algorithm of fig. 2.10, these parameters are used to avoid double-counting. A character with zt 0-taxa and ot 1-taxa has a total of zt-1 independent 00-pairs, and ot-1 independent 11-pairs. These totals are called **IZZPTOT** and **IOOPTOT**. The number of independent accommodated statements already encountered is accumulated in **IACC**. As in the algorithms of fig. 2.10 and 2.11, it is assumed for the sake of presentation that the values of zl, zr, zb, ol, or, and ob for each inner node (cf. 2.4.2) have already been calculated during a previous post order traversal of the tree, but they might as well be calculated along with IACC, IZZPV, and IOOPV.

```
• Initialize IACC, IZZPV, and IOOPV as zero.
• Visit all internal nodes in post order and for each node do the following:
    1. if zl*zr > 0 then add 1 to IZZPV
    2. if ol*or > 0 then add 1 to IOOPV
    3. if (zl*zr > 0) AND (IOOPTOT –ol-or >= 2) then
               if (ol+or = 0)      then add IOOPTOT - IOOPV to IACC
                                   else add IOOPTOT - IOOPV - 1 to IACC
    4. if (ol*or > 0) AND (IZZPTOT-2 < zl+zr) then
               if (zl+zr = 0)      then add IZZPTOT - IZZPV to IACC
                                   else add IZZPTOT - IZZPV - 1 to IACC
• The total number of independent accommodated four-item statements is the current value of
  IACC
```

Fig. 2.23. An algorithm for calculating the number of independent accommodated four-item statements for a given binary character on a given dichotomous tree. See text for explanation.

In the first step of the main part of the algorithm (fig. 2.23), IZZPV is increased with the number of new independent 00-pairs that must be taken into account to be able to deduce all new 00-pairs that are present above the current node. As it is a post order traversal, the left and right daughter nodes of the current node have already been visited, and consequently all 00-pairs above the left daughter and all 00-pairs above the right daughter have already been taken into account. The new 00-pairs above the current node are those with one 0-taxon in the left daughter and the other 0-taxon in the right daughter. If no such pairs exist (zl*zr=0), no new independent 00-pairs are necessary. If such pairs exist (zl*zr>0), then only a single of them must be added to the (zl-1)+(zr-1) independent pairs that were already taken into account at the left and the right daughters. In the second step, OOPV is increased in a similar way by one if ol*or exceeds zero.

In the third step, IACC is increased with the number of new independent statements that must be added to the set of accommodated statements already encountered, so that the resulting set is sufficient to deduce all newly encountered accommodated four-item statements that have their 00-part above the current node and their 11-part below the current node. In the fourth step, the same is done for the new four-item statements having their 11-part above the current node and their 00-part below the current node. In this way, all new accommodated statements are accounted for.

The expression that yields the number of new independent statements that are necessary at step three is obtained as follows: if there is no new independent 00-pair present above the node (zl*zr=0), or if there are no 11-pairs below the node (ol+or>IOOPTOT-2), then no new accommodated statements with their 00-pair above the node are present and nothing happens. Otherwise, the single new independent 00-pair above the current node yields an accommodated four-item statement in combination with any of the ob*(ob-1)/2 11-pairs below the current node, and ob-1 of these are independent. However, some of the resulting statements may have been counted at previous nodes (but then with the 11-pair above), which implies that less than ob-1 independent statements may be new. At this point, IOOPV independent 11-pairs have already been encountered above previously visited nodes, and all associated independent accommodated statements, including those with the new 00-pair above the current node, have already been added to IACC. So only the statements that result from the IOOPTOT-IOOPV remaining independent 11-pairs must be considered further. When no 1-taxa are present above the current node (ol+or=0), then all these remaining pairs are situated below the current node, and all of them yield a new independent accommodated statement in combination with the new 00-pair above the current node. When 1-taxa are present above the current node, then one of the IOOPVTOT-IOOPV remaining independent 11-pairs is a pair that has one of its 1-taxa above the current node, and the second one below. That such a pair exists, follows from fig. 2.22; that it has not been considered yet follows from the fact that the tree is traversed in post order sequence: an independent 11-pair with one taxon above and one taxon below the current node will be added to IOOPV only at some (distant) ancestor of the current node, and these ancestors have not been visited yet. The statement that results from this 11-pair and the new 00-pair above the current node is not accommodated because of the distribution of the four taxa involved: one 0-taxon in the left daughter, one 0-taxon in the right daughter, one

1-taxon below the node, and one 1-taxon above the node, either in the left or in the right daughter. Hence only IOOPVTOT-IOOPV-1 new independent accommodated statements are present. The expression in step 4 is obtained in a similar way.

The algorithm is illustrated by means of the same hypothetical tree and hypothetical character state distribution (fig. 2.12) that was used in 2.4.4 to illustrate the algorithms of figs. 2.10 and 2.11. The values of the the accumulating parameters IZZPV, IOOPV, and IACC that are obtained for the internal nodes when they are visited in the post order sequence a-b-c-d-e-f-g-h-i-j are shown in fig. 2.24. The final number of IACC, 18, is the number of independent accommodated statements for the character on the tree. One possible choice of eighteen independent accommodated statements is indicated in bold in fig. 2.14.

Node a, specifying a sister group relationship between taxa B and C, is the first node to be visited. Taxa B and C both have state one, so there is a new independent 11-pair present above the node, and IOOPV is increased by one (step 2). This pair, (BC), yields one accommodated four-item statement for each 00-pair of taxa below the node. As all six 0-taxa are present below the node (zl+zr=0), there are 6*5/2=15 such statements and any selection of five independent ones is sufficient to deduce the ten remaining ones. Because no 00-pairs have been considered previously (IZZPV=0), all of these independent accommodated statements are new, and IACC is increased by 5 (step 4). The exact identity of the five statements that are selected as independent is irrelevant: as long as they are independent among themselves, they will collectively imply all fifteen statements that they stand for; as an illustration, one possible choice is indicated in bold in fig. 2.14.

The next node, b, unites taxa D and E. Both taxa have state zero, so there is a new independent 00-pair present above the node, and IZZPV is increased by one in step 1. This 00-pair yields one accommodated four-item statement for each of the 11-pairs of taxa below the node. As all six 1-taxa are present below the node (ol+or=0), there are 6*5/2=15 such statements. Not all of these are new, however: the single statement resulting from the 11-pair BC was already counted at node a. Therefore only 4 (IOOPVTOT-IOOPV) new independent statements can be added to the set of statements already counted, and IACC is increased by 4, yielding a total of 9 (step 3; a possible choice is indicated in bold in fig. 2.14).

At node c, no new 00- or 11-pairs are encountered above the node, so no new accommodated statements will be found and nothing happens to the IZZPV, IOOPV, and IACC.

Node d unites taxa H and I. Both taxa have state zero, so there is a new independent 00-pair present above the node, and IZZPV is increased (step 1). Similar to the situation at node b, this pair yields 4 new independent statements, which are added to IACC (a possible choice is indicated in bold in fig. 2.14).

inner node	IZZPV step1	IOOPV step2	IACC step3	IACC step4
a	0	1	0	5
b	1	1	9	9
c	1	1	9	9
d	2	1	13	13
e	2	1	13	13
f	2	2	13	16
g	2	2	16	16
h	3	3	17	18
i	4	4	18	18
j	4	5	18	**18**

Fig. 2.24. The values of the accumulating parameters (algorithm of fig. 2.23) for the character and the cladogram of fig. 2.12, corresponding to postorder traversal a-b-c-d-e-f-g-h-i-j.

Node e does not yield new 00- or 11-pairs above the node, and nothing happens.

Inner node f, specifying a sister group relationship between taxa K and L, is visited next. Taxa K and L both have state one, so there is one new independent 11-pair present above the node, and IOOPV is increased. This pair yields one accommodated four-item statement for each 00-pair of taxa below the node. Two of the total number of indepedent 00-pairs (IOOPV out of IOOPTOT) have been considered previously: DE at node b and HI at node d. The remaining 3 are all present below node f (ol+or=0), yielding three new independent accommodated statements (a possible choice is indicated in bold in fig. 2.14). Note that it is not necessary to know the identity of the two 11-pairs encountered previously, or not even at which nodes they were encountered. All that must be known is their number (IOOPV) and wether or not they are all present below the current node (tested by ol+or=0).

Node g does not yield new 00- or 11-pairs above the node, so there are no new accommodated statements.

The next node, node h, is an example of the most complex case the algorithm has to deal with. As both the left and the right daughter have 0-taxa, there is a new independent 00-pair, and IZZPV is increased (step 1). Similarly, IOOPV is also increased by one in step 2. The new independent 00-pair forms an accommodated statement in combination with any 11-pair below the current node, only some of which are new. The number of new independent statements that are needed to deduce these new accommodated statements is obtained as follows: three (IOOPV) independent 11-pairs have already been considered above previously visited nodes,

including the current one, so maximally two (IOOPTOT-IOOPV) new independent 11-pairs can be below this node (to be correct, the single new 11-pair from the current node is already added to IOOPV in step 2, but it has still to be considered; this will be done in step 4). Of these two, one is a 11-pair that has one of its taxa above node h, and the other one below, leaving only a single new independent statement.
An alternative way of obtaining the same result is as follows: the new independent 00-pair above the node forms an accommodated statement in combination with any 11-pair below. As there are three taxa having state 1 below the node, there are three such pairs, and these can be represented by two independent statements. Of the three (IOOPV) independent 11-pairs already encountered above visited nodes, there is one (IOOPV-(ol+or-1)) that is present below the current node (BC, encountered at node a). Therefore only one new independent 11-pair, either BF or CF, is necessary. Similar reasoning yields one more new independent accommodated statement in step 4 of the algorithm, and finally the results for node i and j are obtained in the same way.

2.6 The problem of mutual exclusiveness

2.6.1 Introduction

In a three-item matrix (e.g. fig. 2.1), a single informative three-item statement A[BC] is represented as a binary character that has a zero-entry for A, a one-entry for both B and C, and missing entries for all remaining taxa. Harvey (1992: 350) noticed that optimization of these missing entries on a cladogram may lead to assignment of wrong states to those remaining taxa. A simple example is presented in fig. 2.25, for a character having state 0 in taxa A and B, and state 1 in taxa C, D, and E.

	X	A	B	C	D	E
full character state distribution:	0	0	0	1	1	1
three-item statement A[CE]:	0	0	?	1	?	1

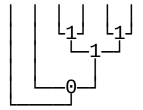


Fig. 2.25. A hypothetical character and tree. Accommodated three-item statement A[CE] wrongly assumes state 0 for taxon B. See text for explanation.

Three-item statement A[CE] is accommodated on the tree that is shown: only a single character state transition is required to explain the character state distribution of the statement. However, in order to arrive at that single step, one has to assume (1) that the three inner nodes between C and E have state 1 and the remaining inner node has state 0, and (2) that the question marks for taxa B and D stand for state 1. The latter is problematic for taxon B, in which state 0 was observed. Both Nelson and Platnick (e.g. Nelson 1992: 358, Platnick 1993: 267) replied that optimizations of these missing entries are completely beside the point: the missing entries in a column of a three-item matrix are inserted merely to be able to use widespread computer programs for testing if the corresponding three-item statement is accommodated or not.

However, Farris et al. (1995) reformulated the problem and elaborated on its consequences for a possible justification of the three-item approach. They pointed out that for any accomodated three-item statement the state assignments to inner nodes follow from the premise that the statement must be explained by inheritance (because otherwise the statement has no evidential value), and they compared results for different statements of a single character. A simple case, using the same character and the same tree as in fig. 2.25, is presented in fig. 2.26. On this tree, four out of the six available three-item statements are accommodated: B[DE], A[CD], A[CE], and A[DC]. Only three of them are independent because A[DC] follows from A[CD] and A[CE].

First consider B[DE]. Without reference to a tree, this statement hypothesizes that the presence of the plesiomorphic state in B and the apomorphic state in D and E is an indication that D and E are more related to each other than either is to B. That the statement is accommodated on the tree is taken as an indication that the tree supports the hypothesis. This is so because the fact that the statement is accommodated implies that the origin of the derived state can be traced back to an ancestor of D and E that is not an ancestor of B, in this case to inner node b; taxon B has retained the plesiomorophic condition that is present at inner nodes a and c. However, a problem arises when the same reasoning is applied to statements A[CD] or A[CE]: starting from C and either D or E, the origin of the derived state is traced back through nodes a and b to node c. So in order to explain A[CD] or A[CE] by common descent it has to be assumed that nodes a and c have the apomorphic state. However, explaining B[DE] by common descent required that a and c have the plesiomorphic state. Since all three statements are about the same character, these results are mutually exclusive: the statements that can be explained by common ancestry on the tree are either both A[CD] and A[CE] or B[DE], but not all three simultaneously.

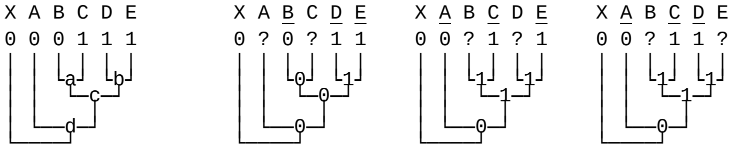


Fig. 2.26. A hypothetical character and tree. Taxon X is the hypothetical outgroup added for three-item analysis. Accommodated three-item statements B[DE] on the one hand and A[CE] or A[CD] on the other assign conflicting states to inner nodes a and c when they are explained by inheritance.

Farris et al. (1995) only discussed three-item analysis, but the criticism applies equally well to four-item analysis. Suppose that taxon X in fig. 2.26 is a real outgroup in stead of the hypothetical one that was added for three-item analysis. In that case, [XB][DE], [XA][CD] and [XA][CE] are a largest set of independent accommodated statements. However, as in the example above, [XB][DE] on the one hand and [XA][CD] and [XA][CE] on the other are mutually exclusive with respect to implied ancestral states. If the rationale for four-item analysis is to maximize the number of independent basic statements that can be explained by common ancestry, then only [XA][CD] and [XA][CE] should be accepted as accommodated statements, and [XB][DE] rejected (an alternative point of view would be to allow polymorphic inner nodes; cf. Farris 1978, Felsenstein 1979). Statements that do not exclude each other, such as [XA][CD] and [XA][CE], will be referred to as compatible statements: they are all compatible with the same set(s) of inner node state assignments. The best trees according to the four-item approach then, are those that maximize the number of compatible independent accommodated four-item statements for the data at hand.

The question then arises how to find these best trees. From the example above it is clear that the maximum number of compatible independent accommodated statements for a character depends upon the states that are assigned to the inner nodes: with states assigned as in the second tree of fig. 2.26, only a single compatible statement is accommodated; with states assigned as in the third tree, two compatible independent statements are accommodated. Therefore, any algorithm to calculate this maximum number must at the same time calculate the inner node assignments that achieve this maximum. The logic for doing so is developed in the following sections.

2.6.2 Calculating the number of independent compatible taxon pairs for fixed inner node state assignments

The algorithm of fig. 2.23 consists of a post order traversal of the given tree, and at each inner node the number of new independent accommodated statements for the given character is calculated from the numbers IZZPTOT and IOOPTOT and the numbers zl, zr, IZZPV, ol, or, and IOOPV for that node. When for each inner node a state **s** is specified, then only some of the independent 00-pairs (IZZPV and IZZPTOT) and 11-pairs (IOOPV and IOOPTOT) will be compatible with these assignments. Appropriate redefinitions of zl, zr, ol, and or will enable redefinitions of IZZPV, IZZPTOT, IOOPV, and IOOPTOT such that only compatible taxon pairs are taken into account. These redefinitions are as follows (the starting 'c' or 'C' stands for compatible; the new part of the definition is in italics; a 00-pair is called compatible if the two taxa involved are connected to each other through a series of inner nodes that each have state 0; a compatible 11-pair is defined analogously):

- **czl**: the number of 0-taxa at or above the left daughter node *that are connected to the current node through a series of inner nodes that all have state 0*
- **czr**: the number of 0-taxa at or above the right daughter node *that are connected to the current node through a series of inner nodes that all have state 0*
- **CIZZPTOT**: the total number of *compatible* independent 00-pairs
- **CIZZPV**: the number of *compatible* independent 00-pairs above the current and previously visited nodes (each pair has also all its connecting nodes above the current or previously visited nodes)
- **col**: the number of 1-taxa at or above the left daughter node *that are connected to the current node through a series of inner nodes that all have state 1*
- **cor**: the number of 1-taxa at or above the right daughter node *that are connected to the current node through a series of inner nodes that all have state 1*
- **CIOOPTOT**: the total number of *compatible* independent 11-pairs
- **CIOOPV**: the number of *compatible* independent 11-pairs above the current and previously visited nodes (each pair has also all its connecting nodes above the current or previously visited nodes)

All these numbers can be calculated using an algorithm that performs a post order traversal of the tree (fig. 2.27). The numbers **cz** and **co** are defined as follows: if the given state s at an inner node is zero (one), then cz (co) is the number of 0-taxa (1-taxa) above that node that are connected to that node through a series of inner nodes that all have state 0 (1), otherwise cz (co) equals zero. In this way the value of cz (co) of any inner node above the basal node is the value of czl (col) or czr (cor) of its ancestor (depending upon the left or right position with respect to the ancestor).

```
• Initialize CIZZPV, and CIOOPV as zero.
• Visit all internal nodes in post order and for each node do the following:
    • initialise cz and co as zero
    • if s=0   then   add czl+czr to cz
                      if (czl>0) AND (czr>0) then add 1 to CIZZPV
               else   add col+cor to co
                      if (col>0) AND (cor>0) then add 1 to CIOOPV
• Do the following for the basal node
    • if the outgroup has state 0 AND (czl+czr > 0) AND (s = 0) then add 1 to CIZZPV
    • if the outgroup has state 1 AND (col+cor > 0) AND (s = 1) then add 1 to CIZZPV
• CZZPTOT is equal to the current value of CIZZPV, COOPTOT to the current value of CIOOPV
```

Fig. 2.27. An algorithm for calculating CIZZPTOT and CIOOPTOT for a given binary character on a given dichotomous tree with a given set of inner node state assignments. See text for explanation.

As the algorithm proceeds, new compatible independent 00-pairs are accumulated in CIZZPV, and new 11-pairs in CIOOPV. Detection of these new compatible independent pairs is as the detection of new independent pairs in the algorithm of fig. 2.23, but with a supplementary test on the state s present at the current node, and using the redefinitions of zl, zr, ol, and or.

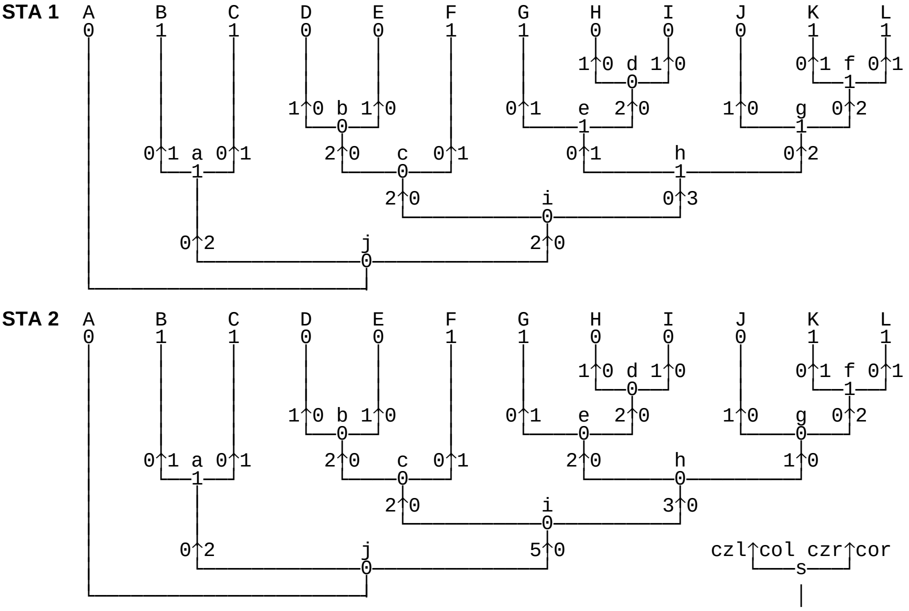


Fig. 2.28. The example cladogram and character of fig. 2.12, showing the numbers czl, czr, col, and cor for two different sets of inner node state assignments, STA1 and STA2.

As a result, only the compatible independent pairs are counted. After all inner nodes have been visited, CIZZPV and CIOOPV contain the number of compatible independent 00-pairs and 11-pairs above the basal node of the tree. Because the outgroup is situated below the basal node, it has still to be checked then if there exists a compatible 00-pair or 11-pair in which the outgroup is involved. If this is the case, either CIZZPV or CIOOPV has to be increased by one to obtain the values of CIZZPTOT and CIOOPTOT.

The definitions and the algorithm are illustrated (figs. 2.28 and 2.29) for two different sets of inner node state assignments (STA1 and STA2; STA2 requires three extra steps and is a most parsimonious reconstruction; STA1 requires one extra step more) for the example character and tree of fig. 2.12. In fig. 2.28, the numbers czl, czr, col, and cor for STA1 and STA2 are indicated along the branches of the tree. The values for CIZZPV and CIOOPV corresponding to post order traversal a-b-c-d-e-f-g-h-i-j are given in fig. 2.29. In both STA1 and STA2, the final value of CIZZPV at basal node j has to be increased by one to obtain CIZZPTOT because (1) the basal node has state 0, (2) the outgroup taxon has state 0, and (3) there are 0-taxa above the basal node that are connected to the basal node through a series of inner nodes that all have state 0.

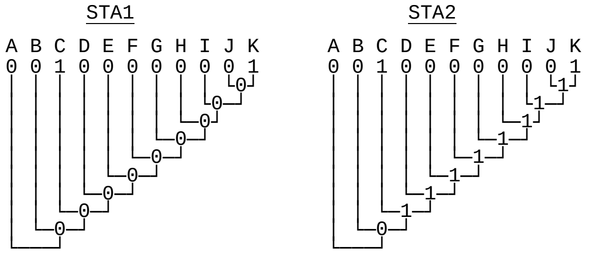
	STA1		STA2	
inner node	CIZZPV	CIOOPV	CIZZPV	CIOOPV
a	0	1	0	1
b	1	1	1	1
c	1	1	1	1
d	2	1	2	1
e	2	1	2	1
f	2	2	2	2
g	2	2	2	2
h	2	3	3	2
i	2	3	4	2
j	2	3	4	2
	CIZZPTOT=2+1=3		CIZZPTOT=4+1=5	
	CIOOPTOT=3		CIOOPTOT= 2	

Fig. 2.29. The values of CIZZPV, CIZZPTOT, CIOOPV, and CIOOPTOT (algorithm 2.27) for the example tree and character of fig. 2.12 and two different sets of inner node state assignments STA1 and STA2 (fig. 2.28); CIZZPV and CIOOPV correspond to postorder traversal a-b-c-d-e-f-g-h-i-j. See text for explanation.

Once czl, czr, col, cor, CIOOPV, CIOOPTOT, CIZZPV and CIZZPTOT have been calculated, the number of accommodated independent four-item statements for the given inner node state assignments could be calculated using an algorithm that is similar to the algorithm of fig. 2.23. However, the compatibility constraint has also repercussions on what can be considered as an independent statement, as will be discussed below, and therefore such an algorithm would overestimate the number of independent compatible accomodated statements.

2.6.3 Dependency revisited

From the above redefinitions it follows that no compatible 00-pair of taxa can have an inner node with state 1 in the path between its two taxa, and no compatible 11-pair can have an inner node with state 0 along the path between its two taxa. As a consequence, any compatible 00-pair will form an accommodated statement with any compatible 11-pair. The resulting CIZZPTOT*CIOOPTOT compatible accomodated statements are a subset of those that are counted in the algorithm of fig. 2.23, which were the independent accomodated statements without compatibility constraint. However, once the compatibility constraint is taken into account, some of these CIZZPTOT*CIOOPTOT accommodated statements may no longer be independent. Indeed, if a single compatible independent 00-pair is coupled with each compatible 11-pair, and a single compatible 11-pair with each compatible 00-pair, then the resulting CIZZPTOT + CIOOPTOT -1 different statements are sufficient to deduce all remaining accommodated statements (note that a similar relationship does not hold when the inner node state assignments are not considered; cf. 2.5.3). Therefore, the set of inner node state assignments that maximizes the number of independent accommodated four-item statements is the one that maximizes the total number of independent pairs (CIZZPTOT + CIOOPTOT) that are retained on the tree. However, there is one severe restriction to this conclusion: if possible, both CIZZPTOT and CIOOPTOT must exceed zero.

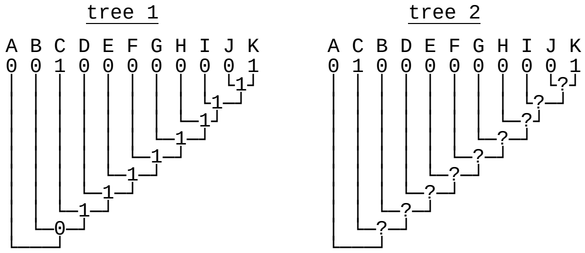


	STA1	STA2
CIZZPTOT	8	1
CIOOPTOT	0	1
total number of independent pairs	**8**	**2**
accommodated statements	**0**	**1** [AB][CK]

Fig. 2.33. The inner node state assignments that retain the highest number of independent pairs (STA1) on a given tree are not necessarily those that retain the highest number of independent accommodated statements (STA2).

As shown in fig. 2.30, this restriction follows directly from the inherent relational aspect of four-item analysis (i.e. 00-pairs and 11-pairs are opposed to each other) and seems to be a purely methodological constraint of four-item analysis as defined thusfar. Whenever either CIZZPTOT or CIOOPTOT drop to zero (fig. 2.30, left), the number of accommodated statements drops to zero also. Therefore any set of inner node state assignments that manages to keep both CIZZPTOT and CIOOPTOT greater than zero will accommodate more statements and should be preferred (fig. 2.30, right).

As a direct consequence of this restriction, four-item analysis may result in very counterintuitive hypotheses concerning the evolution of characters. E.g. in the example of fig. 2.30, the retention of the single accommodated statement [AB][CK] requires that the set of inner node state assignments STA2 is preferred over STA1, which in turn implies that character state zero arose independently in each of the seven lineages leading to taxa D to J. STA1, on the other hand, requires only that character state one arose independently in two lineages, but accepting this reconstruction implies a loss of accommodated statements (the pectinate series D-J can be extended ad libitum, making the implications for the evolution of the character under four-item analysis more and more unrealistic).



accommodated statements 1 [AB][CK] **0**

Fig. 2.31. A marginal increase in the number of accommodated statements may impose extremely strong restrictions on possible inner node states. See text for explanation.

A similar effect is illustrated in fig. 2.31, where two different trees are compared (the two trees differ only in the positions of taxa B and C). With the inner node state assignments as shown in the figure, one four-item statement is accommodated on tree 1 (it is easy to verify that this is the best possible result for this tree: any other assignments lead to a loss of the accommodated statement). However, in order to accommodate this statement, it has to be assumed that state zero arose independently in each of the seven lineages leading to one of the taxa D-J (and once again the pectinate series can be extended ad infinitum). On tree 2, on the other hand, no four-item statements can be accommodated, whatever the inner node state assignments. Therefore, the character state distribution can be explained by a convergent evolution of state one in the two lineages leading to taxa C and K, which requires only two steps. Summarizing, in terms of accommodated statements tree1 is slightly better than tree2, but it implies very unrealistic assumptions about the evolution of the character. Moreover, these assumptions are not implied by tree 2.

2.6.4 Removing the relational constraint: back to the standard approach

So in order to maximize the number of independent accommodated statements for a given character on a given tree, the set of inner node state assignments that maximizes CIZZPTOT + CIOOPTOT has to be identified within the constraint that, if possible, both CIZZPTOT and CIOOPTOT must exceed zero. This constraint has no obvious interpretation and seems to be a methodological constraint of four-item analysis as defined thusfar. Because it may lead to very counterintuitive results in a number of cases (e.g. figs 2.30 and 2.31), it might be proposed to drop it and to identify the set(s) of inner node state assignments that simply maximize(s) (CIOOPTOT + CIZZPTOT). Because the constraint reflects the basic hypothesis that distinghuishes the four-item approach from the standard approach to parsimony analysis, it should come as no surprise that removing the constraint reduces the approach to standard parsimony analysis. Indeed, CIOOPTOT + CIZZPTOT is maximal under any most parsimonious reconstruction of the inner node states as defined by the standard approach, and therefore minimizing homoplasy comes down to maximizing CIOOPTOT + CIZZPTOT.

The following argument clarifies this point. Assume a character that has zt 0-taxa and ot 1-taxa. A priori this character state distribution hypothesizes that there are (zt-1) compatible 00-pairs, and (ot-1) compatible 11-pairs: any tree on which the character has no homoplasy can be subdivided into two parts, one in which all inner nodes have state 0, and a second in which all inner nodes have state 1. The branch between both parts of the tree is the branch along which the single state transition occurs. Next assume a tree on which a most parsimonious reconstruction requires one step of homoplasy. This most parsimonious reconstruction obviously implies two state transitions, and these two state transitions subdivide the tree into three parts: either one part in which all nodes have state zero and two parts in which all nodes have state 1 (one independent compatible 11-pair is lost), or two parts in which all nodes have state zero and one part in which all nodes have state 1 (one independent compatible 00-pair is lost). Which of the two possibilities occurs, depends on the tree and on the most parsimonious reconstruction that is chosen, but in both cases there is exactly one compatible pair that is lost. In a similar way, any subsequent step of homoplasy will imply the loss of one more compatible pair. Summarizing, if h stands for the amount of homoplasy that is implied by the tree, the character and the inner node states, then the amount of compatible pairs that are retained (i.e. CIZZPTOT + CIOOPTOT) is equal to (nz-1) + (no-1) - h. Obviously this expression is maximal when h is minimal, i.e. whenever a most parsimonious reconstruction is chosen. From this point of view, standard parsimony analysis searches for the tree(s) on which the highest amount of compatible independent pairwise similarities is (are) retained, and as such it can be characterized as two-item analysis.

The above considerations do not simply mean that four-item analysis and the standard approach to parsimony analysis are one and the same thing. Indeed, the important fact is that an approach that decomposes a character state distribution into its basic independent compatible statements that are still informative with respect to cladistic grouping behaves very anomalously (cf. figs. 2.30 and 2.31), and that removing these anomalies comes down to dismissing the whole idea of factorization into basic informative statements. This conclusion is not restricted to four-item analysis, but holds also for three-item analysis. This is illustrated by interpreting the trees in fig. 2.31 in terms of accommodated three-item statements, which is easily done by taking taxon A as the all-zero outgroup that is added for the sake of three-item analysis (cf. 2.2). In that case, the only three-item statement that is accommodated on tree 1, B[CK], requires seven independent origins of state zero in the lineages leading to taxa D-J (note that these reversals contradict the Camin-Sokal interpretation that was proposed for three-item analysis in the absence of the compatibility requirement).

2.7 Summary and conclusion

Three-item analysis is a method that was introduced as a novel approach to parsimony analysis in both biogeography (Nelson & Ladiges 1991a, 1991b) and systematics (Nelson & Platnick 1991). In systematics, it rests on the two following assumptions: (1) any informative character state distribution can be decomposed into a series of more basic statements that each are still cladistically informative; (2) the use of such basic statements will improve the sensitivity of parsimony analysis. Following its introduction, three-item analysis has been severely criticized because of three basic defects: (1) it is flawed because it presupposes that character evolution is irreversible; (2) it is flawed because basic statements that are not logically independent are treated as if they are; (3) it is flawed because some of the three-item statements that are considered as independent support for a given tree may be mutually exclusive on that tree. In this chapter, it is shown that these criticisms only relate to the particular way that the approach was implemented by Nelson & Platnick (1991), and an alternative implementation that solves each of the three basic problems is derived. However, the resulting method is not an improvement over standard parsimony analysis. It is identical to the standard approach but for one small constraint, which is a highly unnatural restriction on the maximum amount of homoplasy that may be concentrated in a single character state. As this restriction is a direct consequence of the decomposition of character state distributions into a number of cladistically informative basic statements, it is concluded that any approach that is based on such decompositions will be defective.

3. HOMOPLASY-BASED WEIGHTING SCHEMES[5]

3.1 Introduction

Farris (1990: 92) defined the weight of a character in parsimony analysis as "the numerical change in the parsimony criterion produced by adding one step in that character, and weight is intended to reflect the importance of a step as evidence on phylogenetic relationships". This is in line with his earlier demonstrations (Farris 1969, Kluge and Farris 1969, Farris 1983) that "parsimony does not preclude weighting, but rather ... requires weighting" (Goloboff 1993a: 83; see also Goloboff 1995). Indeed, the principle of parsimony in itself does not imply or presuppose that all characters yield equally strong evidence on phylogenetic relationships and hence should require equal weights (Farris 1983: 11). A most parsimonious cladogram stands as a hypothesis with maximum explanatory power only inasmuch as all characters have been assigned the weights they deserve. Therefore, the relevant question with respect to weighting in parsimony analysis is not if characters should be weighted, but how and how strongly they should be weighted. From this point of view, the common practice to use equal weights for all characters is as much a weighting decision as any other weighting scheme, and its justification should receive as much attention as the justification of any other weighting scheme.

Differential weighting of characters has sometimes been proposed as a means to select among multiple most parsimonious cladograms that are obtained under equal weights (e.g. Rodrigo 1992, Sharkey 1993: 212, Sang 1995, Turner 1995, Turner & Zandee 1995). However, a corollary of the above is that differential weighting is not simply a means to reduce the number of most parsimonious trees under equal weighting, or to impose an ordering upon such trees. Indeed, even if there is only a single most parsimonious tree under equal weights, the issue of character weighting must be considered, and if properly weighted characters indicate more or other trees than found under equal weighting, these should be preferred regardless of the results under equal weights (e.g. Farris 1983: 10-11, Carpenter 1988: 293, 1994: 216, Rodrigo 1989: 101-102, Goloboff 1993a: 83, 1995).

[5] The basic idea of this chapter - minimizing weighted homoplasy in stead of maximizing fit - was presented at the XIVth meeting of the Willi Hennig Society (July 30 - August 3, 1995, College Station, Texas; see De Laet & Smets 1995), albeit mainly in terms of four-item analysis.

Many different approaches to character weighting in parsimony analysis have been proposed in the past (see e.g. Simon et al. 1994: 666-670 and Brower & DeSalle 1994: 703-706 for an overview and some comments), and these various approaches have been grouped under various names and according to various criteria (see e.g. Neff 1986, Rodrigo 1989, Sharkey 1989, Albert & Mishler 1992, Simon et al. 1994, Brower & DeSalle 1994). Goloboff (1993a) stressed the underlying rationale for assessing the reliability of characters, and on that basis he distinguished between a priori, compatibility-based, and homoplasy-based weighting methods.

A priori weighting involves some kind of assessment of character reliability prior to the parsimony analysis: independent of the degree of congruence with other characters in the data matrix, each character is assigned a weight that reflects the confidence in the hypothesis of primary homology (de Pinna 1991) that is expressed by the character. Therefore characters that are better studied deserve in general higher weights than poorly known characters (Neff 1986). This is conceptually sound and clear, but it is far from obvious how to estimate how well-studied a character is, or how to translate this estimate into a numerical weight. From a statistical point of view, there is a linear relationship between the a priori weight of a character and the negative logarithm of a function of its transformational probabilities, at least if these probablilities are not too large (the exact form of the function depends upon the underlying models of evolutionary processes; e.g. Farris 1978, Felsenstein 1981; see Albert et al. 1993: 755-756 for a discussion). However, this logarithmic relationship does not solve the problem of estimating prior character weights, but shifts it to an other level, viz. the estimation of the transformational probablities and the choice and justification of the underlying evolutionary models.

Once assigned, the a priori weights remain fixed throughout the further cladistic analysis and serve to determine the relative importance of one step of homoplasy in different characters. As stated above, parsimony analysis using equal weights is often seen as an unweighted approach, but it involves the a priori decision that all characters deserve equal weights. This seems a valid starting point when all characters in the data matrix are carefully defined and coded, as in that case there are mostly no obvious reasons to expect that some characters should conform better to a general pattern than other. However, the very fact that most analyses of real data indicate a rather wide range of homoplasy in the characters contradicts that expectation (Goloboff 1995: 96).

The two other approaches, compatibility-based and homoplasy-based weighting, estimate the weights of the characters directly from the information that is present in the data set, and not on the basis of some independent assessment, as in the a priori approach (a similar distinction has also been made in phenetic analyses; see Williams & Dale 1964 for an example). It follows that a priori weighting on the one hand and compatibility- or homoplasy-based weighting on the other are not exclusive and can be easily combined.

Both approaches agree that the reliability of an individual character can be estimated from the degree to which the character conforms to the hierarchical pattern that is implied by the data as a whole (this idea was aptly expressed by Patterson 1982: 44: “we do not need to weight homologies: they weight themselves”). However, they differ fundamentally in the precise way this conformation to hierarchical structure is measured. Compatibility-based approaches (e.g. Farris 1969: 381-382, Penny & Hendy 1985, Gauld & Underwood 1986, Sharkey 1989, 1994, Wilkinson 1994c) are tree-independent and derive the weight of a character basically from the number of incompatibilities that the character displays with respect to the other characters in the data set (Le Quesne 1969, 1983). In homoplasy-based weighting schemes, such as successive weighting (Farris 1969; see also Williams & Fitch 1989, 1990; but not Sankoff & Cedergren 1983, as wrongly asserted by Simon et al. 1994) or implied weighting (Goloboff 1993a), the weight is derived from the amount of homoplasy that the character displays on one or more cladograms that are selected during the process of weighting. Arguments against homoplasy-based methods (e.g. Sharkey 1994: 528-529) are twofold: first it may be argued that there are no compelling reasons why measures of reliability should be tree-derived; second, even if such reasons were acknowledged, tree-derived methods would suffer from circular reasoning because they require a priori knowledge of the trees that one is looking for. However, I agree with Goloboff (1993a) that on theoretical grounds and when properly implemented, homoplasy-based weighting schemes are superior to compatibility-based approaches and free from circular reasoning.

The basic reasoning of homoplasy-based weighting schemes (Farris 1969: 374-376, 383-384) is as follows. A set of characters is said to have a high hierarchical correlation if all characters of the set are all highly consistent with a single branching pattern. Common descent is the foremost process that is expected to lead to hierarchical correlation, but other possibilities exist: e.g. characters that are functionally or adaptively correlated may display a hierarchical correlation that is different from the phylogenetic branching pattern. Assume first a data set that contains many characters that are, as far as can be ascertained, independent. Some of these characters will be highly consistent with the phylogenetic branching pattern, and by definition all these characters will have a high hierarchical correlation. Other characters in the data set may be poor indicators of phylogeny because their character state distributions are not consistent with the phylogenetic branching pattern. However, each character that does not fit the phylogenetic pattern very well will deviate in its own idiosyncratic way from the phylogenetic hierarchy, and only by chance two badly fitting characters will deviate in the same direction. Therefore, it may be assumed that over the complete data set the hierarchical correlation due to phylogeny will outweight other hierarchical correlations. As long as enough characters are taken into account, this may be expected to hold even when some of those badly fitting characters are functionally, developmentally, genetically, or in any other way correlated. Badly fitting characters will by definition display a relatively high amount of homoplasy on a branching pattern that depicts the phylogenetic relationships, and therefore the amount of homoplasy can be interpreted as a measure of the cladistic reliability of a character, and characters with high amounts of homoplasy may be downweighted with respect to characters that are less homoplasious. In this way, the differential weighting "is based on the simple idea that characters which have failed repeatedly to adjust to the expectation of hierarchic correlation are more likely to fail again in the future, and so they are less likely to predict accurately the distribution of as yet unobserved characters" (Goloboff 1993a: 84).

However, because the homoplasy of a character cannot be calculated without reference to a branching pattern and because these branching patterns themselves are estimated from the character state distributions, a case of deadlock arises: in order to determine the reliability of the characters, the branching pattern that is implied by the data must be known, but in order to know this implied branching pattern, the reliability of the characters has to be determined first. As noted by Goloboff (1993a), Farris' successive approximations approach to character weighting was the first proposal to solve this apparent dilemma.

In this chapter, I will first discuss successive weighting (Farris 1969) to provide a background for Goloboff's method of estimating implied weights. After treating Goloboff's method, I will propose an extension of his approach by applying the estimated implied weights effectively as weights during parsimony analysis - a step that Goloboff (1993a, 1993c, 1995) did not make or consider - and argue that this approach is more in line with cladistic philosophy (Farris 1983) than Goloboff's original method. Several issues related to weighting are discussed from this point of view.

3.2 Successive weighting

Successive weighting (Farris 1969) consists of finding the most parsimonious trees for a data set using an initial set of character weights, and then reweighting the characters according to their homoplasies on these trees: characters with plenty homoplasy are deemed less reliable than characters that have less homoplasy. Next, the data are reanalyzed using these new weights and the cycle of recalculating weights and reanalyzing the data is repeated until the weights remain unchanged in two successive rounds; the trees that are obtained with these final weights are considered as the most parsimonious trees for the data. Within this general framework several issues require further specification, some of which are shortly discussed below.

3.2.1 Weighting function

A first problem that must be addressed in any homoplasy-based weighting scheme is the exact nature of the weighting function. It is clear that the weight that is assigned to a character should decrease as the character displays more and more homoplasy, but which kind of decreasing function should be used is less obvious. Farris (1969: 379) classified the various functions he considered into three basic groups: convex, linear, and concave (fig. 3.1). Convex weighting functions can be thought of as functions that weight strongly against very unreliable characters, while concave functions weight strongly in favor of very reliable characters; linear weighting functions are intermediate between convex and concave (Farris 1969: 380). On the basis of simulations he concluded that concave weighting functions were superior to linear or convex ones.

A typical concave decreasing function of the homoplasy is provided by the unit consistency index ci (Farris 1969: 375), which was used as weighting function in the

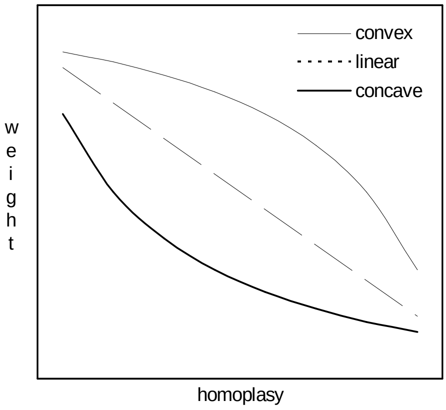


Fig. 3.1. Convex, linear, and concave weighting functions.

computer program PHYSIS (Mickevich & Farris 1982, fide Farris 1989). For any character, the unit consistency index with respect to a particular cladogram is equal to the ratio of m, the minimum number of steps the character can have on any cladogram, and s, the number of steps the character displays on the cladogram under consideration. Because the number of steps on a cladogram is equal to the sum of m and the homoplasy of the character (h), the unit consistency index is a hyperbolic function of the homoplasy.

Besides the consistency index, other concave functions of the homoplasy have been used as weighting functions for successive weighting. In the computer program Hennig86 (Farris 1988), the rescaled consistency index rc is used. This index is defined as the product of the consistency index and the retention index ri (which is in turn defined as (g-m-h)/(g-m), with g the homoplasy of the character on a completely unresolved tree). The rescaled consistency index ri was preferred over ci because rc, and hence the weight that is assigned to the character, equals zero whenever the character displays its maximum amount of homoplasy (Farris 1989: 418). In PAUP (Swofford 1993), the retention index ri is offered as a third option besides ci and rc.

Goloboff (1993a: 86, 89; see 3.3.4) tried to find a rational basis to decide which weighting function should be preferred and noticed that all three indices (ci, ri, and rc) have defects in the sense that characters that display the same amount of homoplasy may receive different weights when they have different values of m and/or g. As an example, the retention index assigns higher weights not only to characters with less homoplasy, but also to characters with more informative variation (g-m). In extreme cases this may lead to a choice of trees on the basis of the amount of informative variation and not on the basis of homoplasy. Similarly, when the consistency index is used as weighting function, not only characters with low homoplasy receive a high weight, but also those characters with a high value of m relative to the homoplasy.

3.2.2 Initial set of weights

In order to obtain a set of initial weights that already reflect as far as possible the cladistic reliability of the characters, Farris (1969:380-382) proposed to estimate the initial weight of each character from a tree-independent measure of its hierarchical correlation with the other characters of the data set. More precisely he suggested to determine the initial weights as a concave function of the number of incompatibilities of the characters (Le Quesne 1969, 1983). However, in practice successive weighting is often performed using equal weighting as starting point (e.g. Carpenter 1988, Anderberg & Ståhl 1995, Endress et al 1996).

Because the final solution under successive weighting may depend on the choice of the initial weights (Farris 1969), the situation may arise that two different sets of plausible initial weights result in two different solutions. This leads to the question how both solutions in such cases should be compared. Indeed, each set of stable final weights results in its own set of trees that are shortest under the weights that are used, but the tree lengths among the two weighting schemes are not comparable precisely because different weights are applied (Goloboff 1993a: 85). Within the framework of successive weighting, it seems the most logic to accept both solutions as equally good though conflicting explanations of the data.

3.2.3 Differential weighting of state transformations in multistate characters

3.2.3.1 Ordered characters

Carpenter (1988) drew attention to the fact that the way in which ordered multistate characters are coded (additive binary or ordinal; cf. Mickevich & Weller 1990) can influence the final stable solution under successive weighting because ordinal coding may give more influence to these characters simply because they are coded that way. The net result of using additive binary coding for ordered multistate characters is that each character state transformation that is allowed by the character state tree is weighted separately on the basis of the amount of homoplasy present in that transformation (Farris 1969: 382). When the ordered multistate characters are not decomposed into their binary constituents, the weighting procedure weights some kind of average character state transformation on the basis of some kind of average character state transformation homoplasy. This is less precise, and therefore ordered multistate characters should be coded in a binary additive way. However, it should be noted that the sensitivity to way of coding is not an intrinsic property of successive weighting itself: it follows from the way it is implemented in particular computer programs, and successive weighting might as well be implemented such that ordinal and additive binary coding give the same results.

3.2.3.2 Unordered characters

In successive weighting, unordered multistate characters (Fitch 1971) are mostly treated in a way that is comparable to the treatment of ordered characters that are not decomposed into their binary constituents (e.g. Anderberg & Ståhl 1995, Endress et al. 1996): the characters are weighted in their entirety on the basis of the homoplasy that is present in the whole character, and no attempts are made to subdivide the homoplasy over the various possible state transformations. However, the possibility for doing so exists: based on Sankoff's (1975, see also Sankoff & Cedergren 1983) cost matrices, Williams & Fitch (1989, 1990) described a successive approximations approach to character weighting - which they called dynamic weighting - that allows within-character differential weighting of state transformations in nucleotide sequences (in addition to weighting full characters similarly as in Farris' approach). The method starts with a cost matrix that describes the initial weights of all possible state transformations (see Williams & Fitch 1990: 618 for various ways of initializing these weights), and in each round of successive weighting the weight of each state transformation is recalculated as a concave decreasing function of the number of times the transformation occurs in all most parsimonious trees in all positions of the sequence. As an option it can be requested that the cost matrices remain symmetric, i.e. that transformations j→k and j←k have identical weights for each pair of states j and k. Because the reweighting procedure can result in cost matrices that are not metric, a special "numbing" procedure ensures that no unobserved intermediate states are postulated (Williams & Fitch 1990: 615; a non-metric cost matrix is a matrix in which the triangle inequality is violated for some sets of three states a, b, and c: the weight for transformation a-b plus the weight for transformation b-c is less than the weight for transformation a-c; as a result the length of a tree may be shortened by assuming that all transformations between states a and c pass through state b, even if this intermediate state is unobserved; cf. Wheeler 1993).

The main reasons that the approach of Williams & Fitch is not often used seem to be the limited distribution of their computer programs and the fact that the use of cost matrices (Sankoff 1975) results in relatively slow algorithms. Sankoff's (1975) original algorithm also allows for dynamic alignment during parsimony analysis, but even if this possibility is not implemented, as in Williams & Fitch's (1990) WTSUBS program or in PAUP's step matrix option (Swofford 1993; PAUP's successive weighting, however, cannot be used in combination with step matrices), a cost matrix algorithm remains slower than Fitch's (1971) basic algorithm for unordered characters. Recently, Goloboff (1996a, 1996b) provided two computer programs for fast approximations of parsimony analysis under Sankoff costs, so it can be expected that the problem of execution time will become less important in the future.

In Williams and Fitch's (1989, 1990) approach, the weight for the state transformation between any pair of two states is recalculated on the basis of the number of times the transformation is present over all positions in the sequence. As such the approach assumes that the substitutional processes are equal across positions, or at least it is not sensitive to possible differences. Because the assumption of equal processes across positions can be violated for a number of reasons (see Swofford et al. 1996: 503 for a summary), Swofford et al. (1996: 503) argued that in some cases it may be necessary to subdivide the positions into a number of classes (e.g. first, second, and third positions of codons in protein coding DNA sequences) and to use a separate cost matrix for each class, which raises the general problem how to determine the exact number and nature of these classes. This problem can be avoided by using a separate cost matrix for each individual position, which requires only a minor modification of Williams and Fitch’s algorithm. The resulting method can be applied to any kind of unordered character, including morphological ones. It could be argued that the assessment of the transformational probabilities of state transformations within a character on the basis of the homoplasies of the transformations suffers from a too low sample size, or that it requires the assumption of constancy of processes across lineages (see e.g. Maddison & Maddison 1992: 63, Swofford et al. 1996: 503). However, such arguments are not specific to within-character differential weighting. They apply equally well to differential weighting of full characters and should be addressed at that level.

3.2.4 Multiple most parsimonious trees

Because the weighting function for a single character can have different values on different trees, a problem arises if more than one tree is present in the set of most parsimonious trees. Different solutions to this problem have been proposed and implemented: Carpenter (1988) used mean values over all most parsimonious trees, as originally proposed by Farris (1969). In Hennig86 (Farris 1988), the maximum value among all most parsimonious trees is used. In PAUP (Swofford 1993) it is possible to choose between the maximum, the minimum and the mean value.

From a logical point of view, using the maximum value is to be preferred over the mean or the minimum value. This is best illustrated using de Pinna’s (1991) distinction between primary and secondary homology (see chapter 1): selecting or delineating a character for cladistic analysis is equivalent to developing a conjecture of primary homology on the basis of observed similarities. The very fact of including a character in a data set expresses the prior expectation that the state distribution of that character will conform to a general hierarchical pattern, and until this is proved false by an analysis of the hierarchical correlation of numerous primary homologies there is no reason to abandon that starting assumption. The degree to which that expectation is refuted beyond doubt in a set of most parsimonious trees is not reflected in the worst or the mean value of the consistency index, but in the best.

Choosing the worst or the mean value for reweighting the character is like admitting halfway that the primary homology conjecture one choose to put in the data set was not that good after all.

When the mean values are used, still another problem arises : programs such as Hennig86 (Farris 1988) and PAUP (Swofford 1993) can produce most parsimonious trees that contain unsupported branches (Coddington & Scharff 1994). Whenever this is the case the calculated means over all trees may be biased.

However, apart from the problem of deciding whether maximum, mean, or worst values should be used, there is a more fundamental problem that arises when multiple most parsimonious trees are present: the final trees that are obtained may not be self-consistent (Goloboff 1993a, 1995; self-consistency is not to be confused with statistical consistency, on which see Kim 1996). This means that on any individual most parsimonious tree there may be characters with high weights that have nevertheless much homoplasy on that cladogram and vice versa.

3.3 Implied weights

Goloboff (1993a: 86) showed that self-consistency is a necessary condition to obtain a sound weighting scheme, and he proposed a new homoplasy-based weighting approach that, contrary to successive weighting, always yields self-consistent cladograms; the character weights that are obtained using this method are called implied weights or character fits.

3.3.1 Self-consistency

The criterion of self-consistency can be described as follows: “A tree which is shortest under the weights it implies is a tree which resolves character conflict in favor of the characters which, on the tree itself, have less homoplasy, and it is therefore *self-consistent*. If the tree is not shortest under the weights it implies, the tree is self-contradictory: it resolves character conflict in favor of exactly those characters the tree is telling not to trust” (Goloboff 1993a: 85). That self-consistency is a necessary condition follows from the basic assumption that the reliability of a character is related to the amount of homoplasy the character displays on the phylogenetic branching pattern (see 3.1) and from the fact that a cladogram is a hypothesis of those phylogenetic relationships. Indeed, any individual hypothesis of phylogenetic relationships should be evaluated on its own merits, and not be criticized because competing hypotheses predict other character reliabilities (which would amount to circular reasoning). Therefore, if homoplasy-based character weights are to be determined to evaluate the strength of any particular cladogram, these weights should be derived solely from that single cladogram.

3.3.2 Character fit as a concave decreasing function of the homoplasy

In a next step, Goloboff argues that self-consistency is automatically obtained in a non-iterative way if (1) the implied weight of a character is determined as a concave function of its homoplasy and (2) the sum of all implied character weights is maximized. To the implied weight of a character he refers as the fit of the character, and to the sum of all character fits as the total fit. The argumentation is as follows (Goloboff 1993a: 86).

First assume that the fit of a character on a tree is measured as a linear decreasing function of its homoplasy on that tree (fig. 3.2, middle). In this case, each step of homoplasy in the character will decrease its fit with an equal amount, and the total fit of the data set will decrease by that same amount. As a result, the trees with the greatest total fit for the data set are simply the most parsimonious trees (using whatever set of a priori weights that is thought to be appropriate). This linear decreasing character fit (and hence standard parsimony analysis) is not sensitive to the different reliabilities of the characters, which is easily shown by example (slightly modified from Goloboff 1993a: 86).

Assume a data set that has only two most parsimonious trees, X and Y, and assume that only two characters, c1 and c2 have different numbers of homoplasious steps (extra steps) on each of those trees: c1 has one extra step on tree X and two extra steps on tree Y, while c2 has fifteen extra steps on tree X, but only fourteen on Y. So even though both trees differ in the distribution of the extra steps over the characters, they have the same total amount of extra steps and therefore are considered equally good explanations of the data, which means that the character conflict between characters c1 and c2 remains unresolved. However, based on the concept of hierarchical correlation (Farris 1969; see 3.1), the character conflict might be resolved in favor of c1, leading to a slight preference for tree X over tree Y. Indeed, character c1 has an almost perfect hierarchical correlation with tree X, and hence with many other characters of the data set (otherwise X would not be a most parsimonious tree). Character c2, on the other hand, has a very poor hierarchical correlation with tree X and hence with all characters that are hierarchically correlated with that tree. On tree Y, character c2 performs better than on tree X in the sense that it requires one step of homoplasy less, but in terms of hierarchical correlation with tree Y it behaves almost as poor as it did on tree X (15 extra steps on tree X, which is bad, compared to 14 extra steps on tree Y, which is bad also).

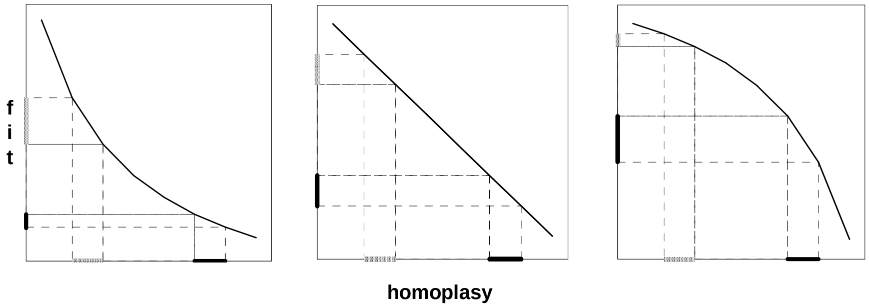


Fig. 3.2. Character fits as concave (left), linear (middle), or convex (right) decreasing functions of homoplasy (after Goloboff 1993a, fig. 1). Concave functions automatically resolve character conflict in favour of reliable characters, i.e. characters with less homoplasy. See text for explanation.

It can be argued that the same difference in homoplasy (1 step) is more important in c1 than it is in c2 because c1 is a character that seems to be a rather good indicator of the overall hierarchical structure of the data set, while c2 is not. Therefore tree X might be preferred over tree Y.

The conflict between characters c1 and c2 is resolved automatically in favor of c1 when the fit of a character is not defined as a linear, but as a concave decreasing function (fig. 3.2, left). Character c1 has a low amount of homoplasy on tree X (1 step), and adding one step of homoplasy to that character (as on tree Y) comes down to a relatively high decrease in the fit of the character. Conversely, character c2 has 14 extra steps on tree Y, and adding one more extra step, as on tree X, will decrease the fit of the character, which is low already, only slightly. The net result is that tree X has a higher total fit than tree Y.

As a third option, the fit could be defined as a convex function of the homoplasy (fig. 3.2, right), but this leads to the absurd situation that character conflicts are resolved in favor of the most homoplasious characters.

3.3.3 Maximizing total fit

In the above, the overall behaviour of concave, linear and convex decreasing functions, and as a result the preference for concave functions, depends crucially on the assumption that the best trees are not the shortest trees (taking into account the implied weights), but those that imply the highest sum of implied weights for all characters. This is an important assumption that is not properly discussed by Goloboff (1993a, 1995), as will be shown below (3.4). He does argue (Goloboff 1993a: 88) that a tree that maximizes the total weight of all characters is the best explanation of the data (Farris 1983) because such a tree treats the available data a priori as non-dismissable, and that therefore maximizing total fit is in line with cladistic philosophy. However, the observation that data should be treated a priori as non-dismissible relates to the question of how strong the weighting function should be and is not relevant to the question if maximizing total fit should be preferred over minimizing total weighted homoplasy.

3.3.4 Choice of a convex decreasing function

If it is agreed that the best trees are those that maximize the total fit, the weighting function should be a concave function of the homoplasy (3.3.2). The next question is what kind of decreasing function should be used (Goloboff 1993a: 89). Goloboff considered both the consistency index ci and the rescaled consistency index rc (the retention index ri is a linear decreasing function of the homoplasy), but he rejected them both because they may give different weights to characters that display the same amount of homoplasy when these characters have different values of m and/or g (see also 3.2.1).

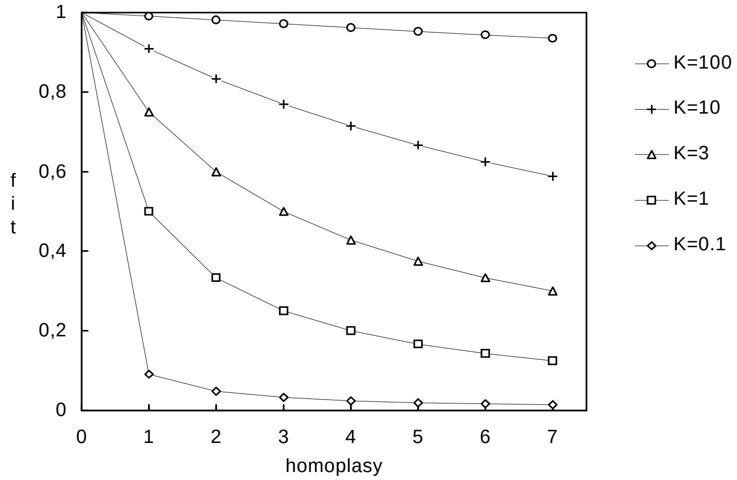


Fig. 3.3. The hyperbolic weighting function K/(K+homoplasy) for different values of K, the concavity constant. See text for explanation.

In order to avoid such differential weighting due to different values of m and/or g, Goloboff ultimately settled on the simple hyperbolic function K/(K+homoplasy), in which the concavity constant K determines the degree of concavity of the function (fig.

3.3; the notation is as in Goloboff 1993c, 1995; Goloboff (1993a) uses k = K-1 instead of K). That m does not appear in this function reflects the opinion that the reliability of a character with a certain number of extra steps is not influenced by the character being binary or multistate, with whatever number of character states. That g does not appear in this function reflects the opinion that the reliability of a character with a certain number of extra steps is not influenced by the total numbers of taxa that have the different states; e.g. a first extra step in a binary character that has two taxa with state zero and eighteen taxa with state one should have the same effect on the total fit as a first extra step in a binary character that has ten taxa with state zero and ten taxa with state one.

In fig. 3.3, the weighting function is shown for several values of K. Lower values of K lead to stronger differential downweighting (fig. 3.3) in the sense that (1) the same amount of homoplasy results in a larger decrease in fit, and (2) the total decrease in fit as a result of having any amount of homoplasy is put increasingly more into the first step of homoplasy that is observed. As K approaches zero, maximizing the total fit for a data set converges to compatibility analysis, and in the limit the method simply reduces to compatibility analysis: all characters with homoplasy receive zero weight, and all characters without homoplasy receive the same non-zero weight; as a result the fittest trees are those that accomodate the largest number of characters without homoplasy. Larger values of K decrease the degree of differential weighting. As K approaches ∞, the weighting function approaches a straight line, and as a result (see 3.3.2) maximizing the total fit converges to standard parsimony analysis. However, in the limit (K= ∞), the weighting function reduces to a vertical straight line, which is a non-decreasing function that assigns the same weight to each character, irrespective of its homoplasy. As a result, all trees fit equally well and the method does not reduce to standard parsimony analysis.

The behavior of the method under extreme values of K (K=0 or K→ ∞) can be described as “all or none” and “no-weighting”, and both are illogical extremes because they leave part of the information in the data unused (Goloboff 1995: 100): setting K to zero ignores that characters that have homoplasy may nonetheless still carry phylogenetic information, while using large values of K ignores that some characters may be more reliable than others. These extremes are to be avoided, but beyond this basic observation it remains to be established which values of K are optimal, or if data sets with different numbers of taxa require different values of K (Goloboff 1993a: 89). One could also argue that the reliability of a character depends upon other factors besides its homoplasy, such as the amount of polymorphism that is present in the character (e.g. Farris 1966). This obviously would lead to a modification of Goloboff’s fitting function, and Szumik (1996) effectively uses a function that takes into account both the amount of homoplasy and the amount of polymorphism to estimate the total fit of his data.

Besides hyperbolic functions, other concave decreasing functions might be considered. As an example, various evolutionary models (see 3.1) lead to the conclusion that the weight of a character behaves as a concave decreasing function that is obtained as the negative logarithm of simple functions of its transformational probabilities. From this point of view, the homoplasy of a character might be used to estimate these probabilities, and logarithmic fitting functions such as -ln((1+h)/C), with **C > g** acting as a concavity constant, might be considered. In fig. 3.4, hyperbolic and logarithmic fitting functions are shown for various values of the concavity constants C and K. The logarithmic weights are scaled such that no homoplasy gives weight one.

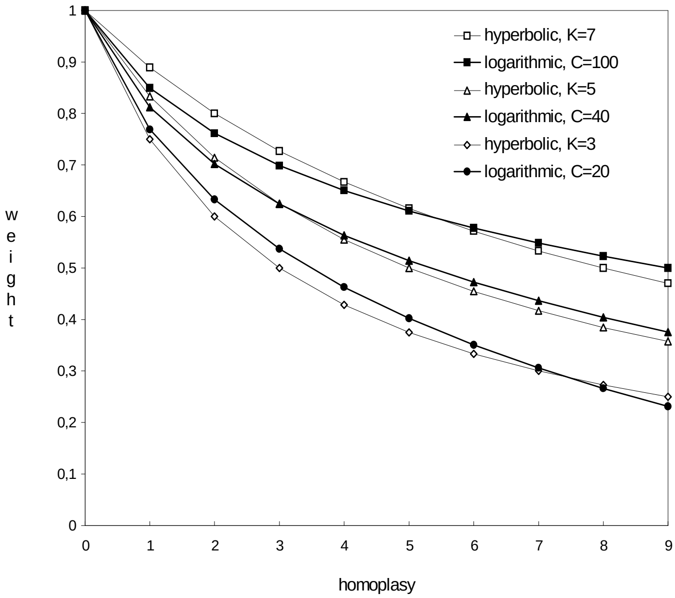


Fig. 3.4.Hyperbolic and logarithmic concave weighting functions. See text for explanation.

3.4 Using implied weights as weights during parsimony analysis

Goloboff (1993a: 88, 90) interpreted the trees with the highest total fit for a data set as the trees that provide the best explanation of that data set, and therefore he considered his approach to be in direct agreement with cladistic ideas. However, this view of what constitutes a best explanation does not agree with what is generally considered as a best explanation in cladistic philosophy (Farris 1983). A modification - or at least a re-interpretation - of Goloboff's approach is proposed.

3.4.1 Maximizing fit versus minimizing weighted homoplasy

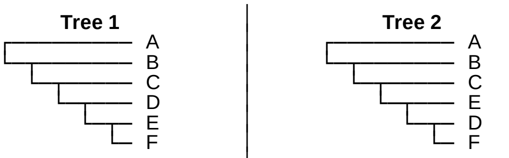
Fittest trees for a data set are trees which imply that the characters have on average as high a weight as possible, i.e. that the characters are maximally reliable. Therefore Goloboff considers them to be the best explanation of the data (1993a: 88, 90). However, in cladistic philosophy it is the most parsimonious trees that are considered to be the best explanation of the data (Farris 1983), and most parsimonious trees are those trees that imply the lowest amount of weighted homoplasy, as Goloboff (1995) himself discusses at length in his rebuttal of Turner & Zandee's (1995) criticism of implied weighting.

The weighted homoplasy of a character on a tree is traditionally calculated as the homoplasy of the character on the tree times its a priori weight (which may change during successive rounds in successive weighting). In the framework of implied weights, this product should in turn be multiplied by the character's implied weight (or fit, f) on the tree under consideration. Assuming equal a priori weights, the most parsimonious trees using implied weights are then those trees that minimize $\Sigma_{i=1..nchar}$ f_i*h_i over all characters of the data set, while the fittest trees are those that maximize $\Sigma_{i=1..nchar}$ f_i. Fittest trees and most parsimonious trees may coincide in particular cases or using particular weighting functions (see 3.4.3), but this is not generally true, as shown in the following example (fig. 3.5).

The first nine characters of the data set of fig. 3.5 specify the relationships between taxa A-F unambiguosly as (A(B(C(D E F)))). Characters c10 and c11 conflict with characters c2-c9 because they both group B together with E and F, but this conflict is resolved beyond doubt in favor of characters c2-c9. Furthermore, characters c10 and c11 are in conflict with character c12 because they resolve the relationships between D, E, and F differently, which results in the two cladograms shown in fig. 3.5.

Because characters c1-c9 are free of homoplasy on both trees, the differences in total fit or total length between both trees depend only on characters c10-c12. In the lower part of figure 3.5, the homoplasy and corresponding implied weights and weighted homoplasies are shown for several weighting functions. Using Goloboff's hyperbolic weighting function, the first tree is the fittest for moderate concavity (K=3 or higher), but stronger concavity (K=1 or lower) indicates the second tree as the fittest. The same holds for the weighted homoplasy using hyperbolic weights: the first tree is the shortest tree when K=3, but the second becomes shorter under stronger concavity (K=1).

```
            11 1
  123456789 01 2
A 000000000 00 0
B 100000000 11 0
C 111110000 00 0
D 111111111 00 1
E 111111111 11 0
F 111111111 11 1
```



		Tree 1 c10	Tree 1 c11	Tree 1 c12	Tree 1 sum	Tree 2 c10	Tree 2 c11	Tree 2 c12	Tree 2 sum
	unweighted homoplasy	1	1	1	**3**	2	2	0	4
K=3	hyperbolic fit,	3/4	3/4	3/4	**2.25**	3/5	3/5	1	2.2
	hyperbolic weighted h	3/4	3/4	3/4	**2.25**	6/5	6/5	0	2.4
K=1	hyperbolic fit	1/2	1/2	1/2	1.5	1/3	1/3	1	**1.67**
	hyperbolic weighted h	1/2	1/2	1/2	1.5	2/3	2/3	0	**1.33**
C=11	logarithmic fit	-ln(2/11)	-ln(2/11)	-ln(2/11)	**5.11**	-ln(3/11)	-ln(3/11)	-ln(1/11)	5.00
	logarithmic weighted h	-ln(2/11)	-ln(2/11)	-ln(2/11)	**5.11**	-2*ln(3/11)	-2*ln(3/11)	0	5.20
C=10	logarithmic fit	-ln(2/10)	-ln(2/10)	-ln(2/10)	**4.83**	-ln(3/10)	-ln(3/10)	-ln(1/10)	4.71
	logarithmic weighted h	-ln(2/10)	-ln(2/10)	-ln(2/10)	4.83	-2*ln(3/10)	-2*ln(3/10)	0	**4.82**

Fig. 3.5. Fittest trees are not necessarily trees with minimal weighted homoplasy. See text for explanation.

The first tree is preferred also by the logarithmic function, -ln((1+h)/C), under moderate concavity (C=11 or higher), both in terms of total fit and total weighted homoplasy. However, when the logarithmic concavity is increased (C=10 or smaller), the fittest tree and the tree with minimum weighted homoplasy are no longer the same: maximal reliability of characters leads to a preference for tree 1, while the weighted homoplasy is minimal on tree 2.

3.4.2 Justifying concave weighting functions

Goloboff's (1993a) rationale for preferring concave decreasing fitting functions and rejecting linear and convex decreasing functions relied on the behavior of such functions when the weight of characters is maximized: concave decreasing functions automatically resolve character conflicts in favor of characters with less homoplasy, while linear functions do not resolve such conflicts and convex functions resolve the conflicts in favor of characters with more homoplasy (3.3.2). However, this rationale does no longer hold when the weighted homoplasy is minimized.

Firstly, there exist concave decreasing functions that resolve character conflicts in favor of characters with more homoplasy. Assume first the concave fitting function 1/h (an example of an unbounded concave function; Farris 1969). In that case, Σf_i*h_i, reduces to $\Sigma 1$, and any tree wil have the same amount of weighted homoplasy: the number of characters in the data set. Next assume a fitting function that decreases more rapidly than 1/h. Now the weighted homoplasy f*h will decrease as h increases, and the net result of minimizing the weighted homoplasy is that character conflicts will be resolved in favor of characters with more homoplasy. An example is the function K/((K+h)*h). In that case minimizing Σf_i*h_i reduces to minimizing Σ K/(K+h), which is identical to searching the trees that imply that the characters are maximally unreliable in terms of Goloboff's approach. Of course, character conflicts can be resolved in favor of reliable characters when the quantity Σ K/(K+h) is maximized in stead of minimized, what is precisely what Goloboff (1993a) proposed, but in terms of weighted homoplasy this comes down to the counterintuitive notion of maximizing homoplasy. Changing the sign (i.e. minimizing -Σ K/(K+h)) does not solve this paradox because this would imply that the fitting function equals - Σ K/((K+h)*h), with the counterintuitive consequence that characters with less homoplasy receive lower implied weights.

Secondly, there exist linear and convex functions that will also resolve character conflicts in favor of characters with less homoplasy. First assume a constant fitting function, i.e. a function that assigns the same constant fit c to every character, irrespective of its homoplasy. In this case Σf_i*h_i reduces to $c*\Sigma h_i$, which comes down to minimizing the unweighted homoplasy. This the border case, and any fitting function that increases as h grows will resolve character conflicts in favor of unreliable characters, while decreasing fitting functions may do the opposite as long as they do not decrease too much in the relevant range of homoplasy.

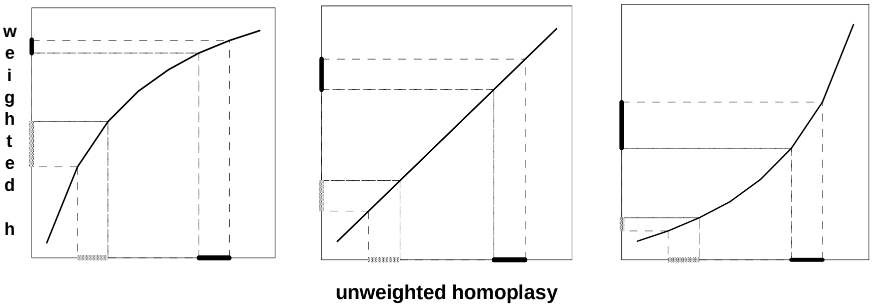


Fig. 3.6. Weighted homoplasy as increasing convex (left), linear (middle), or concave (right) functions. Only fitting functions that result in an increasing convex weighted homoplasy (left) automatically resolve character conflict in favour of reliable characters, i.e. characters with less homoplasy. See text for explanation.

An easy way to conceive of appropriate fitting functions is to focus on the behavior of the weighted homoplasy (fig. 3.6). Firstly, fitting functions that result in a weighted homoplasy that decreases as the homoplasy increases are excluded because they result in the counterintuitive notions of maximizing homoplasy or assigning lower weights to less homoplasious characters. Next, the same kind of reasoning that led to a preference of concave decreasing fitting functions when the total fit was maximized (fig. 3.2) can be used to evaluate the various types of increasing weighted homoplasy (fig. 3.6). Fitting functions that result in a convex increasing weighted homoplasy (fig. 3.6, left) will always resolve character conflict in favor of characters with less homoplasy, and it is of no importance if such fitting functions are themselves linear, convex, or concave; e.g. linear fitting functions a-b*h, with a and b positive and a/(2*b) =< -max$_{i=1..nchar}$(m$_i$-g$_i$)) will result in a convex parabolic increasing weighted homoplasy in the relevant range. Fitting functions that result in a linear increasing weighted homoplasy (fig. 3.6, middle) conform to equally weighted homoplasy, while a concave increasing homoplasy (fig. 3.6, right) will resolve character conflict consistently in favor of characters with higher homoplasy.

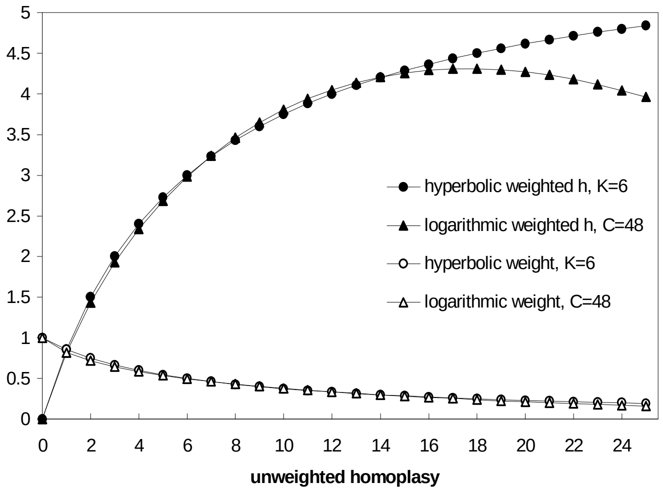


Fig. 3.7. The logarithmic weighted homoplasy becomes a decreasing function at a homoplasy level of approximately C/e (17.6 for C=48), while the hyperbolic weighted homoplasy increases asymptotically towards K. See text for explanation.

While minimizing weighted homoplasy is more in line with cladistic philosophy than maximizing mean character reliability, it puts the question of the appropriate kind of fitting function in a new perspective. A possible argument in favor of concave decreasing weighting functions resulting in convex increasing weighted homoplasy was already mentioned in 3.3.4.: various evolutionary models predict that the weight of a character behaves as a concave decreasing function that is obtained as the negative logarithm of simple functions of its transformational probabilities, given that these probabilities are not too high (e.g. Farris 1978, Felsenstein 1981). From this point of view, hyperbolic decreasing functions might be defended as approximations of such logarithmic functions. One difference between hyperbolic and logarithmic fitting functions is that the resulting weighted homoplasy eventually becomes a decreasing function under logarithmic weighting (at approximately C/e for sufficiently large C, which is obtained by setting the first derivative with respect to the unweighted homoplasy to zero), while it remains increasing towards the asymptotic value K under hyperbolic weighting (see fig. 3.7 for an example). This makes the hyperbolic function more robust against concavity constants that are set too strongly.

3.4.3 A special property of hyperbolic weighting functions

In the example of fig. 3.5, the fittest and the shortest tree are the same for a given concavity constant and using hyperbolic weights: tree 1 is both the shortest and the fittest tree for K=3, while tree 2 is both the shortest and the fittest for K=1. This result is easily generalized for K=1: in that case the fit is 1/(1+h), and the weighted homoplasy h/(1+h). The sum of both is 1, and as a result maximizing the fit and minimizing the weighted homoplasy amount to the same: fittest trees will also have the shortest weighted homoplasy. More generally the order that is imposed on all possible trees by their total fit is exactly the reverse of the order that is imposed by their total weighted homoplasy. This property also holds for other values of K (see appendix B for a general proof).

For logarithmic funcions, on the other hand, maximizing fit and minimizing weighted homoplasy do in general not amount to the same, not even in the range where the weighted homoplasy remains an increasing function, as is clear from the counterexample given in fig. 3.5 (for C=10: tree 1 is fitter than tree 2, but tree 2 is the shortest of both). It remains an open question if equivalence of maximizing fit and minimizing weighted homoplasy is desirable or necessary for a sound weighting function, or if the property holds for a still more limited range of C-values.

3.4.4 Multistate characters

Goloboff (1993a, 1993c) presented his approach of implied weights as a method to differentially weight complete characters, but as was the case for successive weighting (3.2.3) the approach may be extended to differential weighting of within-character state transformations. The same is true for minimization of weighted homoplasy.

3.4.4.1 Ordered characters

As noted by Mészáros et al. (1996; see chapter 5), ordered multistate characters should be coded as series of binary characters when applying the implied weighting approach as implemented in the computer program PeeWee (Goloboff 1993c). As in successive weighting (3.2.3.1), a systematic error arises otherwise: when an ordered multistate character is not decomposed into its binary constituents, its decrease in fit due to homoplasy will be estimated correctly only if exactly one of its state transformations is homoplasious, but it is underrated in any other case. The sensitivity to the way of coding is not an intrinsic property of implied weighting itself: in the computer program ViTA (Appendix A), it is implemented such that ordinal and additive binary coding give the same results.

3.4.4.2 Unordered characters

Implied weights can also be used to obtain within-character differential weighting of state transformations in unordered multistate characters, but some complications arise.

Consider the distribution of steps in an unordered multistate character with ns different states. There are ns*(ns-1)/2 different possible state transformations (the number of pairs of two different states; transformation costs are assumed symmetric and all equal). If the character is fully congruent with a tree, there will be exactly (ns-1) out of those ns*(ns-1)/2 possible transformations that require exactly one step on the tree, while the remaining (ns-1)*(ns-2)/2 do not occur at all. A priori, it is not known which (ns-1) transformations will occur, and different trees on which the character has no homoplasy may imply different sets of (ns-1) transformations that do occur. However, because there is no homoplasy, each of these (ns-1) transformations will always fit the tree perfectly. If in addition the transformations that do not occur are also considered to have a perfect fit, the total fit of the character is ns*(ns-1)/2.

Next assume a tree on which the character has one extra step. There are two possibilities. First consider the case where only (ns-1) different state transformations occur. This means that one of these (ns-1) transformations occurs twice, and therefore that single transformation could be called the homoplasious state transformation. With a hyperbolic fit function, its fit is K/(K+1), compared to 1 for the other transformations. Summed over all transformations, the character has a fit that is equal to (ns*(ns-1)/2-1) + K/(K+1). The other possibility for having one step of homoplasy is that ns different state transformations occur precisely once. In this case the question which one out of the ns occurring state transformations is the homoplasious transformation is difficult to answer, but this is no problem for calculating the total fit of the character: irrespective of the exact identity of the homoplasious transformation, the total fit will always be (ns*(ns-1)/2-1) + K/(K+1). When compared to the case without homoplasy, the fit decrease due to the first step of homoplasy is equal to 1-K/(K+1) = 1/(K+1) in both cases, which is the same result as for a binary character.

With two steps of homoplasy there are more possibilities to distribute the homoplasy over the character. First consider the case where only (ns-1) transformations occur. In that case there is either one transformation that occurs three times and (ns-2) transformations that occur once, or two transformations that occur twice and only (ns-3) that occur once. The transformations that occur more than once are homoplasious, and as a result the total fits are equal to (ns*(ns-1)/2-1)+ K/(K+2) in the first case and (ns*(ns-3)/2-2) + 2*K/(K+1) in the second. If ns different transformations occur, then there will be one transformation that occurs twice, and this transformation has a fit that is equal to K/(K+1). The other step of homoplasy is in any of the other (ns-1) remaining transformations, and therefore the fit of these remaining transformations is (ns-2) + K/(K+1). Adding the perfect fit of the transformations that do not occur, the grand total equals (ns*(ns-2)/2-2) + 2*K/(K+1). Finally, when (ns+1) different transformations occur, similar reasoning leads to a total fit of (ns*(ns-1)/2-2) + 2*K/(K+1) also. Summarizing, the total fit is equal to (ns*(ns-2)/2-1) + K/(K+2) when the two homoplasious steps are concentrated in one transformation, and equal to (ns*(ns-1)/2-2) + 2*K/(K+1) in all other cases. When these results are compared with the fit of a character with one step of homoplasy, the fit decrease that is due to the second step of homoplasy is equal to K/(K+1)-K/(K+2) = K/((K+1)*(K+2)) in the first case and 1+K/(K+1)-2*K/(K+1) = 1/(K+1) in the second. Since K/((K+1)*(K+2)) is lower than 1/(K+1) for any positive value of K, the first case will be preferred over the second whenever there is a choice: conflicts within the character will be resolved in favor of transformations with a higher reliability.

In fig. 3.8, these results are compared with a binary character and with an unordered multistate character according to Goloboff’s (1993a) original approach. The first step of homoplasy in each of these characters gives an identical fit decrease. The fit decreases due to the second step of homoplasy are still identical for the binary and the unordered character according to Goloboff (1993a), but in the new approach the decrease in the unordered character depends upon the distribution of the two homoplasious steps. If the two homoplasious steps are in the same transformation, the decrease due to the second step of homoplasy is as a second step in a binary character or in Goloboff's approach, but otherwise the second step has an identical fit decrease as the first. This means that two binary characters with each one step of homoplasy will have a combined fit that is identical to an unordered character that has two steps of homoplasy such that no transformation occurs three times.

		K=20	K=10	K=5	K=1
binary character (or unordered multistate character according to Goloboff 1993a)					
first step of homoplasy:	1/(K+1)	0.048	0.091	0.17	0.50
second step of homoplasy:	K*/((K+1)*(K+2))	0.043	0.076	0.12	0.17
unordered multistate character					
first step of homoplasy:	1/(K+1)	0.048	0.091	0.17	0.50
second step of homoplasy, added to the transformation that is homoplasious already:	K*/((K+1)*(K+2))	0.043	0.076	0.12	0.17
second step of homoplasy, added to a transformation that is free of homoplasy:	1/(K+1)	0.048	0.091	0.17	0.50

Fig. 3.8. Some fit decreases due to homoplasious steps in binary and unordered multistate characters. See text for explanation.

If the same logic is applied in the general case of any number of homoplasious steps, the fit of an unordered multistate character with ns states is equal to $\Sigma_{j=1..ns}\Sigma_{k=j+1..ns}(K/(K+s_{jk}-l_{jk}))$, and its weighted homoplasy $\Sigma_{j=1..ns}\Sigma_{k=j+1..ns}((s_{jk}-l_{jk})*K/(K+s_{jk}-l_{jk}))$, with s_{jk} the number of times the transformation between states j and k occurs, and l_{jk} equal to 1 for the first (ns-1) smallest $s_{jk}>0$ and equal to 0 for all other s_{jk}; for ns=2 the formulas reduce to those for binary characters. With a logarithmic function, the fit of the character is equal to $\Sigma_{j=1..ns}\Sigma_{k=j+1..ns}-\ln((1+s_{jk}-l_{jk})/C)$ and its weighted homoplasy $\Sigma_{j=1..ns}\Sigma_{k=j+1..ns}(-\ln((1+s_{jk}-l_{jk})/C)*(s_{jk}-l_{jk}))$. Directed versions (separate weighting of transformations $j\rightarrow k$ and $j\leftarrow k$) are easily obtained: e.g. $\Sigma_{j=1..ns}\Sigma_{k=1..ns,\ k\neq j}(K/(K+s_{jk}-l_{jk}))$ for the hyperbolic fit.

When all homoplasy is concentrated into a single transformation, the total fit decrease for any number of steps will be the same as for the same number of steps in Goloboff's approach or in a binary character. If the homoplasy is spread over several transformations, the total fit decrease will be higher, and the more so as the homoplasy is more evenly spread. In the worst case, n steps of homoplasy in one unordered character will have the same total fit decrease as the fit decrease due to n binary characters with each one step of homoplasy.

Algorithms that maximize a function as $\Sigma_{i=1..nchar}\Sigma_{j=1..ns(i)}\Sigma_{k=j+1..ns(i)}K/(K+s_{ijk}-l_{ijk})$ over all characters of a data set will be slow compared to other parsimony algorithms because for every character and every tree all s_{jk} have to be calculated for every possible assignment of states to the inner nodes of the tree (but branch and bound reasoning can be applied such that not all assignments must actually be calculated). Note that not only the shortest reconstructions of the states of the inner nodes must be evaluated: precisely because of the within-character differential weighting there may be reconstructions that require more steps but that are nevertheless fitter than shorter reconstructions.

From this basic approach to differential weighting, several variants can be constructed. E.g. first, second, and third codon positions of protein coding nucleotide sequences could be pooled in order to calculate mean implied weights for the transformations within each of these three classes of characters (in stead of calculating different weights for each character separately). Similar pooling could be done within characters, e.g. to obtain one weight for all transversions and a second weight for all transitions.

3.4.5 Some other possibilities

3.4.5.1 Minimizing weighted length

When character weights remain fixed during parsimony analysis, then the minimized weighted homoplasy and the minimized weighted length of a data set differ by the same constant on any tree (the weighted length exceeds the weighted homoplasy by $\Sigma_{i=1..nchar}$ w_i*m_i, with w_i the weight of the character and m_i its observed variation. As a result, minimizing weighted homoplasy and minimizing weighted length amount to the same. However, when the weights depend on the homoplasy, as is the case for implied weights, this simple relationship between total weighted length and total weighted homoplasy does no longer hold, and it might be argued that minimizing the weighted length is more in line with cladistic philosophy than minimizing the weighted homoplasy. Consequently, the best trees would be those that minimize $\Sigma_{i=1..nchar}\Sigma_{j=1..ns(i)}\Sigma_{k=j+1..ns(i)}(s_{ijk}*K/(K+s_{ijk}-l_{ijk}))$ over all characters of a data set (with l as defined in 3.4.4) in stead of $\Sigma_{i=1..nchar}\Sigma_{j=1..ns(i)}\Sigma_{k=j+1..ns(i)}((s_{ijk}-l_{ijk})*K/(K+s_{ijk}-l_{ijk}))$.

With similar logic as used in appendix B to prove that the tree orders according to increasing total fit and decreasing total weighted homoplasy are identical under hyperbolic weights (cf. 3.4.3), it can be shown that the order according to decreasing $\Sigma_{i=1..nchar}\Sigma_{j=1..ns(i)}\Sigma_{k=j+1..ns(i)}(s_{ijk}*K/(K+s_{ijk}-l_{ijk}))$ also results in that same order for $K > 1$. If the implied weight for multistate characters is calculated as in Goloboff (1993a), then the tree order according to decreasing weighted length will be the same once again as long as K is larger than the largest observed variation that is present in the data set. An advantage of this use of implied weights is that the total weighted length of a data set can be subdivided unequivocally over the various branches of the cladograms.

3.4.5.2 Maximizing weighted similarity

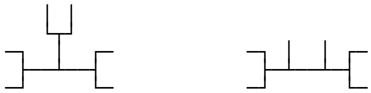
In the previous chapter, parsimony analysis was characterized as two-item analysis, i.e. as a method that identifies those trees that maximize the number of accomodated compatible independent pairwise similarities (2.6.4, p. 67). From this point of view, implied weights could be used to maximize the number of weighted accomodated compatible independent pairwise similarities. A binary character with nz 0-taxa and no 1-taxa has a priori (nz + no - 2) independent pairwise similarities; if there are n taxa and nm missing entries, this is equal to (n-nm-2). In general, a character with ns different states has $\Sigma_{i=1..ns}(n_i-1)$ = (n-nm-ns) independent pairwise similarities, with n_i the number of taxa that have state i. Each step of homoplasy refutes one of these similarities. Because the maximum amount of homoplasy, (g-m), puts an upper bound on the amount of independent similarities that can possibly be refuted, the maximization might be restricted to this number (the difference between maximizing all weighted similarities or only those that can be refuted is similar to the difference between minimizing weighted length and minimizing weighted homoplasy). Using a hyperbolic function and treating multistate characters as in Goloboff (1993a), the latter case yields the optimality function $\Sigma_{i=1..nchar}((g_i-m_i-h_i)*K/(K+h_i))$.

This function will resolve character conflicts not only in favor of characters with less homoplasy, but also in favor of characters with more informative variation (g-m). As a result, whenever two characters have the same amount of homoplasy but different amounts of informative variation (g-m), the character with the highest amount of informative variation will be considered more reliable than the other.

```
outgroup  0000000000  00000   0000000000
A         1111000000  11110   1111110000
B         1000111000  11101   1110001110
C         0100100110  11011   1001101101
D         0010010101  10111   0101011011
E         0001001011  01111   0010110111
          ╙────────────────╜  ╙──────────╜
                g-m=1             g-m=2
```

Fig. 3.9. An indecisive data set for five taxa + an outgroup.

This effect is illustrated by analyzing the indecisive data set that is shown in fig. 3.9. An indecisive data set is a data set that contains every possible informative character precisely once (or in an equal number), and as a result it has the same length on every possible resolved cladogram (Goloboff 1991a; see also chapter 6). In this case, all 105 different rooted cladograms have exactly 51 steps.



			TOPOLOGY 1	TOPOLOGY 2
h	g-m	g-m-h	number of characters	number of characters
0	1	1	3	2
0	2	2	0	1
1	1	0	12	13
1	2	1	6	5
2	2	0	4	4

Fig. 3.10. Distribution of homoplasies of the indecisive data set of fig. 3.9 on the two possible topologies for six taxa.

Because every informative character occurs precisely once in the data set of fig. 3.9, any two trees with the same topology (i.e. with the same connections between the branches, irrespective of the identity of the terminal taxa) will have the same distribution of the homoplasy over the characters with different (g-m) values. For six taxa (A-E + the outgroup) there are two topologies, and the corresponding distributions are given in fig. 3.10.

Based on these distributions, the total weighted number of accomodated pairwise independent similarities is easily calculated. As an example, for K=1 the weighted similarity for the trees with topology one is 12/2, and for the trees with topology two 13/2. Therefore, it is concluded that trees with the second topology are better explanations of the indecisive data set than trees with topology one. The reason is clear from the homoplasy distributions: trees of the second topology have less homoplasy in characters with high (g-m) than trees of the first topology.

3.4.6 ViTA: a computer program for parsimony analysis using implied weights

Two of the hyperbolic optimality functions that are discussed above are available in ViTA, a DOS computer program: minimization of total weighted homoplasy and maximization of total weighted pairwise independent similarity. In order to allow direct comparison with existing methods, the program also provides minimization of total unweighted homoplasy (standard parsimony analysis) and maximization of total fit (Goloboff 1993a). A full discussion of the program is presented in appendix A. A number of practical features, such as calculation of consensus trees or optimization of polytomies, are not available in the program, but it can write trees in a Hennig86 compatible format (Farris 1988). As a result, trees from ViTA can be easily imported in any program that can read Hennig86 trees.

As discussed in 3.4.3 and proved in appendix B, minimizing total weighted homoplasy and maximizing total fit are equivalent under hyperbolic weighting.

Nevertheless, they are retained as separate types of analysis in the program because they differ in interpretation and because the equivalence is not general (e.g. it does not hold for logarithmic functions). The use of implied weights to calculate weighted homoplasies is called **direct weighting**.

The calculation of implied weights and total values differs in the following points from the implementation that is available in Goloboff's (1993c) computer program PeeWee:

1. ViTA rounds the implied weight to two decimals; PeeWee truncates to two decimals and multiplies that result by ten (see appendix A.7).
2. ViTA rounds only after all character fits (or weighted homoplasies) have been summed; PeeWee truncates the character fits before summing (see appendix A.7).
3. ViTA assigns perfect fit to autapomorphies; PeeWee completely excludes autapomorphies from the analysis (see appendix A.8).
4. ViTA automatically decomposes linearly ordered multistate characters into their binary constituents (see 3.4.4.1); this feature is not available in PeeWee.

Some other noteworthy points are the following: ViTA treats unordered multistate characters as in Goloboff (1993a), and not as proposed in 3.4.4.2; contrary to the situation in PeeWee, ViTA does not allow polymorphisms in terminal taxa (when polymorphisms are present in a data set, they are converted automatically into missing entries). Finally, PeeWee is many times faster than ViTA.

In ViTA the maximization of the weighted similarity $\Sigma_{i=1..nchar}((g_i\text{-}m_i\text{-}h_i)*K/(K+h_i))$ is presented as a minimization. The weighted similarity of a character, (g-m-h)*K/(K+h), reaches its maximum value (g-m) when the character has no homoplasy, and as a result the quantity $\Sigma_{i=1..nchar}((g_i\text{-}m_i)\text{-}(g_i\text{-}m_i\text{-}h_i)*K/(K+h_i)) = \Sigma_{i=1..nchar}((K+g_i\text{-}m_i)*h_i/(K+h_i))$ behaves as a weighted homoplasy: it is 0 in the absence of homoplasy and increases with increasing homoplasy. This use of implied weights is called **complex weighting**. The total weighted similarity is equal to (G-M - the total complex weighted homoplasy).

This is illustrated by using the morphological data set of Gentianaceae that is discussed in chapter 5 (table 5.3; Mészáros et al. 1996). The trees that are presented in chapter 5 are the standard shortest trees and the fittest trees under various concavity constants, ordering assumptions, and a priori weights. As discussed in 3.4.3, the fittest trees are also the trees with minimal weighted homoplasy under direct weighting. These trees were calculated using the programs NONA and Pee-Wee (Goloboff 1993b, 1993c), which both accept polymorphisms in terminal taxa. Because ViTA does not accept such polymorphisms, all polymorphisms in the data set of table

5.3 were changed into missing entries for the calculation of the trees with minimal weighted homoplasy under complex weighting (the resulting data set has M=57 and G=204). In order to allow a direct comparison of these trees with shortest and fittest trees, the same modified data set was also used to recalculate the shortest trees and the fittest trees. In each case, all characters were treated as unordered and had equal a priori weights of 1.

The most parsimonious trees were obtained with NONA (Goloboff 1993b), the fittest trees with Pee-Wee (Goloboff 1993c). In both cases, MULT*50 was used to search for the best trees. This instruction carries out 50 replications of randomizing the taxa, creating a tree by stepwise addition and submitting it to branch-swapping by means of tree-bissection reconnection. The trees with the shortest weighted homoplasy were obtained with ViTA (appendix A), using the instruction MULT100 (100 replications of randomizing the taxa, creating a tree and submitting it to branch-swapping by means of subtree pruning and regrafting; see appendix A for details). Apart from the value of the concavity constant (in Pee-Wee and ViTA) and the unordering of multistate characters (in NONA and Pee-Wee) all default settings were retained.

When the polymorphisms in the data set of table 5.3 are changed into missing entries, there are six most parsimonious trees, with length 109 (two steps less than for

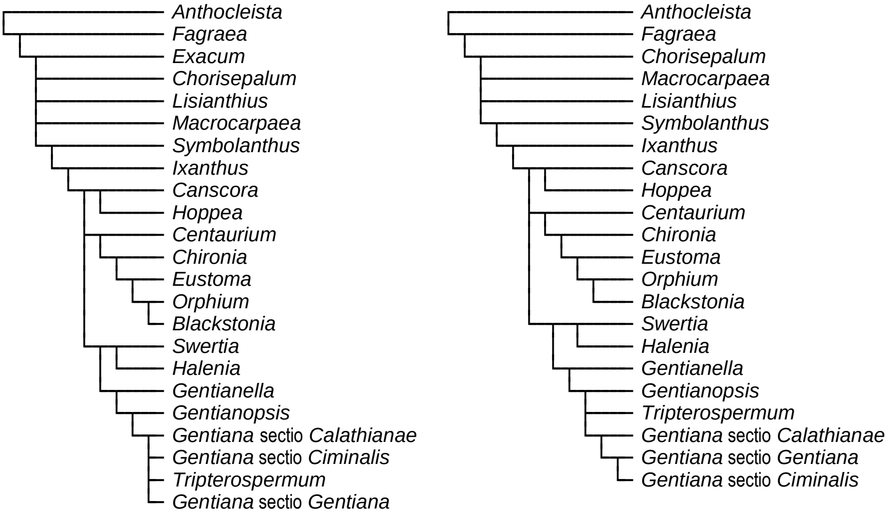


Fig. 3.11. Cladistic analysis of the data set of table 5.3 (all characters unordered, polymorphisms changed into missing entries). Left: strict consensus tree of the six most parsimonious trees (length 109, CI 0.523, RI 0.646). Right tree: strict consensus (with exclusion of *Exacum*) of the seven different fittest trees that are obtained by using concavity constants 1-6. See text for explanation.

the unmodified data, cf. fig. 5.1). The strict consensus of these (fig. 3.11, left) is identical to the strict consensus of the most parsimonious trees for the original data (fig. 5.1) but for one small detail: the sister group relation between *Chorisepalum* and *Lisianthius* is lost. Substituting the polymorphisms for missing entries has not much influence on the fittest trees neither. When the concavity constant is varied between 1 and 6, a total of 7 different fittest trees are found. The strict consensus of these, but with *Exacum* excluded, is shown in fig. 3.11 (right; when *Exacum* is included in the calculation of the strict consensus, *Symbolanthus*, *Ixanthus* and *Exacum* join the polytomy at the base of that tree; a variable position of *Exacum* was also found with the original data). This tree is highly congruent with the unweighted results and with the results obtained in chapter 5 (5.3, e.g. fig. 5.2).

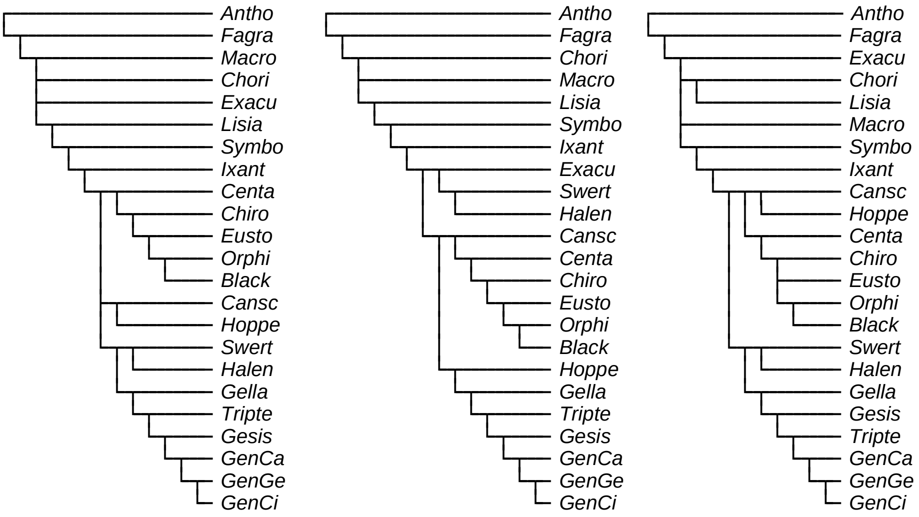


Fig. 3.12. Cladistic analysis of the data set of table 5.3 (all characters unordered, polymorphisms changed into missing entries). Trees with smallest weighted homoplasy, complex weighting. Left: K=1, single best tree with weighted homoplasy = 93.24; Middle: K=6, single best tree with weighted homoplasy = 70.28. Right: K=12, strict consensus of three best trees with weighted homoplasy 62.97. *Gella* and *Gesis* are *Gentianella* and *Gentianopsis*; *GenCa*, *GenCi* and *GenGe* are the sections *Calathianae*, *Ciminalis* and *Gentiana* of the genus *Gentiana*; the other taxa are indicated by their first five characters (see table 5.1 or fig. 3.11 for full names).

Minimization of weighted homoplasy under complex weighting seems to be more sensitive to the exact value of the concavity constant than maximization of fit. In fig. 3.12, the results are shown for three values of K (only the best trees; suboptimal trees were not considered). For K=1, the single best tree (fig. 3.12, left) has a weighted homoplasy equal to 93.24. For K=6, the single best tree (fig. 3.12, middle) has a weighted homoplasy equal to 70.28. Finally, for K=12 there are three best trees, with weighted homoplasy equal to 62.97; the strict consensus of these is shown in fig. 3.12 (right).

A first remark concerns the degree of differential weighting: with higher values of K, this degree decreases, and as K keeps growing, the approach converges slowly towards an unweighted analysis; this can be seen from the weighted homoplasies for the different K-values: 93.24, 70.28, and 62.97 for K equal to 1, 6, and 12 respectively. For still higher values of K, the weighted homoplasy will get closer and closer to the unweighted homoplasy, which is 109-57=52 (the direct weighted homoplasy converges in a similar way towards the unweighted value as K increases, but the deviation is in the other direction: the weighted homoplasy is always less than the unweighted homoplasy).

The trees with the highest K-value (K=12, fig. 3.12, right) are the most congruent with the unweighted and the fittest trees. Better congruence with the unweighted trees could be expected precisely because higher values of K have less differential weighting. The surprising thing is that the lowest tested value of K (K=1) gives a similar tree, while intermediate weighting (K=6) gives a tree that deviates in some conspicuous points from the other trees (an unusual position of the sister pair *Swertia-Halenia*, and a loss of the sister relationship between the genera *Canscora* and *Hoppea*; cf. chapter 5). Various explanations are possible. As an example, calculating the weighted homoplasy with a precision of two decimals might be too precise, and suboptimal trees should be taken into account also. In this case, the above results might simply point to ambiguity in the data. On the other hand, the low and high values of K might be too extreme; after all, as noted by Goloboff (1995: 100), both too strong and too weak forms of character weighting are difficult to defend. Still an other explanation is that complex weighting might give erratic results because it is too sensitive to differences in informative variation between characters. Such questions are difficult to answer on the basis of this small example, and they are left open for future research.

3.5 Summary and conclusions

Based on the concepts of hierarchical correlation and cladistic reliability of characters, Farris (1969) proposed a successive approximations approach to character weighting in which the characters are differentially weighted according to their homoplasy. He preferred concave decreasing functions of the homoplasy, such as the consistency index, over other types of possible weighting functions because these functions performed best in simulation studies. Later, Williams & Fitch (1989, 1990) extended the approach to within-character differential weighting of state transformations in nucleotide sequence data. The theoretical basis of successive weighting and some of the difficulties of the approach are shortly reviewed, and it is discussed how it might be applied to obtain within-character differential weighting in any type of unordered character.

Some years ago, Goloboff (1993a) proposed an alternative homoplasy-based weighting method that is non-iterative and avoids the main problems of successive weighting. In this approach, the fit or the implied weight of a character is defined as a hyperbolic decreasing function of its homoplasy, and the best trees are those that have the highest total fit over all characters of a data set. Concave decreasing weighting functions arise in a natural way in this method because they are the only functions that resolve character conflicts consistently in favor of reliable characters. Goloboff (1993a) considered his approach to be in direct agreement with cladistic ideas, but most parsimonious trees are those trees that imply the lowest weighted homoplasy (Farris 1983), and these are not necessarily the trees that imply that the characters have the highest total fit, as is shown.

When the issue of using implied weights is considered from this point of view - minimization of weighted homoplasy in stead of maximization of character weight - then concave decreasing functions are no longer the only weighting functions that resolve character conflicts in favor of reliable characters, and there even exist concave decreasing functions that resolve character conflicts in favor of unreliable characters. Nevertheless, concave weighting functions might still be preferred because various evolutionary models predict that the weight of a character behaves as a concave decreasing function that is obtained as the negative logarithm of simple functions of its transformational probabilities (given that these probabilities are not too high; e.g. Farris 1978, Felsenstein 1981). From this point of view, hyperbolic decreasing functions might be seen as approximations of such logarithmic functions. It is shown that maximizing total fit and minimizing total weighted homoplasy are equivalent when using hyperbolic decreasing functions, as in Goloboff’s approach. Logarithmic weighting functions, on the other hand, do not have this property.

It is discussed how implied weights may be used to obtain within-character differential weighting of state transformations in both ordered and unordered multistate characters. For ordered characters, the approach comes down to decomposing the characters into their binary constituents; the unordered case is a rather straightforward extension of the binary case.

In standard parsimony analysis, the shortest trees are also the trees that have the lowest homoplasy, and the trees that retain the maximal amount of independent pairwise similarities. However, when using implied weights, these various optimality functions are in general no longer equivalent. The problem is described and some alternative possibilities of using implied weights are proposed. One of these, complex weighting, is sensitive to the amount of informative variation of characters: for equal amounts of homoplasy, a character with more informative variation is estimated more reliable than a character with less informative variation. As a first example of complex weighting, it is shown that indecisive data sets (Goloboff 1991a, see also chapter 6) loose their indecisive nature when this kind of weighting is applied. The chapter is concluded with an analysis of a data set for Gentianaceae (see chapter 5) under complex weighting. This analysis is performed with the computer program ViTA, a newly developed computer program for parsimony analysis using implied weights. For intermediate values of the concavity constant K, the single best tree deviates in some conspicuous points from the shortest unweighted trees and the shortest direct weighted trees. Some explanations for this behaviour are suggested.

4. A COMMENTARY ON THE CIRCUMSCRIPTION AND EVOLUTION OF THE ORDER GENTIANALES, WITH SPECIAL EMPHASIS ON THE POSITION OF THE RUBIACEAE[6]

Recently the inclusion of the Rubiaceae in a monophyletic Gentianalean group has been confirmed in several cladistic analyses of macromolecular as well as morphological data. These developments are discussed against a historical background and some comments are given on the evolution of our understanding of the position of the Rubiaceae within the angiosperms.

4.1 Introduction

This conference[7] is focussing mainly on the intrafamilial relationships of the Rubiaceae. Within this enormous family, many genera need to be revised and many complex subfamilial and tribal classification problems remain unsettled. In comparison with the other large angiosperm families, the Rubiaceae still remain undertreated (Robbrecht 1993a), but the success of the first international Rubiaceae conference (Taylor 1995) and the size of this volume testify that the interest in Rubiaceae systematics is growing. Nevertheless, in this era of cladistics and macromolecular systematics, in which none of the subclasses of the angiosperms as defined by Cronquist (1981, 1988) are save from drastic changes, it would be a missed opportunity not to discuss the higher levels of the classification as well. Moreover, a quick look at the abstract book of this conference reveals several interesting contributions about the possible relatives of the Rubiaceae, and about taxa with a questionable taxonomical position within or near the Rubiaceae (cf. Robbrecht 1993b). With this in mind, it seems appropriate to present some comments on the history and recent developments of our understanding of the evolutionary position of the Rubiaceae within the angiosperms.

4.2 Historical background

During the previous century, hypotheses concerning the relationships of Rubiaceae stressed the strong affinity of this family with Caprifoliaceae, and hence

[6] Reprinted from De Laet J. & E. Smets (1996).
[7]Second International Rubiaceae Conference. Meise, National botanic Garden of Belgium.

with Adoxaceae, Dipsacaceae and Valerianaceae (Wagenitz 1959). Baillon (1880) even included Caprifoliaceae and Adoxaceae in the Rubiaceae. Affinities with Loganiaceae (cf. Wagenitz 1959: 31), Gentianaceae and Apocynaceae (e.g. Le Maout & Decaisne 1868: 159) were acknowledged, but in terms of formal classification these were mostly thought to be less important than the affinities with Caprifoliaceae. By the turn of the century, this general consensus was reflected in Engler's (1897a) Reihe Rubiales, containing Rubiaceae, Caprifoliaceae, Adoxaceae, Valerianaceae and Dipsacaceae (fig. 4.1).

***Reihe* Contortae**
- ***Unterreihe* Oleineae**
  - Oleaceae
  - Salvadoraceae
- ***Unterreihe* Gentianineae**
  - Loganiaceae
  - Gentianaceae (incl. Menyanthaceae)
  - Apocynaceae
  - Asclepiadaceae

***Reihe* Rubiales**
- Rubiaceae
- Caprifoliaceae
- Adoxaceae
- Valerianaceae
- Dipsacaceae

Fig. 4.1. The position of the Rubiaceae in Engler’s system of 1897b (see text for discussion).

The main issue then was not the strong relationship between Rubiaceae and Caprifoliaceae, which was almost undisputed, but the wider relationships of the order Rubiales within the angiosperms. Engler's Reihen Rubiales and Contortae, including the Unterreihe Gentianineae (fig. 4.1) are not next to each other in the linear sequence of his classification, but nevertheless he did consider the possibility that Gentianineae and Rubiales were closely related, as is clear from his own comments (Engler 1897c: 370): "*Es dürften somit die Loganiaceae einen älteren Typus repräsentieren, von dem sich die übrigen Familien der Gentianineae und vielleicht auch die Rubiales abgezweigt haben*". In current terminology Engler's statement implies that he considered the possibility that Loganiaceae as well as Gentianineae were paraphyletic, while the group that consists of Gentianineae + Rubiales might be monophyletic. Bessey (1915: 116-118) had a different opinion. His orders Gentianales and Rubiales contain the same families as Engler's Gentianineae and Rubiales, but he did not consider them to be closely related at all: his Rubiales and Gentianales are representatives of two very distict phyletic sequences, both with an origin in the Ranalean complex.

When Wagenitz (1959) was preparing the treatment of Gentianales and Rubiales for a new edition of Engler's syllabus (Wagenitz 1964), he was struck by the fact that, contrary to common belief, the Rubiaceae were much closer to Engler's Gentianineae than to the other families of the Rubiales. His extensive comparison of existing literature included morphological, anatomical, embryological as well as chemical evidence. The strong affinity between Rubiaceae and Caprifoliaceae, taken for granted for so long, was apparently based on two characteristics only: the presence of an the inferior ovary and the absence of intraxylary phloem. It was probably the inferior ovary that made Engler conclude that Rubiales had reached a higher level of development than the Gentianineae, which in turn justified the recognition of a separate *Reihe* in his approach to classification (Engler 1897a; cf. also Barabé & Vieth 1990). Looking back, the overstatement of the progression to an inferior ovary may have hindered earlier recognition of the importance of the relationship between Loganiaceae and Rubiaceae.

Reihe Gentianales
Loganiaceae
Rubiaceae
Gentianaceae (incl. Menyanthaceae)
Apocynaceae
Asclepiadaceae

Reihe Dipsacales
Caprifoliaceae
Adoxaceae
Valerianaceae
Dipsacaceae

Fig. 4.2. Wagenitz' (1959) orders Gentianales and Dipscales (see text for discussion)

Wagenitz (1959) proposed new circumscriptions for both orders (fig. 4.2). Compared to Engler's (1897b) classification (fig. 4.1), the major changes are the exclusion of the Oleineae from Gentianales and the transfer of the Rubiaceae to Gentianales. For the remaining families of Engler's Rubiales, Wagenitz proposed the name Dipsacales. He considered it highly improbable that Dipsacales were a derived group within Gentianales. From his discussion it is clear that he interpreted both orders to be monophyletic. He put forward several conjectures concerning the relationships of these orders, but seemed to be inclined to the idea that they had separate origins from within Chorisepalous orders.

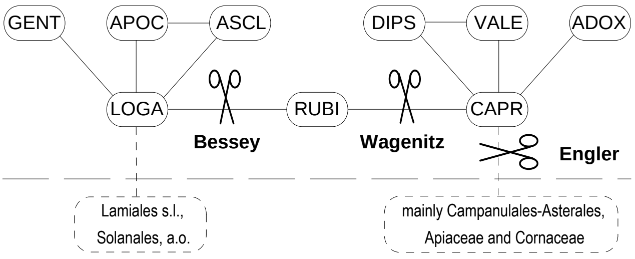


Fig. 4.3. Simplified representation of the relationships around the Rubiaceae as perceived by the end of previous century (complete family names are given in fig. 4.2). When it came to phylogenetic interpretation, Engler (1897c: 370), Bessey (1915), and Wagenitz (1959) cut ties differently. See text for discussion.

In comparison with Bessey's (1915) ideas, Wagenitz' treatment is a contribution toward a better understanding of the systematic position of the Rubiaceae, as the title of his paper implied. On the other hand, when confronted with Engler's conception of a monophyletic group consisting of Gentianineae + Rubiales, he actually raised the problem of the relationships of the Dipsacales (fig. 4.3). Indeed, both Bessey and Engler accepted the close phylogenetic relationship between Rubiaceae and Caprifoliaceae, but they differed in their opinion "which way up" evolution is (Stevens 1986).

In 1964, Wagenitz changed his Gentianales in two ways. Firstly, *Desfontainia*, a genus he had excluded from Loganiaceae and Gentianales in 1959, was now included in Gentianales as a monogeneric family. Secondly, the subfamily Menyanthoideae of Gentianaceae was raised to family level because it proved to be very different from subfamily Gentianoideae anatomically, embryologically as well as phytochemically. The circumscription of Dipsacales remained unchanged, and most recent systems of angiosperm classification (e.g. Takhtajan 1980, Cronquist 1981, 1988, Thorne 1992a, 1992b) agree with the groups included (even though there is no unanimity about the familial delimitations within the order). The only notable exception is Dahlgren (1983a; see also Dahlgren 1989), who transferred Caprifoliaceae, Viburnaceae and Adoxaceae to Cornales and included Calyceraceae in Dipsacales.

There has been less agreement concerning the delimitation of the Gentianales. The main points of discussion are:

1. Several groups with a long history of doubtful position within or near Loganiaceae, particularly *Buddleja* and related genera, and the genera *Desfontainia* and *Retzia* (see Bremer et al. 1994). Leeuwenberg & Leenhouts (1980) included these as the tribes Buddlejeae, Desfontainieae and Retzieae in their broadly circumscribed Loganiaceae. Cronquist recognized *Retzia* as a monospecific family in Gentianales (Cronquist 1981) and later Solanales (Cronquist 1988), while Dahlgren (1983a; see also Dahlgren 1989) and Thorne (1992a, 1992b) stressed the relationship with the Stilbaceae; *Desfontainia* has been recognized as part of Loganiaceae (Cronquist 1981, 1988, Takhtajan 1980), or as a separate family in Gentianales (Dahlgren 1983a, see also Dahlgren 1989) or Hydrangeales (Thorne 1992 a, b). Buddlejaceae are unanimously excluded from Gentianales (Takhtajan 1980, Cronquist 1981, 1988, Dahlgren 1983a, Dahlgren 1989, Thorne 1992a, 1992b) and included in a variously circumscribed Scrophulariales (Bignoniales sensu Thorne 1992b; Lamiales sensu Dahlgren 1989).
2. The species *Dialypetalanthus fuscescens*, that was originally described and put in a new monospecific tribe Dialypetalantheae within the Rubiaceae by Kuhlmann

(1925). Rizzini & Occhioni (1949) stressed the similarities with Myrtaceae and Melastomataceae. They raised *Dialypetalanthus* to family level and included it in Myrtales (see Piesschaert et al. 1997 for a more detailed account). Cronquist (1981, 1988), however, did not agree with a position in Rubiales or Myrtales and by lack of a better solution he included the family in his Rosales. Takhtajan (1980), Dahlgren (1983a; not in his earlier systems; cf. also Dahlgren 1989) and Thorne (1992b) included the family in their Gentianales.

3. Menyanthaceae. This family is still included in Gentianales by Takhtajan (1980) and Dahlgren (1983a), but transferred to Solanales by Cronquist (1981, 1988), to Campanulales by Thorne (1992a, 1992b), and to Cornales by Dahlgren (1989).
4. Rubiaceae. Most authors agree with Wagenitz' inclusion of this family in Gentianales (Takhtajan 1980, Dahlgren 1983a, Dahlgren 1989, Thorne 1992a, 1992b); only Cronquist (1981, 1988) maintains an order Rubiales.

Apart from the problematic delimitation of the order, there are also different opinions concerning the familial delimitations within the order. An example is the species *Saccifolium bandeirae*, that was recently described by Maguire & Pires (1978). Takhtajan (1980) included it with some doubts in the Gentianaceae, while Cronquist (1981, 1988), Dahlgren (1983a; see also Dahlgren 1989) and Thorne (1992a, 1992b) recognize it as a monospecific family in their Gentianales.As for the Rubiaceae, familial delimitation problems are limited to the genus *Theligonum* (*Henriquezia* is included in the Rubiaceae in all systems mentionned). Thorne (1992a, 1992b) is the only one to include *Theligonum* in the Rubiaceae. Takhtajan (1980) and Dahlgren (1983a; see also Dahlgren 1989) recognized a monogeneric family Theligonaceae in the Gentianales, while Cronquist (1981, 1988) did the same in his Rubiales.

As stated above, Wagenitz' (1959) discussion of Gentianales implied monophyly of the order Gentianales. More recently he explicitly put forward the hypothesis that Gentianales, a group that is "tied together by a combination of vegetative, floral and phytochemical characters", may indeed be one of the larger monophyletic groups within the Asteridae (Wagenitz 1992: 210; the loganiaceous tribes Buddlejeae and Retzieae sensu Leeuwenberg & Leenhouts (1980) are excluded; the position of Menyanthaceae is called controversial). It is tempting to evaluate the characters of the order and its different delimitations as given by Takhtajan, Cronquist, Dahlgren or Thorne against this explicit hypothesis. These classifications are indeed often used as if they were cladistic, i.e. "as if the circumscription and the rank of taxa depended on the relative cladistic branching positions" (Stevens 1986: 330). It may not be overlooked, however, that neither the rank of taxa, which is mainly determined by phenetic distance, nor the "box-in-box structure" of these classifications are very meaningful phylogenetic components (Kubitzki 1977: 25). Rank and hierarchy are not intended to reflect cladistic relationships in these systems.

The same is true for the diagrams that are used by Takhtajan, Cronquist, Dahlgren and Thorne to illustrate their systems. Takhtajan (1980: 348) and Cronquist (1981: 853, 1988: 414) use treelike diagrams to depict the "putative relationships" among their orders. Even though the treelike appearance of these diagrams suggests cladistic branching patterns, they basically depict current relationships in terms of relative advancement and are interpreted wrongly when they are thought to express hypotheses of mono- or paraphyly of orders (cf. Heywood 1977: 6). Likewise, Cronquist's (1975: 520, 1988: 439, 445) comments on the origins of Rubiales, Dipsacales, Calycerales and Asterales and their relationship with Gentianales, pointing to close evolutionary relationships between these orders (contrary to Engler, Bessey as well as Wagenitz), are suggestive of a paraphyletic Gentianales, but they do not exclude the monophyly of this order. Dahlgren (1980: 107-109) and Thorne (1992a: 367-369), on the other hand, make use of a cross section through an imaginary evolutionary tree or hedge (Dahlgren) or phyletic shrub (Thorne). The positions of the orders and superorders in the plane of section imply nothing definite about the exact branching pattern below this plane. Hence the question of mono- or paraphyly of the orders is left open.

The differences between evolutionary classifications are often due to different interpretations of the evolutionary significance of characters (Stevens 1986: 327, Barabé 1984; for example, embryological and chemical characters are emphasized much more in Dahlgren's classification than in the system of Cronquist). Wagenitz (1977: 390) rightly pointed out that this was one of the major obstacles to further progress: "we simply often do not know which characters we can rely on as indicating phyletic affinity or only a certain level of evolution". Since then, cladistic analysis has become a major research tool to distinguish between homologous and homoplasious similarity. Besides this methodological advance, macromolecular research has recently started to provide a whole new class of data (cf. Zurawski & Clegg 1993).

4.3 Current developments

Recently the inclusion of the Rubiaceae in a monophyletic Gentianales has been confirmed in several cladistic analyses of macromolecular as well as morphological and phytochemical data. With the exception of Downie & Palmer's (1992) study of restriction site variation of the chloroplast DNA inverted repeat, all the molecular analyses that are relevant to Gentianales are based on the same set of data, viz. the sequence of the *rbc*L gene (e.g. Chase et al. 1993, Olmstead et al. 1992, 1993, Bremer et al. 1994, 1995). They differ in the choice and the number of species included, and in the sophistication of the parsimony analysis (cf. Soltis et al. 1993).

With respect to Gentianales, one of the most interesting is Olmstead et al.'s (1993) thorough parsimony analysis of the Asteridae (fig. 4.4): it combines a relatively high number of species of Gentianales, spread over the order, with a wide range of possible relatives, including a.o. representatives of Dipsacales, Oleales, Cornales, Campanulales and Asterales. Other *rbc*L-based studies do not contradict Olmstead et al.'s main conclusions about Gentianales, which are always recognized as a monophyletic group when Menyanthaceae and part of Loganiaceae sensu Leeuwenberg & Leenhouts (1980) are excluded. Menyanthaceae (*Villarsia* and *Menyanthes*) are part of a Campanulales-Asterales clade (see also Chase et al. 1993, Olmstead et al. 1992, Cosner et al. 1994). This is not expected on the basis of gross floral and vegetative morphology of the family (Cronquist 1981), but restriction site variation of the chloroplast genome (Downie & Palmer 1992) and the presence of several primary and secondary metabolites (Lammers 1992) confirm this result. *Buddleja*, *Nicodemia* (Buddlejaceae), *Retzia* and *Desfontainia* (see Bremer et al. 1994 for *Retzia* and *Desfontainia*) are excluded from Loganiaceae sensu Leeuwenberg & Leenhouts (1980) and Gentianales (*Buddleja*, *Nicodemia* and *Retzia* are allied with Lamiales s.l., while *Desfontainia* is part of Dipsacales). The remaining genera of Loganiaceae that have been sequenced belong to the Gentianales, but they do not form a monophyletic group within the order. This had already been suggested on the basis of restriction site variation of the chloroplast genome (Downie & Palmer 1992) and on the basis of morphological and phytochemical evidence (Bremer & Struwe 1992).

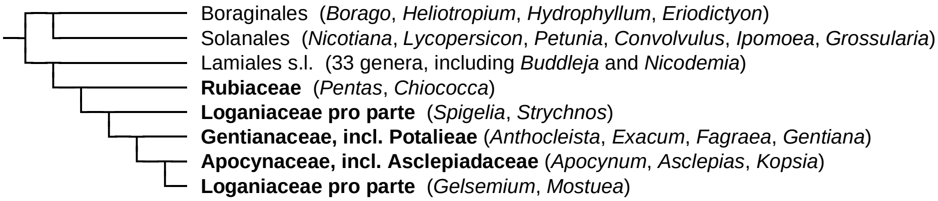

Boraginales (*Borago*, *Heliotropium*, *Hydrophyllum*, *Eriodictyon*)
Solanales (*Nicotiana*, *Lycopersicon*, *Petunia*, *Convolvulus*, *Ipomoea*, *Grossularia*)
Lamiales s.l. (33 genera, including *Buddleja* and *Nicodemia*)
Rubiaceae (*Pentas*, *Chiococca*)
Loganiaceae pro parte (*Spigelia*, *Strychnos*)
Gentianaceae, incl. Potalieae (*Anthocleista*, *Exacum*, *Fagraea*, *Gentiana*)
Apocynaceae, incl. Asclepiadaceae (*Apocynum*, *Asclepias*, *Kopsia*)
Loganiaceae pro parte (*Gelsemium*, *Mostuea*)

Fig. 4.4. Condensed representation of part of fig. 5 of Olmstead et al. (1993). The position of Rubiaceae at the base of Gentianales and the exclusion of Menyanthaceae (not shown) is typical of *rbc*L-based studies.

Struwe et al.'s (1994) cladistic analysis of morphological, anatomical, embryological and phytochemical data, the most comprehensive non-molecular analysis of Gentianales up to date, focusses particularly on the the genera of Loganiaceae sensu Leeuwenberg & Leenhouts (1980). As in the macromolecular analyses, the Rubiaceae are part of a monophyletic Gentianales, and Loganiaceae are a higly unnatural group (fig. 4.5). The Rubiaceae are well nested within Gentianales, with the loganiaceous Gelsemieae as sister group. This is contradicted by the *rbc*L-sequence data as well as by the restriction site variation of the chloroplast DNA inverted repeat, where the Rubiaceae are the sister group of the remaining Gentianales. As suggested by Struwe et al. (1994), further study of non-macromolecular traits as well as further *rbc*L sequencing within Loganiaceae s.l. may help to clarify this issue.

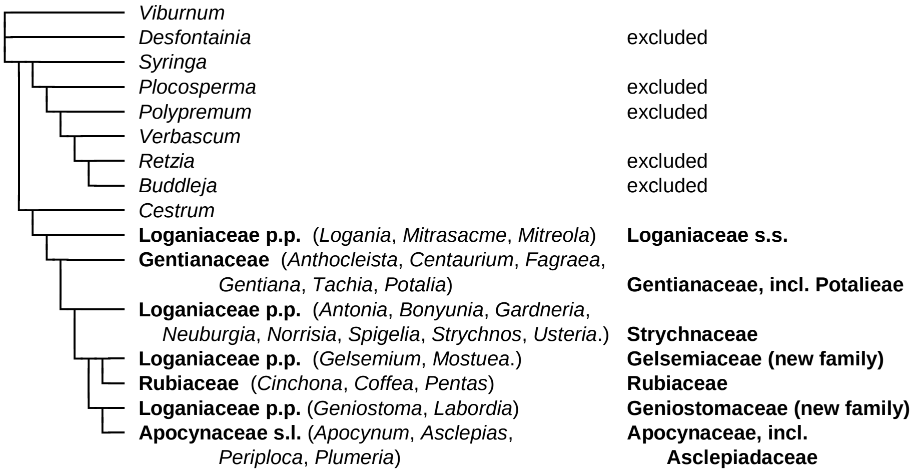

Viburnum
Desfontainia excluded
Syringa
Plocosperma excluded
Polypremum excluded
Verbascum
Retzia excluded
Buddleja excluded
Cestrum
Loganiaceae p.p. (*Logania*, *Mitrasacme*, *Mitreola*) **Loganiaceae s.s.**
Gentianaceae (*Anthocleista*, *Centaurium*, *Fagraea*, *Gentiana*, *Tachia*, *Potalia*) **Gentianaceae, incl. Potalieae**
Loganiaceae p.p. (*Antonia*, *Bonyunia*, *Gardneria*, *Neuburgia*, *Norrisia*, *Spigelia*, *Strychnos*, *Usteria*.) **Strychnaceae**
Loganiaceae p.p. (*Gelsemium*, *Mostuea*.) **Gelsemiaceae (new family)**
Rubiaceae (*Cinchona*, *Coffea*, *Pentas*) **Rubiaceae**
Loganiaceae p.p. (*Geniostoma*, *Labordia*) **Geniostomaceae (new family)**
Apocynaceae s.l. (*Apocynum*, *Asclepias*, *Periploca*, *Plumeria*) **Apocynaceae, incl. Asclepiadaceae**

Fig. 4.5. Cladistics and classification of the Gentianales on the basis of morphological, anatomical, embryological, and phytochemical data (Struwe et al. 1994). The excluded taxa were included in Loganiaceae by Leeuwenberg & Leenhouts (1980).

Struwe et al.'s (1994) analysis confirms the exclusion of *Buddleja*, *Desfontainia* and *Retzia* from Loganiaceae and Gentianales, and adds the problematic genera *Plocosperma* and *Polypremum* to this list. Within Gentianales, the remaining genera of the Loganiaceae are scattered over five different monophyletic groups. Four of these are recognized as distinct families in the newly proposed familial classification of the order, while *Potalia*, *Fagraea* and *Anthocleista* (tribe Potalieae) are included in Gentianaceae. The presence of interpetiolar stipules or stipular lines, the presence of colleters, and the presence of vestured pits in the wood are discriminating characters for the order (Struwe et al. 1994). Other characters are widespread in Gentianales, but there are more or less important exceptions: the opposite, entire leaves, the internal phloem, the regular flowers with an isomerous androecium, the contorted aestivation, the nuclear endosperm formation, and the presence of indole alkaloids belonging to the group of complex seco-iridoids (Wagenitz 1959, 1977, 1992, Jensen 1992).

4.4 Outlook

Formulating his guiding principles for making up a phylogenetic classification, Thorne (1976) also expressed some "strong convictions or concepts about classification". Taking into account that an approach that seeks a compromise between classifications of established authorities is still advocated nowadays (Nicholas & Baijnath 1994), his ninth conviction is worth quoting: "Phylogeny cannot be achieved by consensus. Although some botanists seem to think that a proper system of classification might be derived by popular vote and compromise among the modern phylogenetists this is an unrealistic goal in view of our overwhelming lack of adequate knowledge of the angiosperms. One taxonomist may be right and ten other taxonomists may be wrong. Time and accumulation of more data will have to decide who, if any, was correct. A closer approach to unanimity of phylogenetic thought for the angiosperms should some day be possible, but no one should expect unanimity in our time".

Today, unanimity about the existence of a monophyletic Gentianalean group is almost reached, even though several details of its exact delineation and several important questions about its internal structure remain unsolved. This unanimity was not reached by a majority vote as proposed by Nicholas & Baijnath (1994), but it is based on a steadily increasing data base, provided by monographic work and the study of morphological and molecular characters, as anticipated by Thorne. However, at least as important has been the development of cladistics. Firstly cladistics has stimulated the development of a conceptual framework that enables us to think and talk in a clear way about phylogenetic relationships. In practice, all available data are evaluated simultaneously during cladogram construction; the resulting cladograms permit the distiction between homologous and homoplasious similarities, and between monophyletic groups and groups that reflect merely levels of advancement. Moreover, it would be almost impossible to interpret molecular data such as gene sequences without computerized analyses (other approaches besides cladistics have been developed for molecular data, see Darlu & Tassy 1993 for an overview).

The construction of a data matrix requires an explicit specification of all characters and character states that are used, and of their distribution over the analysed taxa. By this explicit emphasis on characters, it might have been expected that the rise of cladistics during the previous decades would have been accompanied by an increased interest in character research. However, the mere fact that a cladistic analysis is performed, does in itself not guarantee that the used characters are carefully circumscribed and well studied. Morphological and anatomical characters used in higher level systematics of angiosperms are a case in point. Indeed, the classic descriptive terminology was not developed with the intention to reflect topological correspondence (Rieppel 1988) or hypothese of primary homology (de Pinna 1991), but often it is taken for granted that it provides characters that can be used immediately in cladistic analyses. An example are the floral nectaries : e.g. a ring of nectariferous tissue in the flower is very often called a "nectary disk". However, ontogenetically some of these nectary disks are derived from gynoecial tissue, other from receptacular tissue and still other from androecial tissue. Therefore it is better to abandon the superficial similarity in adult morphology and position, and to homologise nectary disks that are derived from gynoecial (receptacular, androecial) tissue with other nectaries that are derived from gynoecial (receptacular, androecial) tissue, even though these can be very dissimilar when adult (Smets 1988b, 1989, Smets & Cresens 1988). The gynoecium provides an other illustration: Igersheim et al. (1994) recently showed that during early development the so-called superior ovary of *Gaertnera* is basically inferior. Only during later stages the ovary deviates from the characteristic development in Rubiaceae, and it becomes secondarily superior. Clearly, a lot of basic morphological work remains to be done in angiosperms, and continuing character research will certainly lead to improved homologisations, not only for relatively recently discovered ultrastructural features such as sieve-element plastids (Behnke & Barthlott 1983, Behnke 1991) or epicuticular waxes (Theisen & Barthlott 1994), but even for very familiar macromorphological features, as the above examples show.

In this way, it can be expected that, also for Gentianales, the quality of morphological data sets will steadily improve, and that the most interesting results indeed will be obtained when molecular and morphological data are confronted with each other (e.g. Hillis 1987, Sytsma 1990, Donoghue & Sanderson 1992, Bremer & Struwe 1992, Patterson et al. 1993, Soltis et al. 1993, Bachmann 1995, Moritz & Hillis 1996).

5. PHYLOGENY OF TEMPERATE GENTIANACEAE: A MORPHOLOGICAL APPROACH[8]

5.1 Introduction

The Gentianaceae is a cosmopolitan family of medium size, with 76 genera (Brummitt 1992) and about 1200 species (Mabberley 1990; see table 5.1). Its oldest known fossils are from the Eocene of North and Central America (Crepet & Daghlian 1981, Graham 1984). Recent cladistic analyses based on on *rbc*L sequence data (Olmstead et al. 1993, Bremer et al. 1994), restriction site variation of the chloroplast genome (Downie & Palmer 1992) and morphological, anatomical, embryological and chemical data (Struwe et al. 1994) indicate that Gentianaceae are one of the principal families of the monophyletic order Gentianales. Results in Bremer et al. (1994) and Struwe et al. (1994) are consistent with the hypothesis (e.g. Downie & Palmer 1992, Bremer & Struwe 1992) that Loganiaceae sensu Leeuwenberg & Leenhouts (1980) are a paraphyletic assemblage with members showing closest relationships to other families both within and outside of the Gentianales. As far as Gentianaceae is concerned, Struwe et al.’s (1994) main conclusion is to formally include *Potalia* Aubl., *Fagraea* Thunb. and *Anthocleista* Afzel. ex R. Br. (tribe Potalieae of Loganiaceae sensu Leeuwenberg & Leenhouts 1980) in the Gentianaceae. This transfer had already been proposed by Bureau (1856) in the previous century and more recently by Fosberg & Sachet (1980) on the basis of gross morphology (although monographers of the Loganiaceae disagreed, e.g. Leeuwenberg & Leenhouts 1980) and by Jensen (1992) on the basis of the presence of advanced iridoid glucosides. It should be noted that the inclusion of *Anthocleista* and *Fagraea* increases the woody paleotropic representation of the family, that is otherwise restricted to *Gentianothamnus* Humbert (Humbert 1937).

While a consensus seems to be emerging about the monophyly of the Gentianales and the inclusion of Potalieae in Gentianaceae, much work remains to be done concerning the interfamilial relationships within the order, including the relationships of the smaller families often included in Gentianales (e.g. Saccifoliaceae, Dialypetalanthaceae) and concerning the infrafamilial relationships of the bigger Gentianales families. We focus on the Gentianaceae.

[8] Reprinted with small modifications from Mészáros, S., De Laet, J. & E. Smets (1996)

Because the broad-based cladistic analyses (e.g. Downie & Palmer 1992, Olmstead et al. 1993, Bremer et al. 1994, Struwe et al. 1994) to date included few representatives of the Gentianaceae, they are not very informative with respect to its problematic and unclear infrafamilial relationships. The three main monographs (Grisebach 1845, Bentham 1876, Gilg 1895) that deal with the systematics of the Gentianaceae all date from previous century. More recent classifications of the family exist, but these are based on taxa occurring in local floras, and not on a worldwide survey (e.g. Garg 1987, Zuyev 1990). Grisebach (1845) and Bentham (1876) use mainly characters of anthers, styles, stigmas and ovaries, while Gilg (1895) based his classification almost exclusively on pollen features. He distinguished two subfamilies: Gentianoideae and Menyanthoideae; within the subfamily Gentianoideae he recognized five tribes: Gentianeae (with subtribes Exacinae, Erythraeinae, Chironiinae, Gentianinae and Tachiinae), Rusbyantheae, Helieae, Voyrieae and Leiphaimeae. Gilg’s classification has been much criticized, major issues being the position of Menyanthoideae and the status of the neotropical (sub)tribes.

The Menyanthoideae proved to be very different from Gentianoideae, and on the basis of anatomical, embryological and phytochemical evidence it was raised to family level by Wagenitz (1964). Gross floral and vegetative morphology point to a close affinity with either Solanales or Gentianales (cf. Cronquist 1981), but both *rbc*L sequence data (Chase et al. 1993, Olmstead et al. 1992, 1993) and restriction site variation of the chloroplast genome (Downie & Palmer 1992) associate the family with Campanulales/Asterales.

Gilg's mainly or exclusively neotropical (sub)tribes Rusbyantheae, Helieae, Voyrieae, Leiphaimeae and Tachiinae have often been criticized for being artificial or redundant groups. Maas (1984a) noted that the neotropical genus *Lisianthius* P. Browne and a number of related neotropical shrubby genera (the “lisanthoid gentians”, Sytsma 1988) are scattered over Helieae, Tachiinae and Rusbyantheae, resulting in a very unnatural grouping of genera. It is now agreed (Weaver 1974, Maas 1984b) that *Rusbyanthus cinchonifolius* Gilg, the only species in Rusbyantheae, is to be included in *Macrocarpaea* Gilg (Tachiinae). *Voyriella* Miq. (Leiphaimeae) is considered to be related to the genera *Curtia* Cham. & Schltdl. and *Tapeinostemon* Benth. (Erythraeinae), and *Leiphaimos* Cham. & Schltdl. (the second genus of Gilg’s Leiphaimeae) is included in *Voyria* Aubl. (Weaver 1974, Maas & Ruyters 1986). In this way, the tribes Rusbyantheae and Leiphaimeae are redundant (Weaver 1974). Wood & Weaver (1982) proposed merging the tribe Helieae and the subtribe Tachiinae, and Fosberg & Sachet (1980) suggested lumping Tachiinae and Potalieeae. Gilg’s (1895) subtribes Exacinae, Erythraeinae, Chironiinae and Gentianinae have been less criticized. The criticisms are mainly restricted to transfers from the neotropical subtribes; e.g. *Hockinia* Gardner (Tachiinae) to Erythraeinae (Maas & Ruyters 1986), *Tachiadenus* Griseb. (Tachiinae) to Exacinae (Klackenberg 1987), and *Eustoma* Salisb. (Tachiinae) and *Coutoubea* Aubl. (Helieae) to Erythraeinae (Kaouadji 1990).

Based on morphological, cytological or chemical data, phylogenetic hypotheses or evolutionary trees (without any cladistic methodology) have been published for *Gentiana* L. and *Gentianella* Moench. (Scharfetter 1953), *Blackstonia* Huds. and *Centaurium* Hill (Zeltner 1970) and for the subtribe Gentianinae (Toyokuni 1963, 1965, Massias et al. 1982). Cladistic analyses exist for *Exacum* L. (Klackenberg 1985), *Tachiadenus* Griseb. (Klackenberg 1987), *Lomatogonium* A. Braun (Liu & Ho 1992), and part of *Lisianthius* P. Browne (Sytsma & Schaal 1985). In order to study xanthone evolution in Gentianaceae, Mészáros (1994) performed a cladistic analysis of a group of 12 genera of Gentianinae, Erythraeinae and Tachiinae.

In this study we extend Mészáros's (1994) data set both in number of characters and number of taxa, and we present a more complete cladistic analysis of the family. We use mainly morphological and anatomical characters and to a lesser extent chemical data. Because of limited availability of the xanthone data, especially for tropical taxa, we focus on temperate representatives of the family.

5.2 Material and methods

5.2.1 Taxa.

A total of 21 genera of Gentianaceae (including Potalieae) were selected. In addition to the principal genera indicated in table 5.1, we included the former loganiaceous genera *Anthocleista* Afzel. ex R. Br. and *Fagraea* Thunb., and the following smaller genera (numbers of species and distributional data from Mabberley 1990): *Blackstonia* Huds. (Erythraeinae; 5-6; Europe), *Chorisepalum* Gleason & Wodehouse (Tachiinae; 5; Guyana highlands), *Eustoma* Salisb. (Tachiinae; 3; southern North to northern South America), *Hoppea* Willd. (Erythraeinae; 2; India), *Ixanthus* Griseb. (Gentianinae; 1; Canary Islands), and *Orphium* E. Mey. (Chironiinae; 1; Southern Africa). Excluding the redundant tribes Rusbyantheae and Leiphaimeae, the chosen genera represent all Gilg's tribes and subtribes except the monogeneric Voyrieae. To reduce problems with polymorphisms (Nixon & Davis 1991), we have split up the genus *Gentiana* and included three European sections (following Pringle 1978) for which xanthone compounds are well documented: *Gentiana* sectio *Calathianae* Froel., *Gentiana* sectio *Ciminalis* (Adans.) Dumort. and *Gentiana* sectio *Gentiana*.

Table 5.1. Principal genera of Gentianaceae listed according to classification of Gilg (1895). Over 75% of total number of species in Gentianaceae belongs to listed genera (about ¼ of total number of genera recognized by Brummitt 1992). Unless indicated otherwise, the numbers of species (second column) and the distributional data (third column) are from Mabberley (1990; his species estimates are low; e.g. in the Exacum monograph of Klackenberg (1985) 65 species are recognized). Genera marked with "*" are included in this study (see text for further details and additional included genera).

Gentianineae		
Exacinae		
Exacum L. *	c. 25	palaeotropics
Sebaea Sol. ex R. Br.	60	Africa to India, Australia, New Zealand
Erythraeinae		
Canscora Lam. *	30	palaeotropics
Centaurium Hill *	30	northern hemisphere, one extending to Australia, one to Chile
Faroa Welw.	17	tropical Africa
Sabatia Adans.	17	northern America, West Indies
Chironiinae		
Chironia L. *	c. 15	subSaharan Africa, Madagascar
Gentianinae		
Frasera Walter *	15	northern America
Gentiana L. *	c. 300	temperate and arctic, usually montane elsewhere but absent from Africa
Gentianella Moench *	125	temperate, excluding Africa
Gentianopsis Ma *	16-25	northern temperate Asia and America
Halenia Borkh. *	c. 70	Eurasian mountains, America
Lomatogonium A. Braun	18	temperate Eurasia
Swertia L. *	50	northern temperate, African mountains
Tripterospermum Blume *	25	from Japan and South Korea to the Himalayas, Sri Lanka and Indonesia (excl. Borneo) (Murata 1989)
Tachiinae		
Lisianthius P. Browne *	27	tropical America
Macrocarpaea Gilg *	30	tropical America
Helieae		
Schultesia Mart.	20	tropical Africa and America
Symbolanthus G. Don *	15	tropical America
Voyrieae		
Voyria Aubl.	30	tropical America, western Africa

The former loganiaceous genera *Anthocleista* and *Fagraea* were included as outgroups (Nixon & Carpenter 1993). Shared synapomorphies with Gentianaceae sensu stricto (i.e. Gentianaceae excluding Potalieae) are the presence of bilobed placentas, the presence of xanthones, and the presence of swertiamarin and other unique seco-iridoids (Struwe et al. 1994). With respect to Potalieae, Gentianaceae sensu stricto (the ingroup) is defined by the absence of stipules and the presence of capsular fruits (characters 30 and 31). However, the assumption that Gentianaceae sensu stricto are monophyletic is contradicted in most broad-based cladistic analyses of molecular and morphological data (e.g. Downie & Palmer 1992, including *Fagraea*, *Exacum*, *Gentiana*, *Lisianthius* and *Obolaria*; Olmstead et al. 1993, including

Anthocleista, *Fagraea*, *Exacum*, and *Gentiana*; Struwe et al. 1994, including *Anthocleista*, *Fagraea*, *Potalia*, *Centaurium*, *Gentiana* and *Tachia*). In these analyses, however, few representatives of Gentianaceae were included, and therefore they may not be very reliable as far as intrafamilial structure of Gentianaceae is concerned: coarse sampling within a clade may lead to a wrong connection of the clade to the rest of the tree (see e.g. Olmstead et al. 1993: in an analysis of Asteridae with few Gentianales included, they obtained a branching sequence within the order that was almost exactly the reverse of what they found when more Gentianales were included; see also Olmstead et al. 1992: 261-263, Struwe et al 1994: 188-189). Nevertheless, it would be too easy to dismiss these results a priori as artifacts of taxon sampling, and we are currently extending our data matrix with additional genera of Loganiaceae and other families of Gentianales to address the question of monophyly of Gentianaceae sensu stricto. In the present analysis, we will arbitrarily depict all cladograms as rooted between *Anthocleista* and the other genera and we will shortly discuss the effect of alternative root positions on our results.

5.2.2 Characters.

32 morphological and anatomical and 8 chemical characters (tables 5.2 and 5.3) were used in the cladistic analysis. Data for morphological character states were compiled mainly from literature, in some cases supplemented with herbarium studies (BP and BR). Literature data were collected either from monographs (mainly Gilg 1895, Kusnezow 1896-1904, Allen 1933, Ewan 1948, Marais & Verdoorn 1963, Weaver 1972, Leeuwenberg 1980, Maguire 1981, Wood & Weaver 1982, Garg 1987, Murata 1989) or from papers dealing with specific characters (mainly Perrot 1898, Hasselberg 1937, Lindsey 1940, Metcalfe & Chalk 1950, Krishna & Puri 1962, Patel et al. 1981, Nishino 1983, Carlquist 1984, Neubauer 1984).

The xanthone data are from the same sources as in Mészáros (1994), to which new information on *Hoppea*, *Chironia* and *Orphium* was added (Stout et al. 1969, Rezende & Gottlieb 1973, Chapelle 1974, Okorie 1976, Carbonnier et al. 1977, Gottlieb 1982, Hostettmann & Wagner 1977, Massias et al. 1977, Sullivan et al. 1977, Ghosal et al. 1978, Hostettmann-Kaldas & Jacot-Guillarmod 1978, Luong et al. 1980, Dreyer & Bourell 1981, Hostettmann-Kaldas et al. 1981, Lin et al. 1982a, 1982b, Massias et al. 1982, Sluis 1985, Lin et al. 1987, Ortega et al. 1988, Khetwal et al. 1990, Bennett & Lee 1991, Wolfender et al. 1991, Wolfender & Hostettmann 1992, Roitman et al. 1992). The majority of the flavanoid data are from Kaouadji (1990) and Hegnauer (1989) (but see also the sources of the xanthone data). The sugar data are from Massias et al. (1978).

Table 5.2. Characters and character states.

1. Life form: trees or shrubs (0) herbs (1)
2. Xylem rays: multi- and uniseriate (0) only uni-(bi-)seriate (1) rayless (2)
3. Nodal anatomy: unilacunar (in *Swertia* sometimes also trilacunar) (0) multilacunar (1)
4. Leaves: petiolate (0) sessile (1) perfoliate (2)
5. Morphological type of stomata: anomocytic (0) paracytic (1)
6. Leaf venation: penninerved (0) parallel veined (1)
7. Calcium oxalate crystals in mesophyll: absent (0) present (1)
8. Calyx symmetry: actinomorphic (0) zygomorphic (1)
9. Inflorescence: dichasium (0) monochasium (1) flowers in clusters (2) solitary flowers (3)
10. Fusion of sepals: scarcely (0) half (1) almost completely (2)
11. Intracalycine membrane: absent (0) present (1)
12. Calyx lateral traces: free (0) fused at origin (1) fused throughout (2)
13. Corolla mery: polymerous (0) pentamerous (1) tetramerous (2)
14. Petal fusion: scarcely (0) half (1) almost completely (2)
15. Corolla aestivation: contorted (0) plicate (1)
16. Pollen: in tetrads (0) in monads (1)
17. Nectaries: none (0) epipetalous (1) gynoecial (2)
18. Anther fixation: basifixed (0) versatile (1)
19. Anther twisting: none (0) moderately (1) largely (2)
20. Anther abortion: none (0) 1-3 aborted stamina (1) only 1 fertile stamen (2)
21. Anther cohesion: free (0) connate (1)
22. Ovary: 4-locular (0) 2-locular (1) unilocular (2)
23. Ovary shape: globular (0) oval (1) long (2)
24. Ovary: sessile (0) stipitate (1)
25. Placentation: axial (0) parietal (1) superficial (2)
26. Carpel ventral traces: free (0) fused at origin (1) fused throughout (2)
27. Seed shape: angular (cubical) (0) globular (1) oval (2) long (3)
28. Seed wing: absent (0) present (1)
29. Flavonoids: flavonol (O-glycosides) (0) flavones (C- or O-glycosides) (1)
30. Sugars: simple (glucose, primverose, rhamnose, galactose) (0) compound (gentianose, gentiobiose) (1)
31. Stipules: absent (0) present (1)
32. Fruit: capsular (0) baccate (1)
33. Seed testa surface: smooth (0) with reticulum of thickened radial cell walls (1)
34. Seed testa-cell shape: isodiametric (in *Exacum* sometimes also star-shaped) (0) elongated (1)
35. Oxygenation of xanthone position C2: absent (0) present (1)
36. Oxygenation of xanthone position C4: absent (0) present (1)
37. Oxygenation of xanthone position C5: absent (0) present(1)
38. Oxygenation of xanthone position C6: absent (0) present (1)
39. Oxygenation of xanthone position C7: absent (0) present (1)
40. Oxygenation of xanthone position C8: absent (0) present (1)

Most of the palynological, embryological and cytological data that we reviewed were excluded from the matrix because of insufficient coverage. Still, 16% of the data matrix cells are scored as missing or inapplicable.

Table 5.3. Data matrix. Numbers of characters and character states refer to table 5.2. "-" indicates polymorphisms in binary characters; polymorphisms in multistate characters are indicated between square brackets; "?" indicates missing values and inapplicable characters.

				111	111111122	2	2222	2223333	333333
	0123	45678	9	012	345678901	2	3456	7890123	456789
Anthocleista	0110	00?00	0	0?0	201??0000	1	00?[02]	0??1110	000--1
Fagraea	0110	?0?00	[12]	0?1	201??000[12]	1	0-?0	0??11??	??????
Symbolanthus	00?0	?0?0[13]	0	0?1	2002?0002	1	01?0	0??00??	??????
Chorisepalum	0??0	?0?0[03]	0	0?0	201200001	2	00?0	0??00??	??????
Lisianthius	-??0	?0100	[01]	020	201210001	2	01?0	0?000??	??????
Chironia	-??1	?1?01	0	001	001002002	0	0101	0?00-10	001111
Orphium	0?01	?1?01	0	0?1	001?01002	1	0101	0??0010	000011
Macrocarpaea	0??0	?0?00	0	001	20-200001	1	00?0	0??00??	100111
Eustoma	1??1	?1101	0	001	001211002	1	0101	0000010	001111
Canscora	1??1	01100	2	002	1010?0102	1	0101	010001?	001111
Hoppea	1??1	01?00	1	012	201??0202	0	0121	0100010	001110
Centaurium	1201	01100	0	001	101002002	2	0101	0000010	-0----
Blackstonia	12?2	?1101	0	0?0	001?01002	1	01??	000001?	000011
Ixanthus	11?2	?1?00	1	0?1	101200002	1	01?[12]	0?0001?	100101
Swertia	1?00	01000	0	001	001110002	[01]	0102	--100-0	---0--
Halenia	1??0	?1001	0	002	001110002	[12]	0102	0?1000?	1110-0
Gentianella	1??1	11000	1	001	101110002	2	1221	010000?	0-10-1
Gentianopsis	1??1	11011	1	122	101110002	2	1213	010001?	000011
Tripterospermum	1??0	?1?01	1	0?1	211200002	2	1201	1?00-??	0-0-1-
Gentiana sectio Gentiana	1210	?1112	2	101	211200012	1	1202	1110011	-00010
Gentiana sectio Ciminalis	1201	11103	1	101	211200012	2	1203	0110011	000011
Gentiana sectio Calathianae	1201	11003	2	101	111200002	2	1203	0110011	000011
Exacum	-2?[01]	-1?00	[01]	02[12]	001?00001	0	0010	01?0010	??????

5.2.3 Coding of the xanthone data.

Xanthones are yellow-coloured dibenzo-γ-pyron compounds that arise biosynthetically from a benzephenone precursor that is derived from acetate (leading to ring A) and shikimate (ring B). With the exception of the widespread compound mangiferin they occur only in a limited number of tracheophyte families; in the angiosperms, they are found mainly in Guttiferae and Gentianaceae (Gottlieb 1982, Frohne & Jensen 1992). Discussions of xanthone evolution center around the degree of oxygenation of the aromatic rings, and are often based on an a priori designation of the primitive type of oxygenation pattern from which the other observed patterns are deduced. Based on the biosynthetic pathway of xanthones, Rezende & Gottlieb (1973) and Gottlieb (1982) suggested that 1,3-dioxygenation of ring A and 5,6- or 6,7-dioxygenation of ring B is the primitive oxygenation pattern in all families that have xanthones. Gottlieb (1982) derived the other observed oxygenation patterns from this type on the basis of a common-is-primitive argument. Carbonnier et al. (1977) and Massias et al. (1982), working on Gentianinae, assumed that trioxygenation is primitive and higher degrees of oxygenation are increasingly derived. Mészáros (1994) was the first to apply cladistic reasoning to this problem, but his analysis was constrained by assuming Camin-Sokal parsimony, which does not allow reversals.

In his study of xanthone evolution in Gentianaceae, Mészáros (1994) did not directly code the oxygenation patterns (absence/presence of oxygenation at the different C-positions), but he used four characters that are derived from patterns as they are observed: minimal grade of substitution, diversity of substitution, specialization of ring A, and specialization of ring B. However, cladistic characters should represent primary hypotheses of homology (de Pinna 1991) and therefore should reflect certain correspondences of parts. This is problematic for these derived characters. For instance, in the character “specialization of ring A”, the states are dioxygenation, trioxygenation and tetraoxygenation, and it is perfectly possible that the state “trioxygenation” refers to different sets of positions in different genera, or even in the same genus (e.g. 1-2-3 or 1-3-4, which would imply the hypothesis that an oxygenated C2 corresponds somehow to an oxygenated C4). For this reason, we choose to code the xanthone data as absence/presence characters describing whether or not each of the different C-positions is oxygenated. Genera in which a certain position is oxygenated in some species or in some xanthones but not in others are coded as polymorphic for that position. We did not distinguish between the different substituents (hydroxyl, methoxyl, O-glycosyl) that may occur on the oxygenated C-positions because this variation seems to be subsidiary to the oxygenation pattern (cf. Hostettmann & Wagner 1977 for *Gentiana* and Wolfender and Hostettmann 1992 for *Chironia*). Positions C1 and C3 of ring A are oxygenated throughout, leaving positions C2 and C4 from ring A and positions C5-C8 of ring B as informative xanthone characters.

5.2.4 Methods.

Standard parsimony analyses with a priori equal weighting of all characters were carried out using NONA (Goloboff 1993b). We also performed analyses using implied weighting (Goloboff 1993a), a method that is based on the concept of cladistic reliability of characters (Farris 1969; cf. Carpenter 1988, 1994). In this approach, characters are non iteratively weighted during tree search by means of a concave function of their homoplasy. It should be noted that implied character weights are different from the weights that can be assigned to the characters prior to the analysis (these a priori weights were mostly kept equal; see below). We refer to Goloboff (1993a) for further theoretical background. Following Goloboff (1993a), we will call the resulting cladograms the fittest cladograms, as opposed to the most parsimonious cladograms of the standard approach. Searches for fittest cladograms were carried out with the computer program Pee-Wee (Goloboff 1993c). In Pee-Wee, the degree of concavity of the weighting function is determined by the concavity constant K (Goloboff 1993c). Beyond the fact that the weighting function should be concave (Farris 1969), it is far from obvious how it should look exactly. For this reason we tried several values of K and compared the results. We varied K between its minimum (1; highest concavity, i.e. strongest differential downweighting of homoplasy) and its maximum (6; lowest concavity, i.e. lowest differential downweighting of homoplasy; this comes closest to the standard approach). In order to avoid confusion, we note that the concavity constants K (Goloboff 1993c) and k (Goloboff 1993a) are not equal (K=k+1).

In most analyses we treated all multistate characters as unordered (cf. Hauser 1992). However, in our data set all multistate characters except 8, 16, and 24 (see table 5.2) represent fairly straightforward morphoclines and hence can be ordered very well using the similarity criterion (cf. Lipscomb 1992; all morphoclines are linear and the numerical codes of the states of these characters in table 5.2 follow the order of the morphoclines). Treating these characters as unordered would imply that some of the observed primary homologies (de Pinna 1991) are dismissed a priori. For this reason we also ran analyses in which these characters were ordered. Carpenter (1988) showed that the way in which ordered multistate characters are coded (additive binary or ordinal) can influence the final stable solution under successive weighting (Farris 1969): using ordinal coding (Mickevich & Weller 1990) distorts the picture because it yields higher weighting of these characters simply because they are coded that way. Although Goloboff (1993a) did not mention it, the situation is similar when using implied weights. To avoid this distortion, we derived a second data set from table 5.3 to perform the ordered analyses. In this data set, we coded the linearly ordered multistate characters in a binary additive way. As the polymorphisms in the ordered characters involved only adjacent states, it was not necessary to expand observed subsets of states to ranges of states.

In all analyses, we used subset coding for polymorphisms. Polymorphisms in terminal taxa may indicate that the terminal taxa are non-monophyletic, a possibility that may not be overlooked when using large and traditionally defined genera. The best way to avoid unwarranted assumptions of monophyly is to split up polymorphic taxa into monomorphic subunits (Nixon & Davis 1991), an approach we informally followed when splitting up *Gentiana* and including three of its sections. Little is known of the effect of using subset coded polymorphic taxa. However, in their analysis of the Gentianales, Struwe et al. (1994) compared subset coding and a version of monomorphic subtaxon recoding and found highly compatible results as far as the global branching pattern is concerned.

Relatively many taxa are polymorphic for the xanthone characters. This may be an indication that the xanthone characters we have delineated do not capture the variation that is relevant for this taxonomic level: they may be more useful at lower taxonomic levels. The degree of polymorphism may actually even be higher than apparent from table 5.3: in many cases, especially for the bigger genera, the xanthone scores are generalized from only a limited number of species. For these reasons it can be argued that the xanthone characters should not get the same a priori weights as the other characters. Following this line of thought, we also did some analyses in which the xanthone characters were excluded or given lower a priori weights than the other characters.

In all analyses with NONA and Pee-Wee, the most parsimonious cladograms or the fittest cladograms were obtained using the instruction MULT*25. This instruction carries out 25 replications of randomizing the taxa, creating a Wagner tree and submitting it to branch-swapping by means of tree-bissection reconnection. Apart from the setting of the a priori weights, the value of the concavity constant, and the ordering of multistate characters, all other default settings were retained in all analyses. By default, NONA collapses all branches that have no unambiguous synapomorphies (a character provides an unambiguous synapomorphy for a branch if a state transition occurs on that branch under every possible optimization of the character on the tree; Goloboff 1993c; see also Coddington & Scharff 1994). The ensemble consistency indices (CI; Kluge & Farris 1969) and ensemble retention indices (RI; Farris 1989) for the standard parsimony analyses were derived from the “minimum” tables, giving minimum and maximum possible steps for each character. Consistency indices are calculated with autapomorphies included (see Yeates 1992). The distribution of the nine autapomorphic states in the matrix is as follows: one in each of the binary characters 15, 35, and 38; one in each of the multistate characters 1, 8, 11 and 21; two in multistate character 19.

In order to evaluate the relative support of clades, we calculated branch support, i.e. the number of extra steps needed to loose a branch in the strict consensus of near-most-parsimonious trees (Bremer 1994; also called “Bremer support” or “decay index”). For similar reasons a bootstrap analysis was performed (Felsenstein 1985; but see Bremer 1994). The calculation of branch support values (‘decay analysis’) and the bootstrap analysis were performed with PAUP (Swofford 1993; characters unordered; initial seed = 1; heuristic search by means of simple addition and tree-bisection-reconnection branch swapping; one tree held at each step during stepwise addition). For the bootstrap analysis MAXTREES was set to 1000 and 100 replicates were run.

5.3 Results

With equal a priori weighting and all characters unordered, the standard parsimony analysis resulted in eight most parsimonious trees (steps 111; CI=0.51; RI=0.64). The strict consensus of these is shown in fig. 5.1. The trichotomy involving clades two, three, and six is present in all of the most parsimonious trees. *Fagraea* and *Symbolanthus* are present in the basal polytomy because *Fagraea* is the sister group of clade one in two of the eight most parsimonious trees, while it branches below the polytomy in the other cases; *Symbolanthus* is the sister group of clade 1 or of clade 1 + *Fagraea* in all trees.

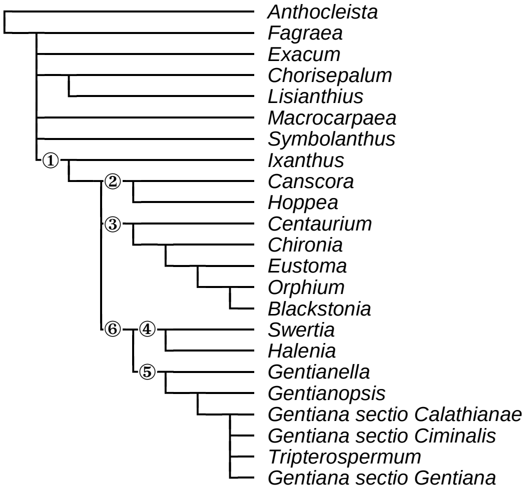
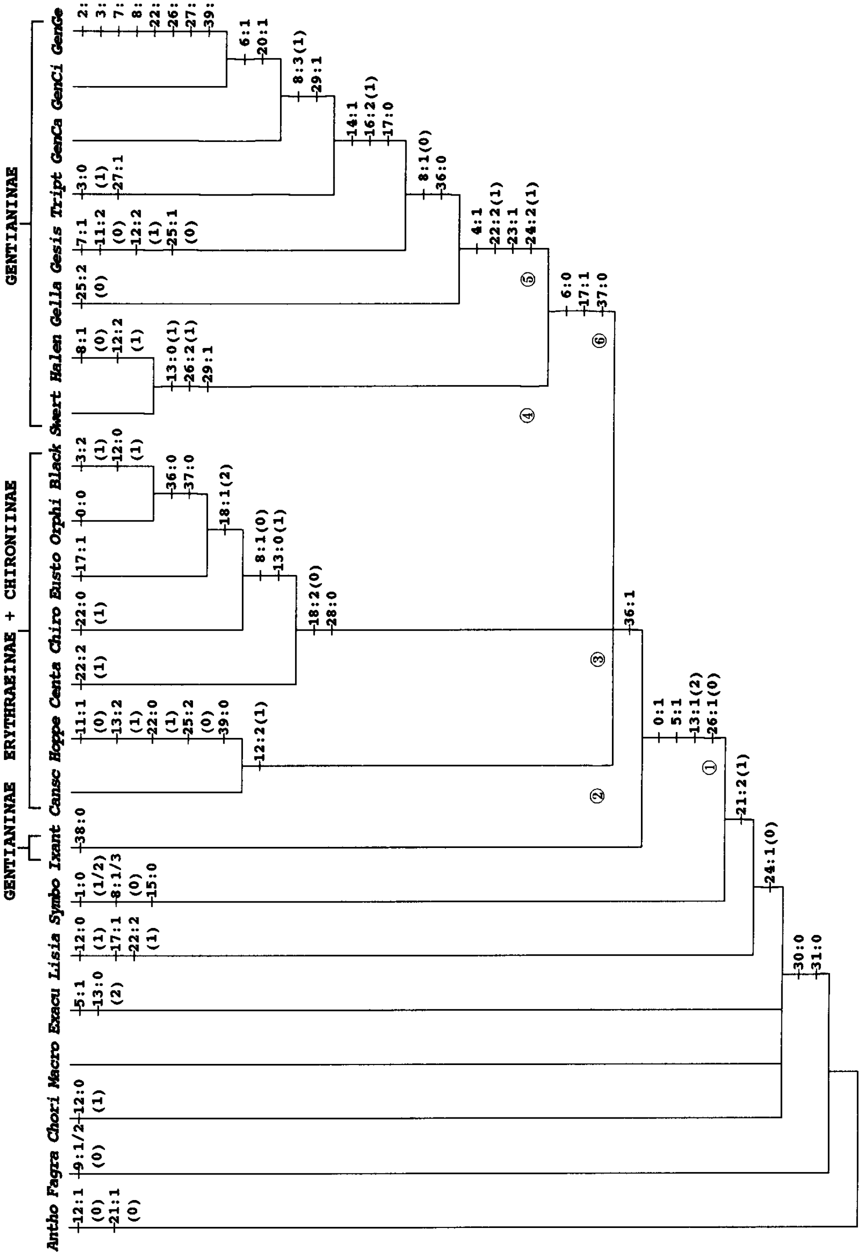


Fig. 5.1. Strict consensus of the eight most parsimonious trees (steps 111) for the data of table 5.3, all characters unordered. Numbered clades are discussed further in text. Bremer branch support value of clade four is 2; all other clades have branch support 1.

In the decay analysis, PAUP found 1160 trees of length ≤ 112. The strict consensus of these is completely unresolved except for a sister-pair relation between *Swertia* and *Halenia*. The search for trees with a length ≤ 113 was stopped prematurely because of memory limitations when 3100 trees were found. The strict consensus of this partial result already refutes the sister-pair relation between *Swertia* and *Halenia*. This comes down to Bremer branch support 2 for *Swertia-Halenia*, while the other branches of the strict consensus of the most parsimonious trees (fig. 5.1) have branch support 1. Low branch support values are typical for morphological data sets (Karis 1995); in this particular case the low values are at least partially due to the fact that the positions of some genera, especially *Exacum*, vary greatly within a topology that is otherwise fairly constant (see Wilkinson 1994a for a general discussion of this problem). This is clear when the strict consensus is calculated with *Exacum* excluded. In this case, a clade containing all taxa of clade 1 is still present in the consensus of all trees of length ≤ 112 and in the consensus of the 3100 trees of length ≤ 113. This implies a branch support value of at least 2 for a group that is nested (Adams 1986) within the set of all terminal taxa excluding *Exacum* and that is composed of all taxa of clade one. In terms of monophyly this means that clade one has a branch support of at least 2 when the question whether *Exacum* belongs to it is left open. In order to overcome the memory limitations, we also ran an alternative analysis in which *Exacum* and *Fagraea*, two genera with strongly varying positions, were excluded from the data matrix. As expected, this resulted in less most parsimonious and near-most-parsimonious trees. In the consensus of the 6 most parsimonious trees (length 104) for this reduced matrix, clade 1 is present and has the same internal structure as in the strict consensus of the 8 most parsimonious trees of the full analysis (fig. 5.1). It survives in trees up to 106 steps long.

On a Performa 450 computer, it took almost 32 hours to complete the bootstrap analysis (MAXTREES was set to 1000 to constrain the duration; in 29 out of the 100 replicates the search for shortest trees was stopped prematurely because of tree-buffer overflow). Only 6 clades are present in at least 50% of all trees that were found. These are (a) the clade that contains all terminal taxa except *Anthocleista* and *Fagraea* (69%), (b) the sister pair *Chorisepalum-Lisianthius* (50%), (c) the sister pair *Swertia-Halenia* (84%), (d) the three sections of *Gentiana* (unresolved; 51%) (e) the three sections of *Gentiana* + *Tripterospermum* (53%) and (f) the previous clade in trichotomy with *Gentianella* and *Gentianopsis* (51%).

Fig. 5.2. (next page) Single fittest cladogram (fit 263.1; K=3) with equal a priori weights, all characters unordered. Apart from the position of *Lisianthius*, the cladogram is identical to one of the eight most parsimonious trees of the unordered standard analysis (fig. 5.1). Numbers of character states and characters refer to table 5.2; analysis is based on the data matrix of table 5.3. Only unambiguous synapomorphies are shown (a:b stands for state b of character a; for multistate characters, number between brackets indicates character state that transforms into the synapomorphic state). *Gella* and *Gesis* are *Gentianella* and *Gentianopsis*; *GenCa*, *GenCi* and *GenGe* are the sections *Calathianae*, *Ciminalis* and *Gentiana* of the genus *Gentiana*; the other taxa are indicated by their five first characters. *Eustoma* is included in Erythraeinae + Chironiinae. Numbered clades are discussed further in text.


GENTIANINAE
ERYTHRAEINAE + CHIRONIINAE
GENTIANINAE
Antho Fagra Chori Macro Exacu Lisia Symbo Ixant Cansc Hoppe Centa Chiro Eusto Orphi Black Swert Halen Gella Gesis Tript GenCa GenCi GenGe
12:1 (0) 21:1 (0)
9:1/2 (0)
12:0 (1)
5:1 13:0 (2)
12:0 (1) 17:1 22:2 (1)
1:0 (1/2) 8:1/3 (0) 15:0
38:0
11:1 (0) 13:2 (1) 22:0 (1) 25:2 (0) 39:0
22:2 (1)
22:0 (1)
17:1
0:0
3:2 (1) 12:0 (1)
8:1 (0) 12:2 (1)
25:2 (0)
7:1 11:2 (0) 12:2 (1) 25:1 (0)
3:0 (1) 27:1
2:1 3:0(1) 7:1 8:2(3) 22:1(2) 26:2(3) 27:1 39:0
6:1 20:1
8:3(1) 29:1
14:1 16:2(1) 17:0
8:1(0) 36:0
4:1 22:2(1) 23:1 24:2(1)
6:0 17:1 37:0
13:0(1) 26:2(1) 29:1
36:0 37:0
18:1(2)
8:1(0) 13:0(1)
18:2(0) 28:0
12:2(1)
36:1
0:1 5:1 13:1(2) 26:1(0)
21:2(1)
24:1(0)
30:0 31:0
①
②
③
④
⑤
⑥

As in the decay analysis, the poor result of the bootstrap analysis is partially due to the varying positions of some genera within a topology that is fairly constant otherwise. This can be illustrated by the following example. PAUP reports that a clade composed of all taxa of clade one appears only in 36% of all trees found during the analysis. Consequently, this clade is not retained in the bootstrap tree. However, a clade composed of all these taxa and *Exacum* appears in 33% of all trees. In combination, this implies that in 69% of all trees a clade composed of all taxa of clade one is nested (Adams 1986) within the set of all terminal taxa excluding *Exacum*. In terms of monophyly this means that the clade is fairly well supported (69%) in the bootstrap analysis, but that we do not know whether *Exacum* belongs to it or not. When the uncertainty of the exact composition of the clade is extended to *Ixanthus*, the support value for this clade is 88%.

Under the same conditions as in the above standard analysis (equal a priori weighting and all characters unordered) the search for the fittest trees resulted in a single fittest cladogram with fit=263.1 for the concavity constant K equal to 3 (fig. 5.2). Apart from the position of Lisianthius, this cladogram is identical to one of the eight fundamental trees of the unordered standard analysis.

Varying the concavity constant (all characters unordered) has little effect for values between 2 and 6: K=6 gives three fittest trees (fit 296.9), K=5 results in two trees (fit 288.9), as does K=4 (fit 278.2). Each time the strict consensus tree is as in fig. 5.2, except that *Lisianthius* is the sister genus of *Chorisepalum*. For K=2 the same fittest tree (fit 239.8) is found as for K= 3 (fig. 5.2). Setting K to 1 (i.e. strong concavity) results in two fittest trees (fit 201.7). The first is similar to fig. 5.2 (only the positions of *Gentianopsis* and *Tripterospermum* are switched), but the second (fig. 5.3a) has a deviant topology: clade three is disrupted to form two subclades and *Exacum* is the sister group of *Hoppea*.

With equal a priori weighting but all multistate characters except 8, 16, and 24 ordered, the standard parsimony analysis resulted in six most parsimonious trees (steps 118; CI=0.48; RI=0.64). In two of these most parsimonious cladograms *Exacum* is the sister group of *Hoppea*, while it is a more basal branch in the other four. Because this results in a highly unresolved strict consensus tree, the strict consenus excluding *Exacum* is shown (fig. 5.3b). *Fagraea* and *Symbolanthus* join the basal polytomy because *Fagraea* is the sister group of clade one in two of the six most parsimonious cladograms. The trichotomy involving clades two, three, and six is present in all of the most parsimonious trees. Under the same conditions (equal a priori weighting but all multistate characters except 8, 16, and 24 ordered), three fittest trees (fit 348.6) are obtained for K=3. The strict consensus is shown in fig. 5.3c.

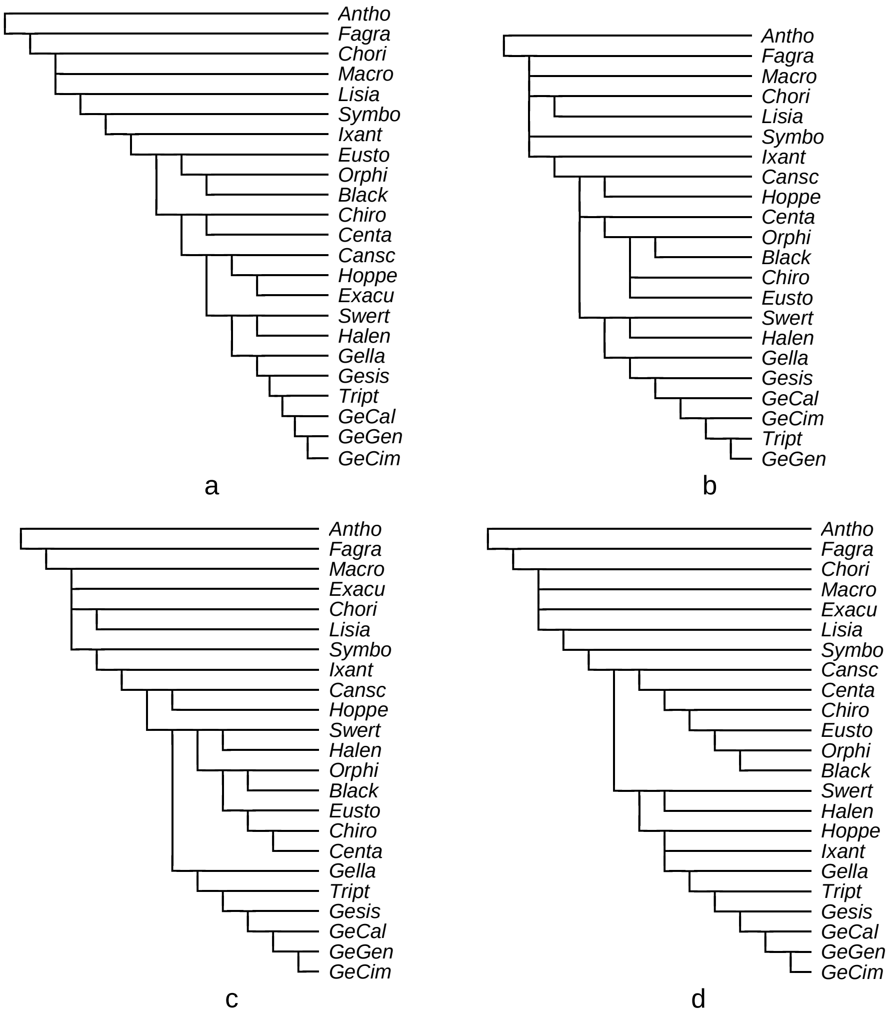


Fig. 5.3. Results of some alternative analyses of the data in table 5.3; a. One of two fittest cladograms with equal a priori weighting, all characters unordered, and K=1 (fit 201.7); the other fittest tree is as in fig. 5.2, but with *Gentianopsis* and *Tripterospermum* switched. b. Strict consensus (excluding *Exacum*) of six most parsimonious trees (steps 118), multistate characters (except 8, 16, and 24) ordered. In two of the most parsimonious cladograms *Exacum* is the sister group of *Hoppea*, while it is a more basal branch in the other four. c. Strict consensus of the three fittest trees (fit 348.6; K=3) with multistate characters (except 8, 16, and 24) ordered. d. Single fittest tree (fit 1215.2; K=3) when the a priori weights for the xanthone characters are set to 1 and for the other characters to 5; all characters unordered.

Varying the concavity constant (all multistate characters except 8, 16, and 24 ordered) has no effect for values between 2 and 6: each time the same three trees are obtained (with a fit of 316.5 for K=2, 348.6 for K=3, 368.4 for K = 4, 382.2 for K =

5, and 392.9 for K=6). Setting K to 1 results in four different fittest trees (fit 263.6). The strict consensus of these is as in fig. 5.2, but with clades three and five differently and/or less resolved, and with *Ixanthus* joining the polytomy between clades two, three, and six. The latter is due to the fact that clade two (*Canscora-Hoppea*) is the sister group of *Ixanthus* in three of the fittest trees, while it forms a polytomy with clades three and six in the fourth one.

Changing the a priori weights in order to downweight the xanthone characters affects the topology when they are downweighted fivefold or more with respect to the other characters (K=3, all characters unordered). When the a priori weights are set to 1 for the xanthone characters and to 5 for the other characters, a single fittest cladogram is obtained (fig. 5.3d; fit 1215.2). The most conspicuous differences with fig. 5.2 are the disruption of clade two and the position of *Ixanthus*. Except for an unresolved *Eustoma-Orphium-Blackstonia* clade, the same tree is obtained when the xanthone characters are excluded completely.

5.4 Discussion

The best supported clade of this study is clade one (fig. 5.2), containing *Eustoma* (Gilg’s Tachiinae) and all included Gentianinae, Erythraeinae and Chironiinae. In the unordered analysis using implied weights (fig. 5.2), its unambiguous synapomorphies are herbaceous life form, parallel leaf venation, globular seeds and intermediate petal fusion. This result confirms Carlquist’s (1984; contra Wood & Weaver 1982: 445) suggestion that a woody habit may be plesiomorphic in the family.

Leaving aside the question whether *Exacum* belongs to it, clade one is present in all analyses. The position of *Exacum* is dubious: it belongs to clade one (as sister genus of *Hoppea*) in two of the six most parsimonious trees with ordered characters and in one of the two fittest cladograms with unordered characters and K=1; it falls outside clade one in all other cases. The uncertain position of *Exacum* is also obvious from the decay analysis and the bootstrap analysis, as shown earlier.

The relationships between clade one on the one hand and the woody tropical representatives of Tachiinae and Helieae on the other are not clear: the cladograms are either poorly resolved below clade one or resolved incongruently among different analyses. The only recurring pattern is the position of *Symbolanthus* (Helieae) as the sister group of clade one. The sole exceptions to this are the few cladograms in which *Fagraea* is the sister group of clade one (two of the eight most parsimonious trees in the unordered analysis and two of the six in the ordered analysis; in these cladograms, *Symbolanthus* remains the sister group of *Fagraea* + clade one). The fact that *Fagraea* appears as the sister genus of clade one in some cladograms leaves open the possibility that Gentianaceae sensu stricto (excluding Potalieae) may be paraphyletic, a result that was also obtained in the higher mentioned broad-based analyses (Downie & Palmer 1992, Olmstead et al. 1993, Struwe et al. 1994). Clade one, however, is unaffected by the alternative root positions as obtained by Downie & Palmer (1992), Olmstead et al. (1993) or Struwe et al. (1994). Therefore we will concentrate on this clade and its internal structure.

The structure of clade one can be visualized as a basal division between *Ixanthus* and the other genera, that in turn belong to two major clades, clade three and clade six. The position of *Canscora* and *Hoppea* within clade one is not clear: when taking into account only the unambiguous synapomorphies, *Canscora* and *Hoppea* are mostly a sister pair that is in a trichotomy with clades three and six (e.g. figs. 5.1, 5.2, 5.3c). Under some conditions, or when ambiguous support is considered, *Canscora* + *Hoppea* appear as the sister pair of *Ixanthus*, as the sister group of clade three + clade six, as part of clade three, or as part of clade six. Lastly, in the analyses with strong downweighting or complete exclusion of the xanthone characters (fig. 5.3d), the sister group relationship between both genera is disrupted: *Canscora* is in clade three, and *Hoppea* in clade six.

The basal division between *Ixanthus* and the other genera is present in most of the analyses. Only under extreme conditions of strong concavity or strong differential a priori weighting other results are obtained (in the ordered analysis with K=1, the basal division is between (*Ixanthus* (*Canscora Hoppea*)) and the other genera in three of the four fittest trees; with strong downweighting or complete exclusion of the xanthone characters, *Ixanthus* is the sister genus of *Gentianopsis* + the three sections of *Gentiana*). *Ixanthus*, a perennial herbaceous plant that becomes woody only at the base of the stem, is an endemic of the laurel forests of the Canary Islands (Bramwell 1972). Taking into account the strong asymmetry of the basal division of clade one and the limited amount of character change on the branch leading to *Ixanthus*, this genus can be interpreted as a kind of living fossil with a character distribution that is intermediate between the mostly woody tropical genera below clade one and the mostly herbaceous temperate taxa within clade one. Indeed, even if the herbaceous life form does occur in the tropical genera, e.g. in *Lisianthius*, it is mostly connected with the Mediterranean or temperate climate that prevails in clade one. The absence of temperate representatives below clade one points to a tropical ancestry of the Gentianaceae. Other, more circumstantial, indications for tropical ancestry are that two thirds of the genera are tropical or have species in the tropical regions (Nilsson

1970, Favarger 1987), that all the other families of the order Gentianales are tropical, and that the oldest fossils represent tropical genera (*Macrocarpaea* and *Lisianthius*; Crepet & Daghlian 1981, Graham 1984). Carlquist (1984) interpreted the presence of interxylary phloem and the lack of rays in the juvenile wood of *Ixanthus* as advanced features which were probably acquired through the adaptation to "winter cold as well as a fluctuation between winter rainfall and summer drought".

Clade three is composed of a mixture of Erythraeinae (*Canscora* and *Hoppea* dubious), Chironiinae and the genus *Eustoma* of Tachiinae. This clade is present in all but one of the reported cladograms, in which it is paraphyletic (one of the two fittest cladograms of the unordered analysis with K=1; fig. 5.3a). The unambiguous synapomorphies of this clade are twisted anthers and the presence of flavonols (instead of flavones). The inner structure of clade three is not clear. The sister group relationship of *Chironia* and *Centaurium* in some of the analyses (figs. 5.3a, 5.3c) is in line with their supplementary geographical areas. However, if *Centaurium* originated in the European paleomediterranean area where its extant diploid species exist (Zeltner 1970), this could also mean that *Centaurium* is paraphyletic.

Clade six, the second major subclade of clade one, is composed of the representatives of subtribe Gentianinae, excluding *Ixanthus*. The unambiguous synapomorphies are the absence of calcium oxalate crystals in the mesophyll, versatile anthers and absence of oxygenetion of C6. A problematic point is the position of the sister pair *Swertia* and *Halenia*, which belongs to clade three in the ordered analyses with K between 2 and 6 (cf. fig. 5.3c), but to clade six in all other analyses. In clade six, *Swertia-Halenia* (clade four) and the other Gentianinae (clade five) are sister groups. This relationship is in agreement with the recent treatment of Liu & Ho (1992). Garg (1987) and Zuyev (1990), on the other hand, considered this group as a separate tribe.

Apart from the position of *Tripterospermum*, the structure of clade five is as depicted in fig. 5.2 in all analyses. Its unambiguous synapomorphies include three well documented features of the ovary, namely superficial placentation, long, and stipitate. Traditionally *Tripterospermum* is mostly included in *Gentiana* (e.g. Marquand 1937) or, mainly since the revision of Smith (1965), treated as a separate genus closely related to *Gentiana*. In our analyses its position varied: in some cladograms it was the sister group of *Gentianopsis* + the three sections of *Gentiana*, in others it was the sister group of the three sections of *Gentiana*, and lastly it sometimes appeared as the sister group of *Gentiana* sectio *Gentiana*. *Tripterospermum* is generally considered to be related to *Gentiana* sectio *Stenogyne* Franch., which is either regarded as "the primitive type of the genus” (Ho & Liu, 1990: 186) or as a "more advanced group than the other sections of the genus" (Yuan & Küpfer 1993: 72; based on chromosome number and karyotype asymmetry). As *Gentiana* sectio *Stenogyne* is not included in our data set, and as the position of *Tripterospermum* was variable in our analysis, it is premature to draw any conclusions concerning the status of *Tripterospermum*.

A lot of questions remain regarding infrafamilial classification. Nevertheless, some tentative proposals can be made. Our results corroborate the traditional composition of the subtribe Gentianinae (probably with the exception of *Ixanthus*). On the other hand, Gilg's (1895) subtribes Erythraeinae and Chironiinae are clearly not monophyletic and we propose merging them, with inclusion of the genus *Eustoma* (Tachiinae). Because this clade (including *Eustoma*) is the sister group of subtribe Gentianinae, it is better to retain subtribal rank (Chironiinae) within the tribe Gentianeae. Following the sequencing convention (Nelson 1972; cf. Wiley 1979), the genus *Ixanthus* can be accomodated in a third subtribe Ixanthinae of a narrowly defined tribe Gentianeae (clade one). *Swertia* and *Halenia* probably belong to subtribe Gentianinae, while the positions of the genera *Canscora* and *Hoppea* within this tribe are not clear. The possibility that they constitute a separate subtribe cannot be excluded. Subtribe Tachiinae (excluding *Eustoma*) falls outside the tribe Gentianeae as defined here, while the status and the position of subtribe Exacinae is not clear. The relationships between Tachiinae, Exacineae, Gentianeae, and Gilg’s (1895) other tribes are not clear. An analysis of a broader array of taxa, including a wider array of outgroup taxa, will be necessary to elucidate these relationships.

5.5 Summary

The intrafamilial relationships of the Gentianaceae are investigated by means of a cladistic analysis based on morphological and to a lesser extent on chemical data. The 21 genera that are selected for the analysis represent all tribes and subtribes except Leiphaimeae, Rusbyantheae and Voyrieae. The large genus *Gentiana* is represented by three of its sections. The former loganiaceous genera *Anthocleista* and *Fagraea* are used as outgroups.

Standard parsimony analyses and analyses using weights that are based on the cladistic reliability of the characters give congruent results as far as the global relationships are concerned. The best supported clade contains *Eustoma* (Tachiinae) and all included Gentianinae, Erythraeinae and Chironiinae. The basal division in this clade is between *Ixanthus* and the other genera. In this way *Ixanthus*, an endemic of the Canary Islands, connects the mostly woody tropical and the mostly herbaceous temperate taxa. Subtribe Gentianinae (excluding *Ixanthus*) is monophyletic, unlike Erythraeinae and Chironiinae. In most analyses, however, both subtribes together (and including *Eustoma*) are the sister-group of Gentianinae. Possibly Erythraeinae, Chironiinae and *Eustoma* should be merged.

The basal parts of the cladograms, involving the woody tropical representatives and *Exacum*, are poorly resolved. More extensive sampling, especially among the tropical representatives, is necessary to elucidate these basal relationships.

6. INDECISIVE DATA AND MISSING ENTRIES[9]

6.1 Introduction

The cladistic decisiveness of a data set can be defined as the degree to which all possible resolved trees for the data set differ in length (Goloboff 1991a, 1991b). The larger this degree, the stronger is the conclusion that the worst cladograms can be safely discarded. Therefore the cladistic decisiveness of a data set stands for the information for tree choice that is present in the data. Data sets that are fully indecisive are data sets for which every possible tree has the same length. For binary characters and when missing entries are not allowed, only data sets that contain every possible informative character state distribution in an equal number are fully indecise (Goloboff 1991a; any amount of uninformative characters may be added). In Goloboff's terminology, which is followed here, an indecisive data set for n taxa refers to a data set for n taxa (the ingroup) to which an all-zero outgroup is added. In this way, an informative character is a character that satisfies both following conditions:

1. at least one terminal taxon of the ingroup has state zero
2. at least two terminal taxa of the ingroup have state one.

Two examples of indecisive data sets are shown in fig. 6.1. For three taxa, the only informative characters are the characters that have the apomorphic state in two taxa. For four taxa, the characters with the apomorphic state in three taxa are also informative. When no missing entries are present, these data sets are essentially the only indecisive data sets that exist for three and four taxa. Besides adding informative characters, the only variation that is possible is to repeat every character for an equal number of times. An indecisive matrix that contains all possible informative characters for n taxa precisely once will be called the **minimal indecisive matrix** for n taxa. For a given number of taxa n, the minimal indecisive matrix contains 2^n-n-2 characters, and because only binary characters are considered this number is also the ensemble observed variation, M (Goloboff 1991a; see also appendix C).

Goloboff (1991a, b) restricted his discussion to data sets in which no missing entries are present. In this chapter I will show how indecisive data sets can be constructed when missing entries are allowed, and discuss some properties. Next I

[9] Based on a presentation held at the XIIIth Meeting of the Willi Hennig Society (August 23 - 26, 1994, Copenhagen, Denmark; see De Laet & Smets 1994b).

will use these data sets to evaluate Goloboff's (1991a) data deciveness index DD, an index that was proposed to measure quantitatively the decisiveness of data.

outgroup	000
A	**011**
B	**101**
C	**110**
	$\underbrace{}_{A_2}$

outgroup	000000	0000
A	**111000**	**0111**
B	**100110**	**1011**
C	**010101**	**1101**
D	**001011**	**1110**
	$\underbrace{}_{A_2}$	$\underbrace{}_{A_3}$

Fig. 6.1. Minimal indecisive data sets for three and four taxa. An A_i character is a character with i 1-entries. The three possible rooted trees for the first data set have all five steps; the fifteen possible rooted trees for the second data set have all eighteen steps.

6.2 Missing entries and indecisiveness

When missing entries are allowed, the basic observation is that an indecisive data set for n+1 taxa can be produced simply by adding a row of missing entries to an indecisive data set for n taxa, which is shown for three taxa in the left part of fig. 6.2. In the character state distributions for taxa A, B and C the indecisive data set for three taxa can be recognized, while taxon D has only missing entries. It may seem rather absurd to add such an uninformative taxon to the data set, but it can be combined with similar sets to obtain less trivial cases, as shown in the second example of fig. 6.2. An other obvious possibility is to combine an indecisive data set that has missing entries with the undecisive data set that has no missing entries, as shown in the third example. In any indecisive data set, the smallest indecisive subsets that contain informative characters will be called **minimal indecisive subsets**.

outgroup	000
A	**011**
B	**101**
C	**110**
D	**???**

outgroup	000 000
A	**011 011**
B	**101 ???**
C	**110 101**
D	**??? 110**

outgroup	000000 0000 000
A	**111000 0111 110**
B	**100110 1011 101**
C	**010101 1101 110**
D	**001011 1110 ???**

Fig. 6.2. Some examples of indecisive data sets that contain missing entries.

A more elaborate example for five taxa is shown in fig. 6.3. The data set consists of four minimal indecisive subsets. For minimal indecisive data sets without missing entries. Goloboff (1991a) derived formulas for M, the ensemble observed variation; S, the total number of steps on a most parsimonious tree, and G, the total length on an unresolved bush, as a function of the number of taxa. Using these formulas, the values of 25, 60 and 51 are obtained for the first indecisive subset.

out	0000000000000000000000000	0000000000	000	000	
A	**1111000000111111000001111**	**1110000111**	**110**	???	
B	**1000111000111000111010111**	**1001101011**	???	**110**	
C	**0100100110100110110111011**	??????????	**101**	**101**	
D	**0010010101010101101111101**	**0101011101**	**011**	???	
E	**0001001011001011011111110**	**0010111110**	???	**011**	
M	**25**	**10**	**3**	**3**	⇒ **41**
G	**60**	**20**	**6**	**6**	⇒ **92**
S	**51**	**18**	**5**	**5**	⇒ **79**

C = 0.52
RI = 0.25

Fig. 6.3. A compound indecisive data set for 5 taxa.

Because the presence of one or more rows of missing entries in any data set does not influence the values of M, S or G, these formulas can also be used for minimal subsets with missing entries. It is sufficient to substitute the total number of taxa for the number of taxa that do not have missing entries, which is four for the second one (M=10, G=20, S=18) and three for the last two (M=3, G=6, S=5). In order to obtain the values of M, G and S for the complete data set, the values of the composing minimal subsets can simply be added. This is obviously true for M and G, that are calculated on a character per character basis. It is also true for S because every minimal data set is indecisive in itself. With the values of M, S and G, the ensemble consistency and retention indices can then be calculated.

The above procedure for calculating the consistency index and the retention index for any possible indecisive data set is straightforward. However, Goloboff's (1991a) formula for calculating the number of steps for the minimal data sets is not practical because it is a recursive formula with many summation operators that are nested in up to two levels (moreover, it is only valid for seven taxa or more). An exact and easy-to-calculate formula (fig. 6.4) is derived in appendix C.

$$S(n) = \frac{1}{9}\left(2^n * (3n+1) - (-1)^n\right) - (n+1)$$

$$G(n) = (n+1) * (2^{n-1} - 1) - \frac{n+1}{2} * \binom{n}{[(n+1)/2]}$$

Fig. 6.4. Number of steps S(n) and ensemble value G(n) for minimal indecisive data sets with n taxa.

As far as G is concerned, Goloboff provided two exact formulas: one for an even and one for an odd number of taxa. Since these formulas are not recursive and contain only a single summation operator, they are relatively easy to calculate. Nevertheless, the summation operator can be eliminated and a simpler formula that is valid for an even as well as for an odd number of taxa is derived in appendic C (fig. 6.4; $\binom{n}{i}$, with 0=<i=<n, stands for n!/(i!*(n-i!)); the square brackets stand for the integer part of the bracketted expression).

Knowing that $M=2^n$-n-2 (Goloboff 1991a, see above) and using these formulas for S and G, it is easy to calculate how an indecisive data set must be composed in terms of minimal submatrices in order to have any specified ensemble consistency index CI or retention index RI (CI = M/S, Kluge & Farris 1969; RI = (G-S)/(G-M), Farris 1989). The only limitation is that the specified values must lie between the upper and lower bounds shown on figs. 6.5 and 6.6.

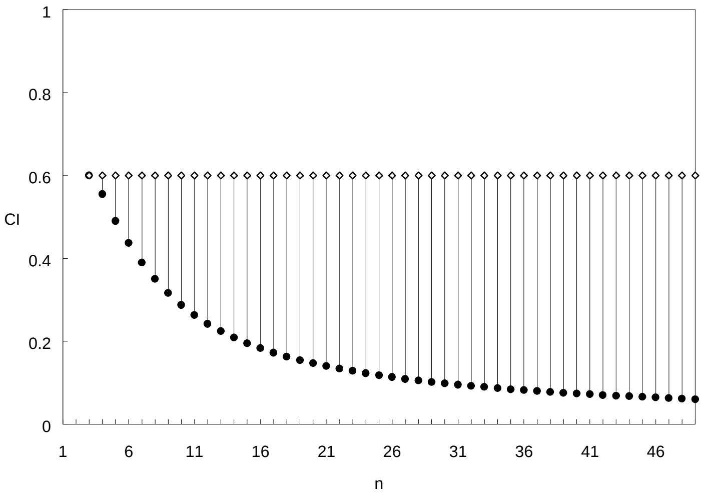


Fig. 6.5. Possible ranges of ensemble consistency index CI(n) for indecisive data sets with n taxa; see text for explanation.

The possible ranges for CI are given in fig. 6.5. The lower bound (black dots) is achieved when no missing entries are present and decreases as n increases (cf. Goloboff 1991a, fig 1). The constant upper bound of CI=0.6 is reached in indecisive data sets that contain only three-taxon statements.

The possible ranges for RI are given in fig. 6.6. The lower bound of RI=0.2 is reached in indecisive data sets that contain only four-item statements, the upper bound of RI=0.33 in indecisive data sets that contain only three-item statements. The retention indices of minimal indecisive data sets without missing entries are indicated as black dots (cf. Goloboff 1991a, fig. 3). Using the above formulas, it is easily verified that RI rises asymptotically to 1/3 when no missing entries are present.

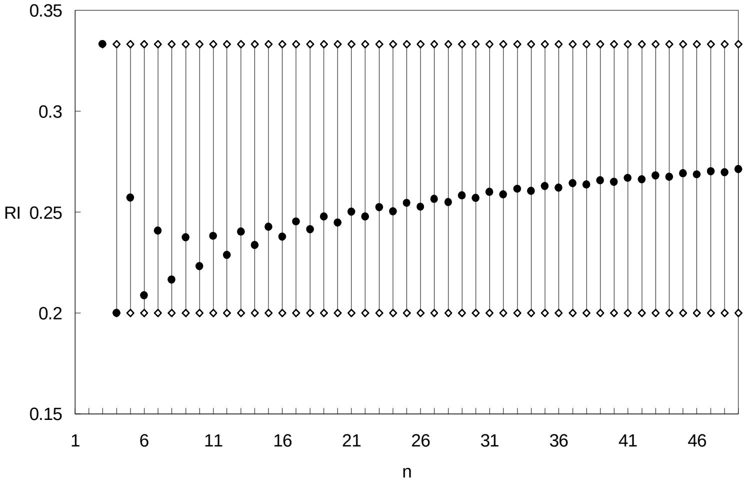


Fig. 6.6. Possible ranges of ensemble retention index CI(n) for indecisive data sets with n taxa.

6.3 Measuring data decisiveness

In the general definition of decisiveness, the question of how the degree of decisiveness of a data set should be be measured is left open. Goloboff (1991a) argued that possible measures should be sensitive to the degree in which the possible trees for a data set differ in length, but insensitive to the degree of homoplasy that is present in it. Therefore, he rejected CI, RI and RC as measures for decisiveness and he proposed a new index, the DD statistic, that is based on the fact that the mean number of steps on all fully resolved trees for a given data set is independent of the amount of homoplasy in it. DD is then defined as the degree to which the length of a most parsimonious tree (i.e. the minimum possible length for the data set, S_{MIN}) deviates from this mean length, and this degree is scaled such that trees that have no homoplasy have a DD value equal to 1:

$$DD = \frac{\bar{S}-S_{MIN}}{\bar{S}-M}$$

However, by using indecisive data sets with missing entries, it can be shown that DD is also directly influenced by the amount of homoplasy: two pairs of data sets are presented, and for each pair the shape of the distribution of the tree lengths is identical, but the amount of homoplasy differs. Because the distributions of tree lengths have an identical shape, the degree to which the possible trees differ in length is identical, and the data sets should have the same decisiveness. However, because the data sets have different amounts of homoplasy, they have different DD-values, as will be shown. The amount of homoplasy in a data set is the difference between the length of a most parsimonious tree for that data set, S_{MIN}, and its ensemble observed variation M. The first pair of data sets (fig. 6.7) have the same minimal length but a different M, in the second example (fig. 6.9) the two data sets have the same M, but a different minimal length.

DATA SET 1

```
out 0000000000000000000000000  0000
A   0110111100011011110011000  1111
B   1011011010101101101010100  1111
C   1101101001110110100110010  1110
D   0001110111000111011110001  1100            M = 29
E   0000000000111111111101111  1000            G = 68
```

DATA SET 2

```
out 00000000000000000000000000000  0000
A   0110111100??????????110???110  1111
B   10110110100110111100???110101  1111
C   11011010011011011010101101???  1110
D   ??????????1101101001011??????  1100        M = 33
E   00011101110001110111???011011  1000        G = 66
```

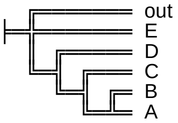


Fig. 6.7. Two data sets with an indecisive part (left) and a decisive part (right, in bold). In both cases the decisive part unambiguously resolves the relationships between taxa A-E as on the tree that is shown.

The two data sets shown in fig. 6.7 each contain an indecisive part and a decisive part. The decisive part is identical in both data sets but the indecisive parts are different. Since the decisive part is identical and consists of characters that are fully congruent among themselves, the same most parsimonious tree is found in both cases. Moreover, the indecisive parts are constructed such that both data sets require the same number of steps on the most parsimonious tree. It follows that also the distribution of tree lengths is identical (fig. 6.8). However, their value for the DD statistic is not identical: DD=0.10 for the first data set and DD=0.12 for the second (the mean number of steps is 6093/105 in both cases).

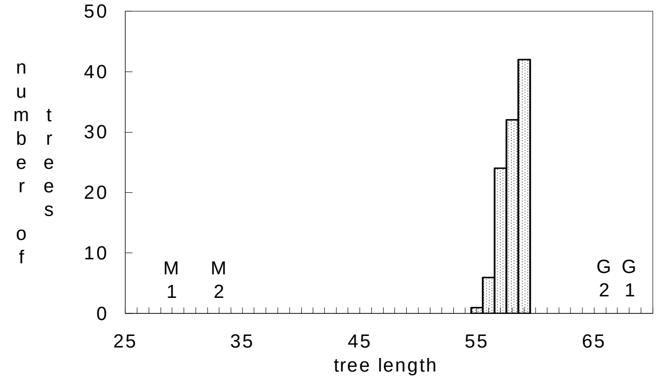


Fig. 6.8. The identical distribution of tree lenghts for both data sets of fig. 6.7. For each data set, the values of M and G are indicated.

Similar as in fig. 6.7, the two data sets of fig. 6.9 have an identical decisive part, but a different indecisive part. The indecisive parts are constructed such that they have the same M and G, but a different amount of homoplasy. As a result, the distributions of tree length for both data sets have the same shape, but they are shifted with respect to each other (fig. 6.10). Also in this case both data sets have a different DD-value: 0.0740 for data set 3 and 0.0625 for data set 4 (the mean number of steps is 53.6 and 57.6 resp.).

DATA SET 3

out	000000000000000000000000000000	**00**	
A	110110110???110110110???110110	**11**	
B	101101???110101101???110101101	**11**	
C	011???101101011???101101011???	**10**	**M = 32**
D	???011011011???011011011???011	**00**	**G = 64**

DATA SET 4

out	000000000000000000000000000000	**00**	
A	011011110001101111000110111100	**11**	
B	101101101010110110101011011010	**11**	
C	110110100111011010011101101001	**10**	**M = 32**
D	000111011100011101110001110111	**00**	**G = 64**

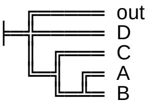


Fig. 6.9. Two data sets with an indecisive part (left) and a decisive part (right, in bold). In both cases the decisive part unambiguously resolves the relationships between taxa D-E as on the tree that is shown.

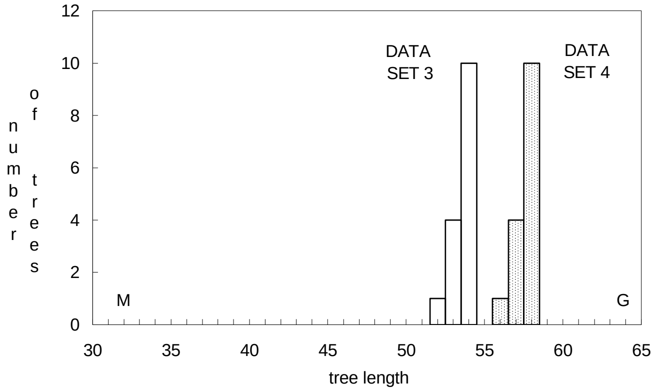


Fig. 6.10. The shape of the distribution of tree lenghts for both data sets of fig. 6.9 is identical, but they are shifted with respect to each other. M and G are the same for both data sets.

The sensitivity of DD to the amount of homoplasy follows from the fact that it is scaled in such a way that data sets without homoplasy have a DD-value of one: the scaling factor (the difference between the mean step number and M) depends on the amount of homoplasy. This is obvious because the mean number of steps minus M is simply the mean homoplasy, and as a result DD can be rewritten as the complement of the ratio of minimal and mean homoplasy:

$$DD = 1 - \frac{H_{MIN}}{\bar{H}}$$

In order to remove this sensitivity to H, the definition of a data decisiveness index should refer only to factors that describe the distribution of the tree length, and as a result such an index would simply describe some aspect of the shape of that distribution. However, in that case one would have to conclude that data set 4 (fig. 6.9) and data set 5 (fig. 6.11), consisting of the decisive part of data set 4, would have the same power to discriminate between trees, which seems difficult to defend.

DATA SET 5

out	**00**	
A	**11**	
B	**11**	
C	**10**	**M = 2**
D	**00**	**G = 4**

Fig. 6.11. Data set 5 consists of the decisive part of data set 4 (fig. 6.9), and therefore data set 4 and 5 have an identical shaped distribution of tree lengths.

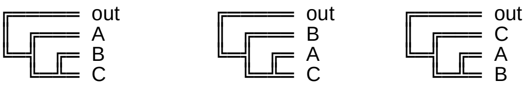
If it is accepted that the absolute level of homoplasy should influence the decisiveness of a data set, the question arises how this influence should be taken into account, and if DD does it in a sensible way. Consider the data sets of fig. 6.12: both data sets have a minimal homoplasy of 2 and a mean homoplasy of 2.66, and as a result they have the same DD-value (0.25). However, because of the distribution of possible minimal homoplasies, it might be argued that data set 6 is more decisive than data set 7 because it allows to discard at least one possible tree, the second one, rather safely: the amount of homoplasy in this tree is twice as much as the amount of homoplasy in the two other trees. Data set 7 has only a single most parsimonious tree (the third one), but the two other trees for this data set are only 1 step worse (which is a smaller difference than in data set 6).

DATA SET 6

out	0000	
A	1100	
B	1111	**M = 4**
C	0011	**G = 8**

DATA SET 7

out	0000	
A	1101	
B	1110	**M = 4**
C	0011	**G = 8**



				$\bar{\mathbf{H}}$
H, DATA SET 6:	2	4	2	2.66
H, DATA SET 7:	3	3	2	2.66

Fig. 6.12. Two data sets with the same minimal and mean homoplasy but different distributions of possible homoplasies.

One could imagine modifications of DD that would make it sensitive to differences as between data set 6 and data set 7, but it seems more appropriate to conclude that the concept of data decisiveness is too complex and elusive to be captured in a single and simple index.

REFERENCES

ADAMS, E. N. 1986. N-trees as nestings: complexity, similarity, and consensus. Journal of Classification 3: 299-317.

ALBERT, V. A. and B. D. MISHLER. 1992. On the rationale and utility of weighting nucleotide sequence data. Cladistics 8: 73-83.

ALBERT, V. A., CHASE, M. W., and B. D. MISHLER. 1993. Character-state weighting for cladistic analysis of protein-coding DNA sequences. Annals of the Missouri Botanical Garden 80: 752-766.

ALLEN, C. K. 1933. A monograph of the American species of the genus *Halenia*. Annals of the Missouri Botanical Garden 20: 119-223.

ANDERBERG, A. and B. STÅHL. 1995. Phylogenetic interrelationships in the order Primulales, with special emphasis on the family circumscriptions. Canadian Journal of Botany 73: 1699-1730.

ARCHIE, J. W. 1996. Measures of homoplasy. In Sanderson, M. J. and L. Hufford. *Homoplasy. The recurrence of similarity in evolution*: 153-188. Academic Press, San Diego.

BACHMANN, K. 1995. Progress and pitfalls in systematics: cladistics, DNA and morphology. Acta Botanica Neerlandica 44: 403-419.

BAILLON, H. 1880. Histoire des plantes 7. Hachette, Paris.

BARABÉ, D. 1984. Les principes directeurs des systèmes modernes de classification des angiospermes. Le Naturaliste Canadien 111: 21-30.

BARABÉ, D. and J. VIETH. 1990. Les principes de systématique chez Engler. Taxon 39: 394-408.

BAUM, D. 1994. *rbc*L and seed-plant phylogeny. Trends in Ecology and Evolution 9: 39-41.

BEHNKE, H.-D. 1991. Distribution and evolution of forms and types of sieve-element plastids in the dicotyledons. Aliso 13: 167-182.

BEHNKE, H.-D. and W. BARTHLOTT. 1983. New evidence from the ultrastructural and micromorphological fields in angiosperm classification. Nordic Journal of Botany 3: 43-66.

BENNETT, G. H. and H.-H LEE. 1991 . *Halenia elliptica* xanthone: a structural revision. Phytochemistry 30: 1347-1348.

BENTHAM, G. 1876. Ordo CIX. Gentianeae. In Bentham and J. D. Hooker, eds., *Genera Plantarum Vol.2*: 799-821. Reeve and Company, London: .

BESSEY, C. E. 1915. The phylogenetic taxonomy of flowering plants. Annals of the Missouri Botanical Garden 2: 109-164.

BOWLER, P. J. 1996. Life's splendid drama. Evolutionary biology and the reconstruction of life's ancestry, 1860-1940. The University of Chicago Press, Chicago.

BRADY, R. H. 1985. On the independence of systematics. Cladistics, 1: 113-126.

BRADY, R. H. 1994. Pattern description, process explanation, and the history of morphological sciences. In Grande, L. and O. Rieppel, eds., *Interpreting the hierarchy of nature. From systematic patterns to evolutionary process theories*: 7-31. Academic Press, San Diego.

BRAMWELL, D. 1972. Endemism in the flora of the Canary Islands. In Valentine, D. H., ed., *Taxonomy, Phytogeography and Evolution*: 141-159. Academic Press, London.

BREMER, B. and L. STRUWE. 1992. Phylogeny of the Rubiaceae and the Loganiaceae: congruence or conflict between morphological and molecular data? American Journal of Botany 79: 1171-1184.

BREMER, B., ANDREASEN, K., and D. OLSSON. 1995. Subfamilial and tribal relationships in the Rubiaceae based on *rbc*L sequence data. Annals of the Missouri Botanical Garden 82: 383-397.

BREMER, B., OLMSTEAD, R. G., STRUWE L., and J. A. SWEERE. 1994. *rbc*L sequences support exclusion of *Retzia*, *Desfontainia*, and *Nicodemia* from the Gentianales. Plant Systematics and Evolution 190: 213-230.

BREMER, K. 1994. Branch support and tree stability. Cladistics 10: 295-304.

BROWER, A. V. Z. and R. DESALLE. 1994. Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. Annals of the Entomological Society of America 87: 702-716.

BRUMMITT, R. K. 1992. Vascular plant families and genera. Kew: Royal Botanic Gardens.

BUREAU, L.-E. 1856. De la famille des Loganiacées, et des plantes qu'elle fournit à la médecine. Thèse pour le doctorat en Médecine. Paris: Faculté de Médecine de Paris.

CAMIN, J. H. and R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution 19: 311-326.

CARBONNIER, J., MASSIAS, M., and D. MOLHO. 1977. Importance taxonomique du schéma de substitution des xanthones chez *Gentiana* L. Bulletin du Museum National d'Histoire Naturelle, Paris, Sciences Physico-chimiques 13: 23-40.

CARLQUIST, S. 1984. Wood anatomy of some Gentianaceae: systematic and ecological conclusions. Aliso 10: 573-582.

CARPENTER, J. M. 1988. Choosing among multiple equally parsimonious cladograms. Cladistics 4: 291-296.

CARPENTER, J. M. 1994. Succesive weighting, reliability and evidence. Cladistics 10: 215-220.

CHAPELLE, J. P. 1974. Constituents chimiques des fenilles d'*Anthocleista vogelii*. Planta Medica 26: 301-304.

CHASE, M. W., SOLTIS, D. E., OLMSTEAD, R. G., MORGAN, D., LES, D. H., MISHLER, B. D, DUVALL, M. R., PRICE, R. A., HILLS, H. G., QIU, Y.-L., KRON, K. A., RETTIG, J. H., CONTI, E., PALMER, J. D., MANHART, J. R., SYTSMA, K. J., MICHAELS, H. J., KRESS, W. J., KARROL, K. G., CLARK, W. D., HEDRÉN, M., GAUT, B. S., JANSEN, R. K., KIM, K.-J., WIMPEE, C. F., SMITH, J. F., FURNIER, G. R., STRAUSS, S. H., XIANG, Q.-Y., PLUNKETT, G. M., SOLTIS, P. S., SWENSEN, S. M., WILLIAMS, S. E., GADEK, P. A., QUINN, C. J., EGUIARTE, L. E., GOLENBERG, E., LEARN, G. H. JR., GRAHAM, S. W., BARRETT, S. C. H., DAYANANDAN, S., and V. A. ALBERT. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. Annals of the Missouri Botanical Garden 80: 528-580.

CODDINGTON, J. and N. SCHARFF. 1994. Problems with zero-length branches. Cladistics 10: 415-423.

COSNER, M. E., JANSEN R. K., and LAMMERS T. G. 1994. Phylogenetic relationships in the Campanulales besed on *rbc*L sequences. Plant Systematics and Evolution 190: 79-95.

CRANE, P. R. 1985. Phylogenetic analysis of seed plants and the origin of the angiosperms. Annals of the Missouri Botnical Garden. 72: 716-793.

CREPET, W. L. and C. P. DAGHLIAN. 1981. Lower Eocene and Paleocene Gentianaceae: floral and palynological evidence. Science 214: 75-77.

CRONQUIST, A. 1975. Some thoughts on angiosperm phylogeny and taxonomy. Annals of the Missouri Botanical Garden 62, 3: 517-520.

CRONQUIST, A. 1981. An integrated system of classification of flowering plants (ed. 2). Columbia University Press, New York.

CRONQUIST, A. 1987. A botanical critique of cladism. Botanical Review 53: 1-52.

CRONQUIST, A. 1988. The evolution and classification of flowering plants (ed. 2). The New York Botanical Garden, New York.

DAHLGREN, G. 1989. The last dahlgrenogram. System of classification of the dicotyledons. In Tan, K., ed., *Plant taxonomy, phytogeography and related subjects. The Davis and Hedge Festschrift*: 249-260. Edinburgh University Press.

DAHLGREN, R. 1983a. General aspects of angiosperm evolution and macrosystematics. Nordic Journal of Botany 3: 119-149.

DAHLGREN, R. 1983b. Monocotyledon evolution. Characters and phylogenetic estimation. In Hecht, M. K., Wallace, B., and G. T. Prance, eds., *Evolutionary Biology* 16: 255-395. Plenum Publishing Corporation.

DARLU, P. and P. TASSY. 1993. Reconstruction phylogénétique, concepts et méthodes. Collection Biologie théorique, Masson, Paris.

DE JONGH, R. 1980. Some tools for evolutionary and phylogenetic studies. Zeitschrift für Zoologische Systematik und Evolutionsforschung 18: 1-23.

DE LAET, J. and E. SMETS. 1994a. Inleiding tot het cladisme. *Belgian Journal of Botany* 127: 207-229.

DE LAET, J. and E. SMETS. 1994b. Indecisiveness, missing entries and three-taxon statements. XIII Meeting of the Willi Hennig Society, Abstracts: 6. Copenhagen.

DE LAET, J. and E. SMETS. 1995. Four item analysis: an undirected implementation of the three-item approach. XIV Meeting of the Willi Hennig Society, Abstracts: 3. College Station, Texas.

DE LAET, J. and E. SMETS. 1996. A commentary on the circumscription and evolution of the order *Gentianales*, with special emphasis on the position of the *Rubiaceae*. In Robbrecht, E., Puff, C., and E. Smets, eds., *Second International Rubiaceae Conference. Proceedings.* Opera Botanica Belgica 7: 11-18.

DE PINNA, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics 7: 367-394.

DONOGHUE, M. J. 1992. Homology. In Keller, E. F. and E. A. Lloyd, eds., *Keywords in evolutionary biology*: 170-179. Harvard University Press, Cambridge.

DONOGHUE, M. J. and M. J. SANDERSON. 1992. The suitability of molecular and morphological evidence in reconstructing plant phylogeny. In Soltis P. M., Soltis D. E. and Doyle J. J., eds. *Molecular systematics of plants*: 340-368. Chapman and Hall, New York.

DONOGHUE, M. J. and P. D. CANTINO. 1988. Paraphyly, ancestors and the goals of taxonomy: a botanical defense of cladism. Botanical Review 54: 107-128.

DOWNIE, S. R. and J. D. PALMER. 1992. Restriction site mapping of the chloroplast DNA inverted repeat: a molecular phylogeny of the Asteridae. Annals of the Missouri Botanical Garden 79: 266-283.

DOYLE, J. J. 1993. DNA, phylogeny, and the flowering of plant systematics. BioScience 43: 380-389.

DREYER, D. L. and J. H. BOURELL. 1981. Xanthones from *Frasera albomarginata* and *F. speciosa*. Phytochemistry 20: 493-495.

ENDRESS, M. E., SENNBLAD, B., NILSSON, S., CIVEYREL, L., CHASE, M. W., HUYSMANS, S., GRAFSTRÖM, E., and B. BREMER. 1996. A phylogenetic analysis of Apocynaceae s. str. and some related taxa in Gentianales: a multidisciplanary approach. In Robbrecht, E., Puff, C., and E. Smets, eds., *Second International Rubiaceae Conference. Proceedings.* Opera Botanica Belgica 7: 59-102.

ENGLER, A. 1897a. Principien der systematischen Anordnung, insbesondere der Angiospermen. In Engler A. and K. Prantl, eds., *Die Natürlichen Pflanzenfamilien. Nachträge zum II.-IV. Teil*: 5-14. Verlag von Wilhelm Engelmann, Leipzig.

ENGLER, A. 1897b. Übersicht über die Unterabteilungen, Klassen, Reihen, Unterreihen und Familien der Embryophyta siphonogama. In Engler A. and K. Prantl, eds., *Die Natürlichen Pflanzenfamilien. Nachträge zum II.-IV. Teil*: 341-357. Verlag von Wilhelm Engelmann, Leipzig.

ENGLER, A. 1897c. Erlaüterungen zu der Übersicht die Embryophyta siphonogama. In Engler A. and K. Prantl, eds. *Die Natürlichen Pflanzenfamilien. Nachträge zum II.-IV. Teil*: 358-380. Verlag von Wilhelm Engelmann, Leipzig.

EWAN, J. 1948. A revision of *Macrocarpaea* a Neotropical genus of shrubby gentians. Contributions from the United States National Herbarium 29: 209-250.

FARRIS, J. S. 1966. Estimation of conservatism of characters by constancy within biological populations. Evolution 20: 587-591.

FARRIS, J. S. 1969. A successive approximations approach to character weighting. Systematic Zoology 18: 374-385.

FARRIS, J. S. 1978. Inferring phylogenetic trees from chromosome inversion data. Systematic Zoology 27: 275-284.

FARRIS, J. S. 1983. The logical basis of phylogenetic analysis. In Platnick, N. and V. A. Funk, eds., *Advances in cladistics vol 2, Proceedings of the second meeting of the Willi Hennig Society*: 7-36. Columbia University Press, New York. [Reprinted in Sober, E. , ed., 1994. *Conceptual issues in evolutionary biology, second edition*.: 333-361. MIT Press, Cambridge]

FARRIS, J. S. 1988. Hennig86. Version 1.5 (computer program and documentation). Port Jefferson Station, New York.

FARRIS, J. S. 1989. The retention index and the rescaled consistency index. Cladistics 5: 417-419.

FARRIS, J. S. 1990. Phenetics in camouflage. Cladistics 6: 91-100.

FARRIS, J. S. 1991. Hennig defined paraphyly. Cladistics 7: 297-304.

FARRIS, J. S., KÄLLERSJO, M., ALBERT, V. A., ALLARD, M., ANDERBERG, A., BOWDITCH, B., BULT, C., CARPENTER, J. M., CROWE, T. M., DE LAET, J., FITZHUGH, K., FROST, D., GOLOBOFF, P., HUMPHRIES, C. J., JONDELIUS, U., JUDD, D., KARIS, P. O., LIPSCOMB, D., LUCKOW, M., MINDELL, D., MUONA, J., NIXON, K., PRESCH, W., SEBERG, O., SIDALL, M. E., STRUWE, L., TEHLER, A., WENZEL, J., WHEELER, Q., and W. WHEELER. 1995. Explanation. Cladistics 11: 211-218.

FAVARGER, C. 1987. Quelques aspects de l'evolution et de la phylogenie dans le famille des Gentianaceae. In Bicchi, C. and C. Frattini, eds., *Atti dei: Seminari di phytochimica 1985 sulle piante contenenti principi amari*: 1-303. University of Torino.

FELSENSTEIN, J. 1979. Alternative methods of phylogenetic inference and their interrelationship. Systematic Zoology 28: 49-62.

FELSENSTEIN, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biological Journal of the Linnean Society 16: 183-196.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783-791.

FELSENSTEIN, J. 1988a. The detection of phylogeny. In Hawksworth, D., ed., *Prospects in systematics*: 112-127. Systematics Association, Clarendon Press. [Reprinted in Sober, E., ed., 1994. *Conceptual issues in evolutionary biology, second edition*.: 363-376. MIT Press, Cambridge]

FELSENSTEIN, J. 1988b. Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics 22: 521-565.

FITCH, W. M. 1971. Towards defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20: 406-416.

FITCH, W. 1981. A non-sequential method for constructing trees and hierarchical classifications. Journal of Molecular Evolution 18: 30-37.

FOSBERG, F. R. and M.-H. SACHET. 1980. Systematic studies of Micronesian plants. Smithonian Contributions to Botany 45: 1-40.

FRIEDMAN, W. E. 1992. Evidence of a pre-angiosperm origin of endosperm. Implications for the evolution of flowering plants. Science 255: 336-339.

FROHNE, D. and U. JENSEN. 1992. Systematik des Pflanzenreichs unter besonderer Berücksichtigung chemischer Merkmale und pflanzlicher Drogen. 4th ed. Stuttgart: Gustav Fisher Verlag.

FUNK, V. A. 1985. Phylogenetic patterns and hybridisation. Annals of the Missouri Botanical Garden 72: 681-715.

GARG, S. 1987. Gentianaceae of the North West Himalaya (a revision). *International Bioscience Monograph* 17: 1-342. Today and Tomorrow's Printers and Publishers, New Delhi.

GAULD, I. and G. UNDERWOOD. 1986. Some applications of the Lequesne compatibility test. Biological Journal of the Linnean Society 29: 191-222.

GHOSAL, S., JAISWAL, D. K., and K. BISWAS. 1978. New glycoxanthones and flavanone glycosides of *Hoppea dichotoma*. Phytochemistry 17: 2119-2123.

GILG, E. 1895. Gentianaceae. In Engler, A. and K. Prantl, eds. *Die natürlichen Pflanzenfamilien IV/2*: 50-108. Verlag von Wilhelm Engelmann, Leipzig.

GOLOBOFF, P. A. 1991a. Homoplasy and the choice among cladograms. Cladistics 7: 215-232.

GOLOBOFF, P. A. 1991b. Random data, homoplasy and information. Cladistics 7: 395-406.

GOLOBOFF, P. A. 1993a. Estimating character weights during tree search. Cladistics 9: 83-91.

GOLOBOFF, P. A. 1993b. NONA version 1.1. Program and documentation distributed by the author. Tucumán, Argentina.

GOLOBOFF, P. A. 1993c. Pee-Wee version 2.1. Program and documentation distributed by the author. Tucumán, Argentina.

GOLOBOFF, P. A. 1995. Parsimony and weighting: a reply to Turner and Zandee. Cladistics 11: 91-104.

GOLOBOFF, P. A. 1996a. SPA. Sankoff parsimony analysis. Version 1.1. Program and documentation distributed by the author. Tucumán, Argentina.

GOLOBOFF, P. A. 1996b. PHAST. Phylogenetic analysis for Sankovian transformations. Version 1.1. Program and documentation distributed by the author. Tucumán, Argentina.

GOTTLIEB, O. R. 1982. Evolution of xanthones in Gentianaceae and Guttiferae. In Gotlieb, O. R., *Micromolecular evolution, systematics and ecology*: 89-93. Springer, Berlin.

GRAHAM, A. 1984. *Lisianthius* pollen from the Eocene of Panama. Annals of the Missouri Botanical Garden 71: 987-993.

GRISEBACH, A. H. R. 1845. Gentianaceae. In de Candolle, A., *Prodromus Systematis Naturalis Regni Vegetabilis IX* : 38-141. Fortin, Masson, Paris.

HALL, B. K., ed. 1994. Homology: the hierarchical basis of comparative biology. Academic Press, San Diego.

HARVEY, A. W. 1992. Three-taxon statements: more precisely, an abuse of parsimony. Cladistics 8: 345-354.

HASSELBERG, G. B. E. 1937. Zur Morphologie des vegetativen Sprosses der Loganiaceen. Symbolae Botanicae Upsalienses 2: 3.

HAUSER, D. L. 1992. Similarity, falsification and character state order - a reply to Wilkinson. Cladistics 8: 339-344.

HEGNAUER, R. 1989. Chemotaxonomie der Pflanzen 8. Basel: Birkhäuser Verlag.

HENDY, M. D. and D. PENNY. 1982. Branch and bound algorithms to determine minimal evolutionary trees. Mathematical Biosciences 59: 277-290.

HENNIG, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutsche Centralverlag, Berlin.

HENNIG, W. 1966. Phylogenetic Systematics. University of Illinois Press, Urbana, Illinois.

HENNIG, W. 1982. Phylogenetische Systematik. Pareys Studientexte 34. P. Parey, Berlin.

HEYWOOD, V. H. 1977. Principles and concepts in the classification of higher taxa. Plant Systematics and Evolution, Supplement 1: 1-12.

HILLIS, D. M. 1987. Molecular versus morphological approaches to systematics. Annual Review of Ecology and Systematics 18: 23-42.

HO, T.-N. and S.-W. LIU. 1990. The infrageneric classification of *Gentiana* (Gentianaceae). Bulletin of the British Museum of Natural History (Botany) 20: 169-192.

HOSTETTMANN, K. and H. WAGNER. 1977. Xanthone glycosides. Phytochemistry 16: 821-829.

HOSTETTMANN-KALDAS, M. and A. JACOT-GUILLARMOD. 1978. Xanthones et C-glucosides flavoniques du genre *Gentiana* (sous-genre *Gentianella*). Phytochemistry 17: 2083-2086.

HOSTETTMANN-KALDAS, M., HOSTETTMANN, K., and O. STICHER. 1981. Xanthones, flavones and secoiridoids of American *Gentiana* species. Phytochemistry 20: 443-446.

HUMBERT, H. 1937. Un genre nouveau de Gentianacées Chironiinées de Madagascar. Bulletin de la Societé Botanique de France 84: 386-90.

HUMPHRIES, C. J. and J. A. CHAPPILL. 1988. Systematics as a science: a response to Cronquist. Botanical Review 54: 129-144.

IGERSHEIM, A., PUFF, C., LEINS, P., and C. ERBAR. 1994. Gynoecial development of *Gaertnera* Lam. and of presumably allied taxa of the Psychotrieae (Rubiaceae): secondaruly "superior" vs. inferior ovaries. Botanische Jahrbucher für Systematik 116: 401-414.

JENSEN, S. R. 1992. Systematic implications of the distribution of iridoids and other chemical compounds in the Loganiaceae and other families of the Asteridae. Annals of the Missouri Botanical Garden 79: 284-302.

KAOUADJI, M. 1990. Flavonol diglycosides from *Blackstonia perfoliata*. Phytochemistry 29, 1345-1347.

KARIS, P. O. 1995. Cladistics of the subtribe Ambrosiinae (Asteraceae: Heliantheae). Systematic Botany 20: 40-54.

KHETWAL, K. S., JOSHI, B, and R. S. BISHT. 1990. Tri- and tetraoxygenated xanthones from *Swertia petiolata*. Phytochemistry 29: 1265-1267.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Systematic Biology 45: 363-374.

KIRIAKOFF, S. G. 1956. Beginselen der dierkundige systematiek voor hoogstudenten en biologen. De Sikkel, Antwerpen.

KIRIAKOFF, S. G. 1960. Filosofische grondslagen van de biologische systematiek. Natuurwetenschappelijk Tijdschrift 42: 35-57.

KITCHING, I. J., 1992. Tree building techniques. In Forey, P. L., Humphries, C. J., Kitching, I. J., Scotland, R. W., Siebert, D. J., and D. M. Williams, eds., *Cladistics. A practical course in systematics*: 44-71. Clarendon Press, Oxford.

KLACKENBERG, J. 1985. The genus *Exacum* (Gentianaceae). Opera Botanica 84: 1-144.

KLACKENBERG, J. 1987. Revision of the genus *Tachiadenus* (Gentianaceae). Adansonia 9: 43-80.

KLUGE, A. G. 1993. Three-taxon transformations in phylogenetic inference: ambiguity and distortion as regards explanatory power. Cladistics 9: 246-259.

KLUGE, A. G. 1994. Moving targets and shell games. Cladistics 10: 403-413.

KLUGE, A. G. and J. S. FARRIS. 1969. Quantitative phyletics and the evolution of Anurans. Systematic Zoology 18: 1-32.

KRISHNA, G. G. and V. PURI. 1962. Morphology of the flower of some Gentianaceae with special reference to placentation. Botanical Gazette: 124, 42-57.

KUBITZKI, K. 1977. Some aspects of the classification and evolution of higher taxa. Plant Systematics and Evolution, Supplement 1: 21-31.

KUHLMANN, J. G. 1925. Contribuçao para a conhecimento de algumas plantas novas, contendo tamben um trabalho de critica ennovas combinçoes. Archivos do Jardim Botânico do Rio de Janeiro 4: 347-365.

KUSNEZOW, N. J. 1896-1904. Subgenus *Eugentiana* Kusnez. generis *Gentiana* Tournef. Acta Horti Petropolitani 15, 1-507.

LADIGES, P. Y., PROBER, S. M., and G. NELSON. 1992. Cladistic and biogeographic analysis of the 'blue ash' Eucalypts. Cladistics 8: 103-124.

LAMMERS, T. G. 1992. Circumscription and phylogeny of the Campanulales. Annals of the Missouri Botanical Garden 79: 388-413.

LE MAOUT, E. and J. DECAISNE. 1868. Traité général de botanique descriptive et analytique. Librairie de Firmin Didot frères, Fils et Cie, Paris.

LE QUESNE, W. J. 1969. A method of selection of characters in numerical taxonomy. Systematic Zoology 18: 201-205.

LE QUESNE, W. J. 1983. The uniquely derived concept as a basis for character compatibility analysis. In Felsenstein, J., ed., *Numerical taxonomy*. NATO ASI series vol. G1: 296-303. Springer-Verlag, Berlin.

LEEUWENBERG, A. J. M., ed. 1980. Die natürlichen Pflanzenfamilien. Band 28 b I. Angiospermae: Ordnung Gentianales Fam Loganiaceae. Duncker and Humblot, Berlin.

LEEUWENBERG, A. J. M. and P. W. LEENHOUTS. 1980. Taxonomy. In Leeuwenberg, A. J. M., ed., *Die natürlichen Pflanzenfamilien. Band 28 b I. Angiospermae: Ordnung Gentianales Fam Loganiaceae*: 8-96. Duncker and Humblot, Berlin.

LIN, C.-N., CHANG, C.-H., ARISAWA, M., SHIMIZU, M., and N. MORITA. 1982a. Two new xanthone glycosides from *Tripterospermum lanceolatum*. Phytochemistry 21: 205-208.

LIN, C.-N., CHANG, C.-H., ARISAWA, M., SHIMIZU, M., and N. MORITA. 1982b. A xanthone glycoside from *Tripterospermum taiwanese* and rutin from *Gentiana flavomaculata*. Phytochemistry 21: 948-949.

LIN, C.-N., CHUNG, M.-I., GAN, K.-H., and J. R. CHIANG. 1987. Xanthones from Formosan gentianaceous plants. Phytochemistry 26: 2381-2384.

LINDSEY, A. A. 1940. Floral anatomy in the Gentianaceae. American Journal of Botany 27: 640-651.

LIPSCOMB, D. L. 1992. Parsimony, homology and the analysis of multistate characters. Cladistics 8: 45-65.

LIU, S.-W. and T.-N. HO. 1992. Systematic study on *Lomatogonium* A. Br. (Gentianaceae). Acta Phytotaxonomica Sinica 30: 289-319.

LUNDBERG, J. G. 1972. Wagner networks and ancestors. Systematic Zoology 21: 398-413.

LUONG, M. D., FOMBASSO, P., and A. JACOT-GUILLARMOD. 1980. Contribution à la phytochimie du genre *Gentiana* XXV. Helvetica Chimica Acta 63: 244-249.

MAAS, P. J. M. 1984a. Systematic studies in neotropical Gentianaceae - The *Lisianthius* complex. Introduction. Acta Botanica Neerlandica 32: 371.

MAAS, P. J. M. 1984b. Systematic studies in neotropical Gentianaceae - The *Lisianthius* complex. Conclusion. Acta Botanica Neerlandica 32: 373-374.

MAAS, P. J. M. and P. RUYTERS. 1986. *Voyria* and *Voyriella* (saprophytic Gentianaceae). Flora Neotropica 41: 1-93.

MABBERLEY, D. J. 1990. The plant book. (ed. 2.) Cambridge: Cambridge University Press.

MADDISON, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Systematic Zoology 40: 315-328.

MADDISON, W. P. and D. R. MADDISON. 1992. MacClade. Analysis of phylogeny and character evolution (version 3). Sinauer Associates Inc., Sunderland, Massachusetts.

MAGUIRE, B. 1981. Gentianaceae. The botany of the Guayana Highland, 11. Memoirs of the New York Botanical Garden 32: 330-388.

MAGUIRE, B. and PIRES J. M. 1978. The botany of the Guyana Highland ,10. Saccifoliaceae. Memoirs of the New York Botanical Garden 29: 230-245.

MARAIS, W. and J. C. VERDOORN. 1963. Gentianaceae. In Dyer, R. A., Codd, L. E., and H. B. Rycroft, eds., *Flora of Southern Africa Vol.26*: 171-243. Botanical Research Institute, Pretoria.

MARQUAND, C. V. B. 1937. The gentians of China. Bulletin of the Miscellaneous Information of the Royal Botanic Gardens, Kew for 1937: 134-180.

MASSIAS, M., CARBONNIER, J, and D. MOLHO. 1977. Xanthones de *Swertia speciosa* Wall. Contribution à chimitaxonomie du genre. Bulletin du Museum National d'Histoire Naturelle, Paris, Sciences Physico-chimiques 13: 55-61.

MASSIAS, M., CARBONNIER, J., and D. MOLHO. 1978. Implications chimiotaxonomiques de la répartition des substances osidiques dans le genre *Gentiana* L. Bulletin du Museum National d'Histoire Naturelle, Paris, 3me série, No. 504: 41-53.

MASSIAS, M., CARBONNIER, J., and D. MOLHO. 1982. Chemotaxonomy of *Gentianopsis*: xanthones, C-glycosyl-flavonoids and carbohydrates. Biochemical Systematics and Ecology 10: 319-327.

MÉSZÁROS, S. 1994. Evolutionary significance of xanthones in Gentianaceae: a reappraisal. Biochemical Systematics and Ecology 22: 85-94.

MÉSZÁROS, S., DE LAET, J., and E. SMETS. 1996. Phylogeny of temperate Gentianaceae: a morphological approach. Systematic Botany 21: 153-168.

METCALFE, C. R. and L. CHALK., ed. 1950. Anatomy of the Diocotyledons. London: Clarendon Press.

MICKEVICH, M. F. and J. S. FARRIS. 1982. Phylogenetic analysis system (PHYSIS) (FORTRAN V software system of cladistic and phenetic algorithms).

MICKEVICH, M. F. and S. J. WELLER. 1990. Evolutionary character analysis: tracing character change on a cladogram. Cladistics 6: 137-170.

MORITZ, C. and D. M. HILLIS. 1996. Molecular systematics: context and controversies. In Hillis D. M. and Moritz C., eds., *Molecular systematics, (second edition)*: 1-10. Sinauer Associates Inc., Sunderland, Massachusetts.

MORRONE, J. J. and J. M. CARPENTER. 1994. In search of a method for cladistic biogeography: an empirical comparison of component analysis, Brooks parsimony analysis, and three-area statements. Cladistics 10: 99-153.

MURATA, J. 1989. A synopsis of *Tripterospermum* (Gentianaceae). Journal of the Faculty of Sciences. Imperial University of Tokyo, Section 3 (Botany) 14: 273-339.

NEFF, N. A. 1986. A rational basis for a priori character weighting. Systematic Zoology 35: 110-123.

NELSON, G. J. 1972. Phylogenetic relationships and classification. Systematic Zoology 21: 227-231.

NELSON, G. 1992. Reply to Harvey. Cladistics 8: 355-360.

NELSON, G. 1993. Reply. Cladistics 9: 261-265.

NELSON, G. 1994. Homology and systematics. In Hall, B. K., ed., *Homology: the hierarchical basis of comparative biology*: 101-149. Academic Press, San Diego.

NELSON, G. 1996. Nullius in verba. Published by the author. 24 pp.

NELSON, G. and P. Y. LADIGES. 1991a. Three-area statements: standard assumptions for biogeographic analysis. Systematic Zoology 40: 470-485.

NELSON, G. and P. LADIGES. 1991b Standard assumptions for biogeographic analysis. Australian Systematic Botany 4: 41-58. (addendum: 5:247).

NELSON, G. and P. LADIGES. 1992. Information content and fractional weight of three-item statements. Systematic Biology 41: 490-494.

NELSON, G. and P. Y. LADIGES. 1993. Missing data and three-item analysis. Cladistics 9: 111-113.

NELSON, G. and P. Y. LADIGES. 1994. Three-item consensus: empirical test of fractional weighting. In Scotland, R. W., Siebert, D. J., and D. M. Williams, eds., *Models in phylogeny reconstruction*: 193-209. Oxford University Press (Systematics Association Special Volume 52), Oxford.

NELSON, G. and P. Y. LADIGES. 1995. TAX: MSDos computer programs for systematics. New York and Melbourne. Published by the authors.

NELSON, G. and P. Y. LADIGES. 1996. Paralogy in cladistic biogeography and analysis of paralogy-free subtrees. American Museum Novitates 3167: 1-58.

NELSON, G. and N. I. PLATNICK. 1991. Three-taxon statements: a more precise use of parsimony. Cladistics 7: 351-366.

NEUBAUER, H. F. 1984. Knotenbau and Blattgrundvaskularisation bei einigen Gentianaceae. Plant Systematics and Evolution 144: 1-7.

NICHOLAS A. and H. BAIJNATH. 1994. A consensus classification for the order Gentianales with additional details on the suborder Apocyneae. Botanical Review. 60: 440-482.

NILSSON, S. 1970. Pollen morphological studies in the Gentianaceae. Acta Universitatis Upsaliensis. Abstracts of Uppsala Dissertationes in Science 165: 1-18.

NISHINO, E. 1983. Corolla tube formation in the Tubiflorae and Gentianales. Botanical Magazine Tokyo 96: 223-243.

NIXON, K. C. and J. I. DAVIS. 1991. Polymorphic taxa, missing values and cladistic analysis. Cladistics 7: 233-241.

NIXON, K. C. and J. M. CARPENTER. 1993. On outgroups. Cladistics 9: 413-426.

OKORIE, D. A. 1976. A new phtalide and xanthones from *Anthocleista djalouensis* and *Anthocleista vogelii*. Phytochemistry 15: 1799-1800.

OLMSTEAD, R. G., BREMER, B., SCOTT, K. M., and J. D. PALMER. 1993. A parsimony analysis of the Asteridae sensu lato based on *rbc*L sequences. Annals of the Missouri Botanical Garden 80: 700-722.

OLMSTEAD, R. G., MICHAELS, H. J., SCOTT, K. M., and J. D. PALMER. 1992. Monophyly of the Asteridae and identification of their major lineages inferred from DNA sequences of *rbc*L. Annals of the Missouri Botanical Garden 79: 249-265.

ORTEGA, E. P., LOPEZ-GARCIA, R. E., RABENAL, R. M., DARIAS, V., and S. VALVERDE. 1988. Two xanthones from *Ixanthus viscosus*. Phytochemistry 27: 1912-1913.

PAGE, R. D. M. 1993. COMPONENT: Tree comparison software for Microsoft Windows, version 2.0. Natural History Museum, London.

PATEL, R. C., J. A. INAMDAR, and N. V. RAO. 1981. Structure and ontogeny of stomata in some Gentianaceae and Menyanthaceae complex. Feddes Repertorium 92: 535-550.

PATTERSON, C. 1982. Morphological characters and homology. In: Joysey, K. A. and A. E. Friday, eds. Problems of phylogenetic reconstruction. The Systematics Association special volume no 21: 21-74. Academic Press, London.

PATTERSON, C. and G. D. JOHNSON. 1995. The intermuscular bones and ligaments of telostean fishes. Smithsonian Contributions to Zoology 559: 1-85.

PATTERSON, C., WILLIAMS, D. M., and C. J. HUMPHRIES. 1993. Congruence between molecular and morphological phylogenies. Annual Review of Ecology and Systematics 24: 153-188.

PENNY, D. and M. D. HENDY. 1985. Testing methods of evolutionary tree construction. Cladistics 1: 266-278.

PENNY, D., HENDY, M. D., and M. A. STEEL. 1992. Progress with methods for constructing evolutionary trees. Trends in Ecology and Evolution 7, 3: 73-79.

PERROT, M. E. 1898. Anatomie comparée des Gentianacées. Annales des Sciences Naturelles. Botanique 7: 105-292.

PIESSCHAERT, F., ROBBRECHT, E., and E. SMETS. 1997. *Dialypetalanthus fuscescens* Kuhlm. (Dialypetalantaceae): the problematic taxonomic position of an Amazonian endemic. Annals of the Missouri Botanical Garden 84: 201-223.

PLATNICK, N. I. 1993. Character optimization and weighting: differences between the standard and three-taxon approaches to phylogenetic inference. Cladistics 9: 267-272.

PLATNICK, N. I., GRISWOLD, C. E., and J. A. CODDINGTON 1991. On missing entries in cladistic analysis. Cladistics 7:337-343.

PRINGLE, J. S. 1978. Sectional and subgeneric names in *Gentiana* (Gentianaceae). Sida 7: 232-247.

PRUDNIKOV, A. P., BRYCHKOV, Y. A., and O. I. MARICHEV. 1988. Integrals and series. Gordon and Breach Science Publishers, New York. (translated from the Russian by N. M. QUEEN).

REECK, G. R., DE HAËN, C., TELLER, D. C., DOOLITTLE, R. F., FITCH, W. M., DICKERSON, R. E., CHAMBON, P., MCLACHLAN, A. D., MARGOLIASH, E., JUKES, T. H., and E. ZUCKERKANDL. 1987. Homology in proteins and nucleic acids: a terminology muddle and a way out of it. Cell 50: 667.

REMANE, A. 1952. Die Grundlagen des natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik. Akademische Verlagsgesellschaft, Leipzig.

REZENDE, C. M. A. M. and O. R. GOTTLIEB. 1973. Xanthones as systematic markers. Biochemical Systematics and Ecology 10: 111-118.

RIEPPEL, O. C. 1988. Fundamentals of comparative biology. Birkhäuser Verlag, Basel.

RIZZINI, C. T. and O. OCCHIONI. 1949. 'Dialypetalanthaceae'. Lilloa 17: 243-288.

ROBBRECHT, E. 1993a. Introduction. In Robbrecht E., ed., *Advances in Rubiaceae macrosystematics.* Opera Botanica Belgica 6: 7-18.

ROBBRECHT, E. 1993b. On the delimitation of the Rubiaceae. A review. In Robbrecht E., ed., *Advances in Rubiaceae macrosystematics.* Opera Botanica Belgica 6: 19-30.

RODRIGO, A. G. 1989. An information-rich character weighting procedure for parsimony analysis. New Zealand Natural Sciences 16: 97-103.

RODRIGO, A. G. 1992. Two optimality criteria for selecting subsets of most parsimonious trees. Systematic Biology 41: 33-40.

ROITMAN, J. N., WOLLENWEBER, E., and F. J. ARRIAGA-GINER. 1992. Xanthones and triterpene acids as leaf exudate constituents in *Orphium frutescens*. Journal of Plant Physiology 139: 632-634.

SANG, T. 1995. New measurements of distribution of homoplasy and reliability of parsimonious cladograms. Taxon 44: 77-82.

SANKOFF, D. 1975. Minimal mutation trees of sequences. SIAM Journal of applied Mathematics 28: 35-42.

SANKOFF, D. and R. J. CEDERGREN. 1983. Simultaneous comparison of three or more sequences related by a tree. In Sankoff, D. and J. B. Kruskall, eds., *Time warps, string edits and macromolecules: the theory and practice of sequence comparison*: 259-263. Addison-Wesley, Reading.

SATTATH, S. and A. TVERSKY. 1977. Additive similarity trees. Psychometrika 42: 319-345.

SCHARFETTER, R. 1953. Biographien von Pflanzensippen. Wien.

SCHOCKAERT, E. 1992. Moderne trends in de systematiek. Jaarboek V.O.B.: 105-112. De Sikkel, Antwerpen.

SCOTLAND, R. W. 1992. Cladistic theory. In Forey, P. L., Humphries, C. J., Kitching, I. J., Scotland, R. W., Siebert, D. J., and D. M. Williams, eds., *Cladistics, a practical course in systematics*: 3-13. Clarendon Press, Oxford.

SHARKEY, M. J. 1989. A hypothesis-independent method of character weighting for cladistic analysis. Cladistics 5: 63-86.

SHARKEY, M. J. 1993. Exact indices, criteria to select from minimum length trees. Cladistics 9: 211-222.

SHARKEY, M. J. 1994. Discriminate compatibility measures and the reduction routine. Systematic Biology 43: 526-542.

SIMON, C., FRATI, F., BECKENBACH, A., CRESPI, B. LIU, H., and P. FLOOK. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Annals of the Entomological Society of America 87: 651-701.

SLUIS, W. G. 1985. Secoiridoids and xanthones in the genus *Centaurium* Hill (Gentianaceae): a pharmacognostical study. Drukkerij Elinkwijk, Utrecht

SMETS, E. 1988a. Florale nektariën van de Magnoliophytina: karakterizering en systematische betekenis. Proefschrift, Instituut voor Plantkunde, K.U.Leuven.

SMETS, E. 1988b. La présence des "nectaria persistentia" chez les Magnoliophytina (Angiospermes). Candollea 43: 709-716.

SMETS, E. 1989. The distribution and the systematic relevance of caducous nectaries and persistent nectaries in the Magnoliophytina (angiosperms). Acta Botanica Neerlandica 38: 100.

SMETS, E. F. and E. M. CRESENS 1988. Types of floral nectaries and the concepts "character" and "character-state" - a reconsideration. Acta Botanica Neerlandica 37: 121-128.

SMITH, H. 1965. Notes on Gentianaceae. Notes from the Royal Botanic Garden Edinburgh 26: 237-258.

SOLTIS, D. E., CHASE, M. W., and R. G. OLMSTEAD. 1993. Introduction. Annals of the Missouri Botanical Garden 80: 526-527.

STEVENS, P. F. 1980. Evolutionary polarity of character states. Annual Review of Ecology and Systematics 11: 333-358.

STEVENS, P. F. 1986. Evolutionary classification in botany, 1960-1985. Journal of the Arnold Arboretum 67: 313-339.

STEWART, C.-B. 1993. The powers and pitfalls of parsimony. Nature 361:603-607.

STOUT, G. H., REID, B. J., and G. D. BRECK. 1969. The xanthones of *Macrocarpaea glabra*. Phytochemistry 8: 1427.

STRUWE, L. , ALBERT, V. A., and B. BREMER. 1994. Cladistics and family level classification of the Gentianales. Cladistics 10: 175-206.

STUESSY, T. F. 1990. Plant taxonomy. Columbia University Press, New York.

SULLIVAN, G., STILES, F. D., and K. A. ROSLER. 1977. Phytochemical investigations of xanthones of *Eustoma grandiflorum* (Raf.) Shinners. Journal of Pharmaceutical Sciences: 66: 828-831.

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

SWOFFORD, D. L. and W. P. MADDISON. 1987. Reconstructing ancestral states under Wagner parsimony. *Mathematical Biosciences* 87: 199-229.

SWOFFORD, D. L. and W. P. MADDISON. 1992. Parsimony, character-state reconstructions, and evolutionary inferences. In Mayden, R. L., ed., *Systematics, historical ecology, and North American freshwater fishes*: 186-223. Stanford University Press, Stanford.

SWOFFORD, D. L., OLSEN, G. J., WADELL, P. J., and D. M. HILLIS. 1996. Phylogenetic inference. In Hillis, D. M., Moritz, C. and B. K. Mable, eds., *Molecular systematics, second edition*: 407-514. Sinauer Associates Inc., Sunderland, Massachusetts.

SYTSMA, K. 1988. Taxonomic revision of the Central American *Lisianthius skinneri* species complex (Gentianaceae). Annals of the Missouri Botanical Garden 75: 1587-1602.

SYTSMA, K. J. 1990. DNA and morphology: inference of plant phylogeny. Trends in Ecology and Evolution 5: 104-110.

SYTSMA, K. and B. A. SCHAAL. 1985. Phylogenetics of the *Lisianthius skinneri* (Gentianaceae) species complex in Panama utilizing DNA restriction fragment analysis. Evolution 39: 594-608.

SZUMIK, C. A. 1996. The higher classification of the Order Embioptera: a cladistic analysis. Cladistics 12: 41-64.

TAKHTAJAN, A. L. 1980. Outline of the classification of flowering plants (Magnoliophyta). Botanical Review 46: 225-359.

TAYLOR, C. M. 1995. First international conference on the Rubiaceae: introduction. Annals of the Missouri Botanical Garden 82: 355-356.

THEISEN, I. and W. BARTHLOTT. 1994. Mikromorphologie der epicuticularwachse und die Systematik de Gentianales, Rubiales, Dipsacales und Calycerales. Tropische und subtropische Pflanzenwelt 89: 1-62.

THORNE, R. F. 1976. A phylogenetic classification of the Angiospermae. Evolutionary Biology 9: 35-106.

THORNE, R. F. 1992a. Classification and geography of the flowering plants. Botanical Review 58: 225-348.

THORNE, R. F. 1992b. An updated phylogenetic classification of the flowering plants. Aliso 13: 365-389.

TOYOKUNI, H. 1963. Conspectus Gentianacearum Japonicarum. Journal of the Faculty of Science, Hokkaido University Series 5, 7: 137-259.

TOYOKUNI, H. 1965. Systema Gentianinarum novissimum. Facts and speculation relating to the phylogeny of *Gentiana*, sensu lato and related genera. Symbolae Asahikawensis 1: 147-158.

TURNER, H. 1995. Cladistic and biogeographic analyses of *Arytera* Blume and *Mischarytera* gen. nov. (Sapindaceae) with notes on methodology and a full taxonomic revision. Blumea Supplement 9: 1-230.

TURNER, H. and R. ZANDEE. 1995. The behaviour of Goloboff's tree fitness measure F. Cladistics 11: 57-72.

UDOVICIC, F., MCFADDEN, G. I., and P. Y. LADIGES. 1995. Phylogeny of *Eucalyptus* and *Angophora* based on 5s rDNA spacer sequence data. Molecular Phylogenetics and Evolution 4: 247-256.

WAGENITZ, G. 1959. Die systematische Stellung der Rubiaceae. Ein Beitrag zum System der Sympetalen. Botanische Jahrbucher für Systematik 79: 17-35.

WAGENITZ, G. 1964. Gentianales and Dipscales. In Melchior H., ed., *A. Engler's Syllabus der Pflanzenfamilien, vol. 2 (ed. 12)*: 405-424, 472-478. Verlag von Wilhelm Engelmann, Leipzig.

WAGENITZ, G. 1977. New aspects of the systematics of Asteridae. Plant Systematics and Evolution, Supplement 1: 375-395.

WAGENITZ, G. 1992. The Asteridae: evolution of a concept and its present status. Annals of the Missouri Botanical Garden 79: 209-217.

WAGNER, W. H. 1961. Problems in the classification of ferns. Recent Advances in Botany 1: 841-844.

WATERMAN, M. S. 1995. Introduction to computational biology. Maps, sequences and genomes. Chapman and Hall, London.

WEAVER, R. E., J. 1972. A revision of the Neotropical genus *Lisianthius* (Gentianaceae). Journal of the Arnold Arboretum 53: 76-100 and 234-311.

WEAVER, R. E., J. 1974. The reduction of *Rusbyanthus* and the tribe Rusbyantheae (Gentianaceae). Journal of the Arnold Arboretum 55: 300-302.

WEINBERG, S. 1997. The first elementary particle. Nature 386: 213-215.

WESTON, P. H. 1988. Indirect and direct methods in systematics. In Humphries, C. J., ed., *Ontogeny and systematics*: 27-56. British Museum (natural history), London.

WEYNANTS, C. 1993. Studie van de sytematische verwantschappen binnen de Primulanae door middel van een cladistische analyse. Eindverhandeling, Instituut voor Plantkunde, K.U.Leuven.

WHEELER, W. 1993. The triangle inequality and character analysis. Molecular Biology and Evolution 10: 707-712.

WILEY, E. O. 1979. The annotated Linnean hierarchy, with comments on natural taxa and competing systems. Systematic Zoology 28: 308-337.

WILKINSON, M. 1994a. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. Systematic Biology 43: 343-368.

WILKINSON, M. 1994b. Three-taxon statements: when is a parsimony analysis also a clique analysis? Cladistics 10: 221-223.

WILKINSON, M. 1994c. Weights and ranks in numerical phylogenetics. Cladistics 10: 321-329.

WILKINSON, M. 1995. Arbitrary resolutions, missing entries, and the problem of zero-length branches in parsimony analysis. Systematic Biology 44, 1: 108-111.

WILLIAMS, P. L. and W. M. FITCH. 1989. Finding the minimal change in a given tree. In Fernholm, B., Bremer, K. and H. Jörnvall, eds., *The hierarchy of life. Proceedings from Nobel Symposium*. Excerpta Medica 70: 453-470, Elsevier Science Publ., Amsterdam.

WILLIAMS, P. L. and W. M. FITCH. 1990. Phylogeny determination using dynamically weighted parsimony method. Methods in Enzymology 183: 615-626.

WILLIAMS, W. T. and M. B. DALE. 1964. An objective method of weighting in similarity analysis. Nature 201: 426.

WOLFENDER, J.-L. and K. HOSTETTMANN. 1992. Search for xanthones in *Chironia* species: a rapid method for the screening of crude extracts. Planta medica 58 (Suppl. 7): A673-674.

WOLFENDER, J.-L., HAMBURGER, M., MSONTHI, J. D., and K. HOSTETTMANN. 1991. Xanthones from *Chironia krebsii*. Phytochemistry 30: 3625-3630.

WOOD, C. E., J. and R. E. WEAVER J. 1982. The genera of Gentianaceae in the southeastern United States. Journal of the Arnold Arboretum 63: 441-487.

YEATES, D. 1992. Why remove autapomorphies? Cladistics 8: 387-389.

YUAN, Y.-M. and P. KÜPFER. 1993. Karyological studies on *Gentiana* sect. *Frigida* s.l. and *Stenogyne* (Gentianaceae) from China. Bulletin de la Societé Neuchâteloise des Sciences Naturelles 116: 65-78.

ZELTNER, L. 1970. Recherches de biosystématique sur les genres *Blackstonia* Huds. et *Centaurium* Hill. (Gentianacées). Bulletin de la Societé Neuchâteloise des Sciences Naturelles 93: 1-164.

ZURAWSKI, G. and M. T. CLEGG. 1993. *rbc*L sequence data and phylogenetic reconstruction in seed plants: foreword. Annals of the Missouri Botanical Garden 80: 523-528.

ZUYEV, V. V. 1990. On the systematics of the Gentianaceae family in Siberia (in Russian). Botanicheskii Zhurnal 75: 1296-1305.

A. VITA, A PROGRAM FOR PARSIMONY ANALYSIS USING IMPLIED WEIGHTS

version 0.9c (310197)
for use with 286 personal computers or higher and MS-DOS version 6.0 or higher

A.1 Introduction

ViTA is a DOS-program for parsimony analysis in which two weighting schemes proposed in chapter 3 are available (in the following, 'weighting' without qualification refers to implied weights, and not to a priori weights): lowest weighted homoplasy (direct weighting) and highest weighted similarity, which is presented as a mimimized homoplasy (complex weighting; cf. 3.4.6, p. 94). To allow direct comparison with existing methods, standard parsimony analysis and parsimony analysis using Goloboff fits are provided as well (see fig. 1 for a summary; see appendix B and 3.4.3, p. 88, for the equivalence of direct weighing and Goloboff fits)[10].

NO WEIGHTING	$\sum_{ch=1}^{N} apw_{ch} * h_{ch}$
GOLOBOFF FIT	$\sum_{ch=1}^{N} apw_{ch} * \left[(1-ad_{ch}) * \frac{K * mi_{ch}}{K + h_{ch}} + ad_{ch} * \sum_{tr=1}^{m_{ch}} \frac{K * mi_{ch_{tr}}}{K + h_{ch_{tr}}} \right]$
DIRECT WEIGHTING	$\sum_{ch=1}^{N} apw_{ch} * \left[(1-ad_{ch}) * \frac{K * h_{ch}}{K + h_{ch}} + ad_{ch} * \sum_{tr=1}^{m_{ch}} \frac{K * h_{ch_{tr}}}{K + h_{chtr}} \right]$
COMPLEX WEIGHTING	$\sum_{ch=1}^{N} apw_{ch} * \left[(1-ad_{ch}) * \frac{(K + g_{ch} - m_{ch}) * h_{ch}}{h_{ch} + K} + ad_{ch} * \sum_{tr=1}^{m_{ch}} \frac{(K + g_{ch} - m_{ch}) * h_{ch_{tr}}}{h_{ch_{tr}} + K} \right]$

Fig. 1. Optimality functions available in ViTA. Optimization is over all cladograms for the taxa in the data set. Apw_{ch} is the a priori weight of character ch (1<=ch<=N, with N the number of characters in the data set; inactive characters are treated as if their apw is 0); h_{ch} is the homoplasy of a character; m_{ch} and g_{ch} are the minimum and maximum number of steps of character ch on any possible cladogram; ad_{ch} is 0 for unordered and 1 for ordered characters (linear character state trees according to numerical state codes); mi_{ch} is 0 if g_{ch}=m_{ch}, otherwise it is 1; for ordered characters, subscript tr refers to a branch of the character state tree; K is the concavity constant; in the weighted analyses, the value of the optimality function is rounded after summation to two signifcant digits following the decimal point .

[10] A similar program, ViTA2, optimizes the total number of accomodated four-item statements; cf. chapter 2.

In this version, only the basic functions that are necessary to obtain a practically useful system are provided. Features such as calculation of consensus trees or optimization of polytomies will be added in future versions. The interface of ViTA is based on the interface of Hennig86 (Farris 1988), which is also used in NONA and Pee-Wee (Goloboff 1993b, 1993c). Likewise, most of the commands in ViTA are based on commands available in these programs. Users that are acquainted with any of these programs should have little problems in using ViTA. The program is written in Pascal, which was choosen primarily because of its clarity and simplicity. The program has been tested on PC's running MS-DOS 4.0, 6.0, and 6.1, but no difficulties are expected with other versions. The current version of the program uses only conventional RAM. The conventional RAM that is still available after a data set has been read is used as a buffer for storing trees. When the tree buffer is not empty, one of its trees is called the 'current' tree, to which a number of commands refer. The program is distributed as a single executable file, *vita.exe*. As with other programs, ViTA will be easiest to use when the directory that contains *vita.exe* is placed in the DOS search path (cf. the PATH command in the c:\autoexec.bat file).

A.2 Defining the optimality criterion

In the default mode, all characters are treated as unordered, active, and with equal a priori weights; the tree-finding commands search for the most parsimonious trees without applying implied weights. These default settings can be changed using the following commands:

- SET WEMODE: switch between the three available weigthing modes that use implied weights, or switch off the use of implied weights (most commands that deal with tree statistics accept options that select any of these modes directly, without having to change the default).
- SET CCODE: set the character additivities, activities, and a priori weights
- SET K: set the concavity constant for the implied weights
- SET DEVIATION: retain also suboptimal trees, to a specified degree
- SET SEMODE: switch between searching for best trees (SEMODE on) and searching for worst trees (SEMODE off); note that the **best** trees are those with **highest** fit for Goloboff weighting, but the **lowest** homoplasy in all other cases; the **worst** trees are those with **lowest** fit for Goloboff weighting, but the **highest** homoplasy otherwise; a tree that is **optimal** can be either a worst or a best tree, depending on the status of SEMODE.

In the current version, the a priori weights are set for full characters, but for the additive characters in the standard approach they might as well be set for each state transformation individually (as is effectively possible when using additive binary coding). These possibilities will be added in future versions.

A.3 Tree finding algorithms

Tree finding algorithms are often divided into exact algorithms and heuristics algorithms (e.g. Swofford et al. 1996, Darlu & Tassy 1993). Exact algorithms are algorithms that guarantee that all optimal trees will be found. They can do so either by checking explicitly every possible tree (implicit enumeration) or by traversing the search space of all possible trees in such a way that subspaces can be safely skipped without loosing optimal trees (branch and bound algorithms; cf. Hendy & Penny 1982). The problematic point with these exact algorithms is their execution time: the number of unrooted dichotomous trees for n terminal taxa equals $\Pi_{i=3..n}(2i-5)$ (e.g. Waterman 1995:346), and as the number of taxa increases, the number of trees soon becomes so large that they cannot be all enumerated and evaluated in a reasonable time (e.g. for 5 taxa there are 15 trees, for 10 taxa 2 027 025 and for 15 already over 10^{12}). With this respect, branch and bound methods perform better than implicit enumeration, but even then the execution time soon becomes a limiting factor as the number of taxa increases. Depending “upon the efficiency of the algorithm used, the amount of homoplasy in the data, computer speed and patience of the analyst” branch-and-bound methods are feasible only up to about 25 taxa (Kitching 1992: 65).

This execution-time limitation is the reason why the so-called heuristic algorithms have been developed: even though they do not guarantee that all optimal trees are found (or even that the the resulting trees are optimal at all), they run faster, and may provide good results when the exact algorithms are useless from a practical point of view. In general, these algorithms start from one or more initial trees and then proceed by “hill-climbing” towards better trees: the branches of the initial trees are rearranged (“swapped”) in the hope that better trees will be found; when this is the case, these better trees are in turn used as starting points for a new swapping round. The procedure goes on untill no better trees are obtained. As a rule, the more rearrangements tried for each tree, the higher the chance of finding all optimal trees, but the slower the algorithm. Besides the kind of rearrangements that are considered during swapping, numerous other details ultimately define the strength of a full heuristic algorithm (e.g. how are the initial tree(s) obtained; how many trees are kept during one swapping round; are suboptimal trees retained during swapping, ...).

From the above, it is easy to see why branch swapping may not be able to find the optimal trees: whenever the optimal trees are several rearrangements away from the starting tree, the possibility exists that one of these rearrangements is worse than the trees already considered, and the algorithm will stop there, having found a local rather than a global optimum. Similarly, the fact that the algorithm finds one optimal tree, does not imply that it will find all optimal trees: there may be optimal trees that are several rearrangements away, and these rearrangements may be far from optimal. This kind of problem is not restricted to parsimony analysis, but is encountered in many optimization procedures; in parsimony analysis, it is known as the problem of multiple islands of most parsimonious trees (Maddison 1991).

As has been stressed by Swofford et al. (1996) and Penny et al. (1992), the problem of evaluating a single tree under a particular optimality criterion and the problem of finding optimal trees according to that criterion are two separate problems. Therefore, the same tree finding algorithms can be used for the weighted as well as for unweighted analyses (but see below). At present, two different but closely related heuristic algorithms are available in ViTA: MULT and MAX (variants of the same instructions in NONA and Pee-Wee; Goloboff 1993b, 1993c). Both do branch-swapping by subtree pruning- regrafting (SPR), a swapping strategy that is relatively fast but only of moderate strength (Swofford et al. 1996). In subtree pruning-regrafting, the rearrangements tried are those that can be achieved by pruning any subtree from the starting tree and reattaching it to any branch of the remaining part of the tree (note that the set of subtrees depends on the current outgroup: when the tree is cut in two parts, the part that does not contain the current outgroup taxon is considered the subtree). The number of rearrangements for each tree is higher than in nearest-neighbor interchange (NNI), but lower than in tree bissection and reconnection (TBR), two other widely used swapping algorithms (Swofford et al. 1996).

MULT creates its own starting trees by means of stepwise addition (multiple random addition is supported), while MAX looks for starting trees in the tree buffer. The other most important difference is that MULT retains only one tree during swapping, while MAX retains all best trees it has found so far. More details are given below.

A.4 Restrictions on data set size

In the current version, the number of characters must be in the range 1-200, the number of taxa in the range 4-50, and the maximum number of states per character is restricted to four. The main reasons for these restrictions are the limited power of the current tree finding algorithms and the current implementation of memory use. These restrictions will be relaxed in future versions.

A.4.1 Tree finding capacities

As discussed higher, the tree-finding possibilities are restricted to two variants of branch swapping by means of subtree pruning-regrafting (MULT and MAX), a search strategy that is relatively fast but not very strong. Until more powerful search strategies are available, it does not seem very useful to allow larger data sets, as they cannot be properly processed.

ViTA's current branch swapping capacities have been evaluated using some published matrices and comparing the results with Hennig86 and Pee-Wee. For up to 25 taxa, ViTA's command sequence "MULT25; MAX" mostly succeeds in finding the same trees as Hennig86' "mh;bb*" (ViTA's WEMODE switched off, and taking into account that Hennig86 has a different strategy for collapsing zero-length branches).

As discussed by Goloboff (1993c), the landscape of islands (Maddison 1991) can be very different when different optimality criteria are used. The fact that a particular algorithm performs well in the unweighted standard approach is no guarantee that it will work equally well when using implied weighting. Therefore comparisons with Pee-Wee were made also. With SET WEMODE switched to Goloboff weighting, and the concavity set as in Pee-Wee, ViTA's command sequence "MULT25; MAX" mostly finds the same trees as Pee-Wee's "mult*25;max*" for data sets up to 25 taxa (differences may be due also to the fact that Pee-Wee truncates before summation, while ViTA rounds after summation; see below).

It remains to be determined how efficient the available tree finding algorithms are with direct and complex weighting and/or with data sets that have more than 25 taxa.

A.4.2 Memory use

Memory is currently assigned in a rather inflexible way: it is assumed that all data sets always have the maximum number of characters, and all characters the maximum number of states. In this way, a lot of unused memory may become unnecessarily blocked and as a result the number of trees that can be kept in memory may be sharply reduced, especially for small data sets with mainly binary characters. More flexible and optimized memory use will overcome these problems in future versions.

A.5 Suppression of zero-length branches

Branches can be collapsed according to the rules that are also available in NONA and Pee-Wee (only active non-zero weight characters are used to assess branch support): according to the first rule (SET AMBIGUOUS-), a branch is supported when it is assigned non-zero length by all most parsimonious reconstructions (MPR's, Swofford & Maddison 1987) of at least one character; according to the second rule (SET AMBIGUOUS=), a branch is supported when at least one character has at least one MPR that assigns non-zero length to the branch (cf. Coddington & Scharff 1994 and Wilkinson 1995 for a general discussion of the problem of zero-length branches).

As in NONA and Pee-Wee, the trees in the tree buffer are always completely dichotomous, and the unsupported branches are suppressed only when the trees are shown. Reported tree lengths ALWAYS refer to the dichotomous representation. This implies that, as in Pee-Wee and NONA, collapsed trees may be worse than their dichotomous representation in memory under SET AMBIGUOUS-. Such overcollapsed trees can be removed automatically from the tree buffer using the command EXACT (as an alternative, they could be rejected immediately during tree search, but as they might serve as stepping stones between islands, it is better to retain them).

A.6 Polymorphisms

Data sets in which polymorphisms are indicated between square brackets (as in NONA and Pee-Wee) are properly read, but the polymorphisms are automatically converted to missing entries in the current version. In the case of DNA and RNA data (DREAD), the one-letter IUPAC codes for polymorphisms are not yet recognized and will produce an error.

A.7 Precision of weighted parsimony scores

In all weighted analyses, the optimality score of a single character (or character state transformation for ordered multistate characters) is a rational value, and not an integer. This raises the question of how precise or fine-grained the weighting should be (Goloboff 1993a). As an example, should a tree that has an ensemble Goloboff fit of only 100.000 be considered really worse than a tree with a global Goloboff fit of 100.001? In order to prevent that the weighting becomes sensitive to such minute differences, NONA and Pee-Wee retain only two positions following the decimal point, which results in fit 100.00 for both trees (which appears as fit 1000.0 in the output of NONA and Pee-Wee because these programs scale all character fits to a maximum of 10). When an increased precision is wanted, this can be obtained by increasing the a priori weigths (e.g. giving all characters an a priori weight of 10 amounts to retaining three decimal positions in stead of two).

ViTA essentially provides the same solution, but there are two slight differences: while NONA and Pee-Wee truncate each character fit to two decimal positions before summation over all characters, ViTA does the summation first, and then rounds the ensemble value. A lower precision can be obtained by considering also suboptimal trees (SET DEVIATION), while an increased precision can be obtained by increasing the a priori weights.

The difference between truncating and rounding is basically a shift in the position of the fit or length intervals. This shift leaves the precision unaffected, and the difference between truncating and rounding is merely the difference between two conventions of indicating intervals; e.g. a rounded fit of 100.00 means a fit between 99.50000... and 100.49999..., while a truncated fit of 100.00 means a fit between 100.00000... and 100.99999... The fit intervals are equally large in both cases, only the starting point differs.

While the choice between truncating and rounding is essentially arbitrary, the situation is different when it comes to choosing between dropping decimal positions before or after summation. Consider e.g. a data set having seven characters, and the concavity constant K set to 6. On a first tree there are two characters having six steps of homoplasy (each with fit K/(K+h) = 6/12 = 0.5), while the other characters are free of homplasy (fit 1); on a second tree, each of the seven characters has one step of homoplasy (i.e. character fits equal to 6/7 = 0.857142...). Rounding (or truncating) to two decimal positions after summation correctly gives an ensemble fit of 6.00 for both trees. Rounding and truncating before summation, however, both still give a fit of 6.00 for the first tree, but fits of 6.02 and 5.95 respectively for the second one. The deviations from 6.00 follow directly from the fact that decimal positions are dropped before summation. Moreover, these errors are not random but systematic; e.g. in the case of truncating to two decimal positions, the error equals [100K mod (K+h)], and as [100K mod (K+h)] increases, the fit of a character will be increasingly more underestimated. This implies that from two trees that are equally fit, the one that minimizes the sum of [100K mod (K+h)] over all characters will be preferred. In order to avoid this systematic distortion, dropping decimal positions should be done only after summation.

A.8 Uninformative characters and Goloboff fit

In Pee-Wee, only informative characters are used to calculate fit. In ViTA a distinction is made between uninformative characters (uninformative character state transformations for additive characters) that have an observed variation $m > 0$ on the one hand (autapomorphies), and uninformative characters having $m = 0$ on the other hand (symplesiomorphies). A character having $m>0$ is considered to have a fit equal to 1 (or whatever its a priori weight is), while a character having $m=0$ is not considered at all (i.e. it is assigned fit 0). This distinction is based on and motivated by the influence of both types of uninformative character on the ensemble consistency index $C = M/S$ (Kluge & Farris 1969): symplesiomorphies do not influence C, while autapomorphies do (e.g. Yeates 1992). Analogous to the calculation of C, the fit of a data set to a tree can be expressed as the ratio between the maximum fit of the data set (no homoplasy) and the fit that is observed on the tree. When autapomorphies are given fit 1 and symplesiomorphies fit 0, the influence of the uninformative characters on this ratio is the same as their influence on the consistency index.

It is of course possible to calculate C with autapomorphies excluded. This corresponds to giving the autapomorphies a zero-fit, which is the Pee-Wee's option. ViTA does not provide automatic calculation of fits with uninformative characters excluded, but XLENGTH, the command that shows character fits, automatically flags uninformative characters and uninformative character state transformations. This permits an easy correction of the calculated ensemble fits if exclusion of autapomorphies is considered necessary .

Remark: when comparing Goloboff fits reported by Pee-Wee and ViTA, one should be aware of still a third difference besides the differences reported in this and the preceding section: ViTA automatically decomposes additive multistate characters into their constituent binary additive ‘sub-characters’, which are all weighted seperately (see chapter 3). Pee-Wee, on the other hand, lumps the homoplasies of the different transformations to weigh the original undivided character.

A.9 File input and output

By default, input is from the keyboard, and output is to the screen. These defaults can be overruled with the commands SET PROCEDURE *fname* (input from file *fname*), SET LOG *fname* (all following output to file *fname*) and TSAVE *fname* (save all trees currently in the tree buffer to file *fname*). For SET PROCEDURE, *fname* must be the name of an existing filename, possibly preceded by a valid

DOS-path. When the path is not specified, the file is expected to be in the current directory. The input files must be organized in lines of maximally 255 characters long. This restriction on line length is a difference with a.o. NONA and Pee-Wee, and may cause problems when tsave files from these programs are directly read by ViTA.

TSAVE and SET LOG create a file with name *fname* in the current directory if no path is specified, or in the specified path otherwise. When file *fname* already exists, it is overwritten (when SET CONFIRM is on, the user is prompted for confirmation to overwrite an existing file). Output from TSAVE and SET LOG commands is by default organized in lines of maximally 80 characters (but see SET SWIDTH).

Trapping of input/output errors is rather primitive, but hopefully robust. The specified files may be located on any drive, but when a drive other than C: is used, the input/output actions require supplementary safety tests (e.g. disk drive ready) that slow down the input/output process. The drive-ready test creates (or opens) a file *vita.tmp* in the root directory of the drive being used. This file is not deleted by the program, so it will still be found there after ViTA has been closed.

A.10 Commands

In this section, ViTA's commands are described in detail, grouped by topic. The section is closed with a summary. In the command descriptions, the **KEywords** are put in bold, with the minimum truncation in capitals. Following the keywords there may be arguments or options or both. Arguments mostly provide information that is necessary to perform the default functioning of the command, while options are used to change this default. To indicate that the options (or sometimes also the arguments; cf. the SET commands) are optional, they are put between square brackets.

Options are mostly single characters (indicated between 'q'uotes in the following descriptions; not case-sensitive) or numbers (indicated as N in the descriptions). Numbers or single-character options may but must not be separated by blanks or linefeeds from each other or from the command keyword. Filenames must always be preceded and followed by a blank or a linefeed.

In general, different commands must be separated by a semicolon, but commands that do not accept options may also be terminated by a line feed (or pressing enter); commands that accept options are only performed after the command is closed using a semicolon. When more than one option can be specified, a fixed system of option priorities is followed whenever conflicting options are requested. The order in which different options are entered is immaterial, but the order of options vs. arguments and the order of different arguments are fixed and cannot be changed.

As in Hennig86, some commands take scopelists as arguments. Scopelists are lists of scopes, separated by blanks or linefeeds. In general, a single scope is of the form "N.M", with N and M two natural numbers and N<=M. The minimum and maximum values for M and M depend on the command; e.g. in KEEPTREES, N and M must be between 1 and the number of trees currently in the tree buffer. When N equals the minimum, it can be dropped, and when M equals the maximum it can be dropped also. When N equals M, it suffices to enter N.

Some examples to illustrate the concept of scopes and scopelists:

Keeptrees 1 5: keep trees 1 and 5
Keeptrees 1.5 8.10: keep trees 1 to 5 and 8 to 10
Keeptrees .25: keep the first 25 trees
Keeptrees .: keep all trees

A.10.1 Reading data

A.10.1.1 Taxon and character data

XRead [title] nchar ntaxa data_set_description

XREAD echoes the title to the output, reads the number of characters (nchar), the number of terminal taxa (ntaxa) and the data set. After reading the data into memory, XREAD reports the maximal number of trees that can be stored in the remaining memory (in the current implementation this number depends only on the number of taxa; only conventional RAM is used).

The title is optional and must be enclosed in single quotes; the title itself may not contain quotes. Nchar must be in the range 1-200, ntaxa in the range 4-50, and the number of states per character is restricted to four.

The data must be described taxon per taxon (interleaved input is not supported). The description of the character states of one taxon must start with the name of the taxon, followed by at least one blank or linefeed (indicating the end of the name), followed by the character states. The name of a taxon must start with an alphabetical character and may not contain blanks. The specified name may be of any length that fits on one line, but only the first twelve characters are retained and used for output.

The codes for the character states must be integers in the range 0-3. Polymorphisms must be specified by using the square brackets '[' and ']' (single states may also be enclosed in square brackets). Missing entries (because of inapplicable or unknown states) may be represented by either ‘-’ or ‘?’. The character codes may be separated by blanks or linefeeds.

When a data set is read successfully, all characters are made active, nonadditive, and are assigned an a priori weight of one (these default settings can be changed with the CCODE command). SET AMBIGUOUS, SET WEMODE, SET SEMODE, SET K, SET OUTGROUP, SET RSEED, and SET DEVIATION are reset to their defaults. As in Hennig86, all information and results from a previous data set are lost.

DRead [title] nchar ntaxa data_set_description

As XREAD, but with (limited) special provisions for DNA and RNA data. The codes for character states must be A, C, G, T or U, either in upper- or lower-case (output in upper-case). U is read as T and will appear as such in the output.

Gaps may be specified as either ‘-’ or ‘?’. In the current version they are treated as missing values. IUPAC one-letter codes for polymorphisms or ambiguities are not yet supported, so these must be explicitly specified by using the square brackets ‘[‘ and ‘]’.

When data are read using the DREAD command, the characters can not be made additive in a later CCODE statement.

A.10.1.2 Trees

TRead [title] tree_description_part

TREAD echoes the title to the output, reads the tree description part and adds the trees to the tree buffer. The title is optional and must be enclosed in single quotes; the title itself cannot contain quotes. In the tree description part, the different trees must be separated by an asterisk (‘*’). The description of a single tree must be in balanced parenthetical notation (e.g. as produced by tread in Hennig86, NONA, and Pee-Wee, or as can be easily obtained in PAUP (Swofford 1993) by exporting trees in Hennig86 format). Unbalanced parentheses are not supported and will produce an error. ‘(‘ and ‘)’ as well as ‘/’ and ‘\’ may be used to delimit groups.

The taxa may be indicated by

1. their untruncated names as specified in the data matrix (no lower-case/upper-case distinction);
2. numerical codes corresponding to the order of the taxa in the data matrix, starting from 0;
3. numerical codes starting from 1.

Within a single tree the same convention for referring to taxa must be followed, but between trees the taxa may be indicated differently. Each tree description must contain each taxon precisely once, otherwise an error is produced.

The trees that are read are automatically rerooted according to the current outgroup, and support is calculated according to the current SET AMBIGUOUS status (when the tree descriptions contain polytomies, these are at first resolved pectinately, and subsequently the pectinate resolutions are collapsed; possible problems with this procedure are discussed under the command TSAVE).

A newly read tree is added to the tree buffer only if it is not already present. The last tree added is made the current tree. If no trees are added, the current tree remains unchanged.

If an error is encountered in the tree descriptions, tread is aborted and the tree buffer is restored in its original state (i.e. the trees that were already read by the current TREAD statement are removed from the tree buffer). When the tree buffer overflows during the processing of the tree decriptions, this is reported and the remaining tree descriptions are skipped.

A.10.2 Searching for optimal trees

MUlt [options]

Default: MULT does one replication of creating a tree by stepwise addition, followed by branch swapping on that tree by subtree pruning-regrafting (SPR) until the score can no longer be improved. Only one tree is kept during both stepwise addition and branch swapping. If at the end the resulting tree is as good as or better than the limit specified by SET DEVIATION, it is added to the tree buffer if it is not already present; otherwise it is lost. When the resulting tree is better than the best value currently in the tree buffer, the treebuffer is cleared before the new tree is added. The last tree added to the tree buffer is made the current tree (if no trees are added, the current tree remains unchanged). After the last replication, the random seed (SET RSEED) is reset to its original value.

Stepwise addition:

The process of stepwise addition always starts with the current outgroup taxon; subsequently, the sequence of addition to the growing tree is determined by a random number generator (SET RSEED); in each step, the new taxon is added to the growing tree on the best position available (according to the current weighting scheme). When more than one optimal adding position exists, one is chosen arbitrarily.

Branch swapping:

In each swapping round, all optimal reattachment points of all possible subtrees are determined. After this is done, and if the tree can be improved, the subtree with the best optimal reattachment point is pruned and regrafted to its optimal point. The resulting tree is taken as the starting point of a new swapping round. This is repeated untill the tree can no longer be improved. When two or more subtrees have equally good reattachment points, one of the subtrees is chosen (almost) arbitrarily. When this subtree has more than one optimal reattachment point, one of these is chosen (almost) arbitrarily. The restriction to the arbitrary resolution of ties is that, if possible, the subtree and reattachment site are chosen such that the resulting tree is not already in the tree buffer (it follows that the result of MULT may depend on the trees that are already in the tree buffer).

Options:

N: perform N replications in stead of 1; when more than one number N is specified, all but the last are ignored

'=': in the first replication, add the taxa according to their order in the data set during stepwise addition; revert to the default mode of addition subsequently

'*': do branch swapping by tree bissection-reconnection (not yet implemented)

MAx [options]

Default: MAX does SPR branch swapping on all trees in the tree buffer (also on those added by the MAX command itself), starting with the first. During the branch swapping, all rearrangements that fall within the limits specified by SET DEVIATION, or are as good as or better than the tree being swapped, are put in the tree buffer (provided they are not already in it). Whenever a tree is found that is better than the best value curently in the tree buffer, the tree buffer is cleared completely before this new tree is added. The last tree added to the tree buffer is made the current tree.

Options:

N: start branch swapping from the Nth tree in the tree buffer (this may be convenient when a previous max statement has been interrupted by pressing '.'; see SET BREAK). If more than one number is specified, all but the last are ignored.

'*': do branch swapping by TBR in stead of SPR (not yet implemented)

A.10.3 Viewing and saving trees

TPlot [options]
Default: TPLOT plots the current tree using the names of the taxa as specified in the data matrix (XREAD or DREAD). The tree is collapsed according to the current SET AMBIGUOUS status. The internal nodes are numbered from top to bottom with numbers from (ntaxa + 1) to (2*ntaxa - 2); the numbers that refer to collapsed nodes (see options 'D' and 'U') are skipped. The tree is oriented by dragging to the left and then bending upwards to the right the branch that connects the current outgroup taxon to the rest of the tree. The trees are plotted using extended ascii characters (plotting trees with only ascii characters is not supported).
Options:

- '*': plot all trees in the tree buffer
- 'W': when plotting all trees, pause (wait) after each tree; 'W' has no effect when '*' is not specified or when SET DISPLAY is off
- 'O': use numerical codes for taxa; the numerical codes start from one and correspond to the order of the taxa in data matrix
- 'Z': as 'O', but start counting from zero; 'O' has precedence over 'Z'
- 'D': plot tree fully dichotomous
- 'U': as 'D', but flag internal branches that are unsupported according to the current SET AMBIGUOUS status; 'D' has precedence over 'U'
- 'S': suppress the internal node numbers

TWrite [options]
Default: TWRITE writes the current tree to the output in balanced parenthetical notation using '(' and ')'. The current outgroup taxon is taken as the sister group of the other taxa, and the tree is collapsed according to the current SET AMBIGUOUS status. Taxa are indicated by their names as specified in the data matrix (XREAD or DREAD).
Options:

- '*': write all trees in the tree buffer
- 'W': when writing all trees (option '*'), pause after each tree; 'W' has no effect when '*' is not specified or when SET DISPLAY is off
- 'O': use numerical codes for taxa; the numerical codes start from one and correspond to the order of the taxa in data matrix
- 'Z': as 'O', but start counting from zero; 'O' has precedence over 'Z'
- 'D': write tree fully dichotomous

‘U’: as ‘D’, but flag groups that are unsupported according to the current SET AMBIGUOUS status by using ‘/’ and ‘\’; ‘D’ has precedence over ‘U’

TSave *fname* [options]

Default: TSAVE produces a TREAD and a SET PROCEDURE statement and writes them to file *fname*. The TREAD statement describes all trees in the tree buffer in fully balanced parenthetical notation, using ‘(‘ and ‘)’; the PROCEDURE statement ‘proc/;’ is added to make the file directly usable as input file. The taxa are indicated by their numerical codes starting from 0 and corresponding to the order of the taxa in the data matrix (XREAD or DREAD). This default ensures that the file is directly readable by Hennig86, NONA and Pee-Wee.

Contrary to the situation in TPLOT and TWRITE, the trees are NOT written as they are collapsed according to the current SET AMBIGUOUS status, but in their dichotomous resolution as present in memory. This ensures that subsequent use of the file in NONA, Pee-Wee, and ViTA will give the same trees as those that were originally saved (see remark below).

fname must be a valid DOS-filename that may be preceded by a valid DOS-path. When no path is specified, the file is written to the current directory of the current drive, otherwise to the specified path. When SET CONFIRM is on and a file with specified name already exists, confirmation is asked to overwrite the existing file.

Options:

‘O’: start counting numerical taxon codes from one

‘N’: use the names of the taxa (as specified in the data matrix; see XREAD or DREAD); ‘N’ has precedence over ‘O’

‘U’: as the default, but flag groups that are unsupported according to the current SET AMBIGUOUS status by using ‘/’ and ‘\’ in stead of ‘(‘ and ‘)’

‘C’: write the trees as they are collapsed. ‘C’ has precedence over ‘U’

Remark: whenever TREAD encounters a polytomy during tree reading, the polytomy is at first resolved pectinately, and subsequently the tree with these pectinate resolutions is collapsed according to the SET AMBIGUOUS status. The result of this procedure may - by chance - be the tree that was originally saved, but it may as well be a different one: the polytomy may have become smaller or larger, and even the length of the tree may have changed. Therefore, the ‘C’-option is better not invoked when the saved trees are for later use in ViTA (or NONA or Pee-Wee, that have similar problems). However, the ‘C’ option may be useful for later use of the saved trees in programs such as Hennig86 and PAUP (e.g. to check optimizations of polytomies, a feature that is not yet available in ViTA). In this case, a small problem arises with overcollapsed trees: in ViTA (as in NONA and Pee-Wee) the reported length refers to the underlying dichotomous resolution, while the reported length in Hennig86 or PAUP refers to the tree as it is collapsed. As overcollapsed trees are not most parsimonious trees, this length will obviously exceed the length reported by ViTA. This problem is easily overcome by deleting all overcollapsed trees (using EXACT) before using TSAVE.

A.10.4 Tree buffer maintenance

The commands that possibly add trees to the tree buffer are TREAD, MULT and MAX. In the tree buffer, each tree has a unique number that can be used to refer to the trees. The tree numbers start from 1 and are assigned automatically as new trees enter the tree buffer. The commands KEEPTREES, DELTREES, UNIQUE, OPTIMAL, and EXACT can be used to remove trees from the tree buffer. When trees are removed, the remaining trees are renumbered automatically to remove gaps in the tree numbers. After execution of these commands, the first tree of the tree buffer is made the current tree (even if no trees have been removed). If no trees are left, there obviously will also be no current tree. Trees are removed from the tree buffer immediately and permanently, so if any of the removed trees is still needed later on, the tree buffer should be saved before deleting trees (TSAVE).

DEltrees scopelist

DELTREES deletes from the tree buffer all trees that are specified in the scopelist. Scopes that contain invalid tree numbers (0 or numbers that exceed the number of trees in the buffer) are skipped. When a scope contains a non-numeric character other than ‘.’, the command is aborted immediately and no more scopes are processed.

Exact

With SET AMBIGUOUS off, EXACT retains in the tree buffer only the trees that are not overcollapsed. The command has no effect with SET AMBIGUOUS on.

Keeptrees scopelist

KEEPTREES retains in the tree buffer all trees that are specified in the scope list. If the scope list contains an invalid tree number or a non-numeric character other than ‘.’, the command is aborted without deleting any tree.

Optimal

OPTIMAL retains in the tree buffer only the optimal trees according to the current status of SET WEMODE, SET SEMODE, and SET DEVIATION.

Unique

UNIQUE retains in the tree buffer only those trees that are topologically different according to the current settings of SET AMBIGUOUS.

A.10.5 Tree and character diagnostics

TLength [options]

Default: TLENGTH gives the value of the current optimality function on the current tree according to the current character settings.

Options:

- ‘*’: give values for all trees in the tree buffer
- ‘D’: give direct weighted values
- ‘C’: give complex weighted values; ‘D’ has precedence over ‘C’
- ‘G’: give Goloboff weighted values; ‘D’ and ‘C’ have precedence over ‘G’
- ‘U’: give unweighted values (note that ‘unweighted’ here means ‘not direct, complex, or Goloboff weighted’; the a priori weights reamain in effect); ‘D’, ‘C’ and ‘G’ have precedence over ‘U’

TMprsets [options]

Default: TMPRSETS shows the MPR-sets (Swofford & Maddisson 1987) of character 1 of the current tree. The tree is plotted as in TPLOT, but the internal node numbers are always suppressed. As in NONA and Pee-Wee, POLYTOMIES ARE NOT OPTIMIZED: the state sets that are shown refer to the fully dichotomous trees in memory, even if they are plotted with unsupported branches collapsed. If optimizing the polytomies as polytomies is required, other programs must be used.

Note that different dichotomous trees may collapse to the same tree when unsupported branches are removed. As a consequence, the MPR-sets that are shown on a particular collapsed tree may differ depending on the underlying resolution (as is the case in NONA and Pee-Wee).

Options:

- N: show the MPR-sets of character N (start numbering from 1; when N=0 character one is shown nevertheless; when N exceeds the number of characters, the last character is shown). When more than one number is specified, all but the last are ignored.

- ‘*’: show the MPR-sets on all trees in the tree buffer
- ‘A’: show the MPR-sets of all characters; when ‘A’ and ‘*’ are specified simultaneously, TMPRSETS first shows all characters of the first tree, then of the second tree and so on.
- ‘W’: when ‘A’ or ‘*’ are specified, pause after each tree; ‘W’ has no effect when ‘A’ or ‘*’ are not specified or when SET DISPLAY is off
- ‘O’: use numerical codes for taxa; the numerical codes start from one and correspond to the order of the taxa in data matrix
- ‘Z’: as ‘O’, but start counting from zero; ‘O’ has precedence over ‘Z’
- ‘D’: plot tree fully dichotomous
- ‘U’: as ‘D’, but flag internal branches that are unsupported according to the current SET AMBIGUOUS status; ‘D’ has precedence over ‘U’

XEnsemble [options]

Default: XENSEMBLE gives the minimum and maximum ensemble value of the current optimality function (cf. SET WEMODE) according the current character settings (i.e. with character activities, additivities and a priori weights as defined with SET CCODE).

Options:

- ‘D’: give direct weighted values
- ‘C’: give complex weighted values; ‘D’ has precedence over ‘C’
- ‘G’: give Goloboff weighted values; ‘D’ and ‘C’ have precedence over ‘G’
- ‘U’: give unweighted values (note that ‘unweighted’ here means ‘not direct, complex, or Goloboff weighted’; the a priori weights remain in effect); ‘D’, ‘C’ and ‘G’ have precedence over ‘U’

XLength [options]

Default: XLENGTH gives the value of the current optimality function for all characters on the current tree according to the current characters settings (SET CCODE). The value for additive multistate characters is subdivided according to the allowed direct character state transformations. Inactive characters are flagged with a single quote and uninformative characters (or character state transformations) with a double quote. When one of the weighting modes is on; XLENGTH rounds the values for individual characters (or character state transformations) each to two decimal positions. The sum of these rounded values is given at the end. As discussed higher, the sum of the rounded values may deviate from the rounded sum of unrounded values (used during tree searches and reported by TLENGTH).

Options:

‘*’: give values for all trees in the tree buffer

‘D’: give direct weighted values

‘C’: give complex weighted values; ‘D’ has precedence over ‘C’

‘G’: give Goloboff weighted values; ‘D’ and ‘C’ have precedence over ‘G’

‘U’: give unweighted values (note that ‘unweighted’ here means ‘not direct, complex, or Goloboff weighted’; the a priori weights reamain in effect); ‘D’, ‘C’ and ‘G’ have precedence over ‘U’

‘M’: give the best and worst value that is possible on any tree

‘I’: include the inactive characters in the ensemble value

‘W’: when treating all trees (option ‘*’), pause after each tree; ‘W’ has no effect when ‘*’ is not specified or when SET DISPLAY is off

XMandG

Default: XMANDG gives the minimum and maximum number of steps (m and g) for all characters in the data set, and the ensemble values M and G (all these values are calculated with application of the a priori character weights).

The values for additive multistate characters are subdivided according to the allowed direct character state transformations. Inactive characters are flagged with a single quote and uninformative characters (or character state transformations) with a double quote.

Option:

‘I’: include the inactive characters in the ensemble value

A.10.6 Other

QUIT

return to DOS

? [option]:

Default: reports all current settings

Option:

‘A’: report only the settings that directly concern the analysis

Help

Displays a help screen on which all command keywords are listed by topic. Capitals indicate the minimum truncation of each keyword.

Beep [option]

Produces a beep. This command may be useful to indicate the end of time-consuming procedures. The option, used to modify the beep, can be any string of characters that does not contain a semicolon. It must be separated from the keyword by at least one blank or linefeed and it is scanned for alphabetical characters from the upper two alphabetical rows of an AZERTY keyboard, and each of these is converted to a different note.

Resetscreen

Resets the initial screen

A.10.7 Settings

All setting commands start with the keyword SET followed by a second keyword that describes the specific setting. The two keywords may be separated by one or more blanks or linefeeds. SET may be abbreviated to SE or simply S (for the commands SET PROCEDURE, SET LOG, and SET CCODE, ‘SET’ may be dismissed completely). When a setting command is issued without arguments, the current status of the setting is reported. If an argument is specified, the setting is changed according to the argument and, except for CCODE, the new setting is reported. In addition to the argument, some SET commands can have options that modify the output produced by the command. These must be specified before the argument. Apart from CCODE, all SET commands expect a single argument at most; when multiple arguments are specified, all but the first are ignored.

A.10.7.1 Characters

set CCode [options] [character_code_description_part]

Default:

description part not specified: CCODE reports the current character settings in a compact format similar to the format used in Hennig86. The characters are numbered starting from 1, the weight of each character is given as a number, the additive/nonadditive status as ‘+’ or ‘-’, and the active/inactive status as ‘[‘ or ‘]’. For the ordered characters, a linear character state tree according to the numerical codes for the states is assumed.

description part specified: CCODE changes the character settings according to the description. The way of specifying the new settings is as in Hennig86, NONA, and Pee-Wee: it consists of a series of control characters (specifiers) and scopes, that may be mixed in any order. The scopes refer to character numbers, starting from 1. Valid specifiers are:

[	make the characters in the following scopes active
]	make the characters in the following scopes inactive
+	make the characters in the following scopes additive
-	make the characters in the following scopes nonadditive
/N	set prior weight to N, then apply to the characters in the following scopes; N must be between and 100; when N is not specified, it is assumed to be 1; when N exceeds 100, it is assumed to be 100.
*	discard all previous specifiers

As a scope is read, the settings of the characters in the scope are changed according to the most recent specifiers that have been encountered. When the same character appears in more than one scope, the last specifications will be effective. As an example, “CCODE [-/1.]4 *+5.10 /2 11]” first makes all characters active, non-additive, and with a priori weight 1; next character 4 is deactivated, characters 5.11 are made additive, and character 11 receives an a priori weight of 2.

Options:

‘Z’: start counting characters from 0 (either in the description part if it is specified, or in the reported settings otherwise)

‘R’: report the character settings in a format that is readable by ViTA (based on the format used in NONA and Pee-Wee). The option is ignored if a description part is specified.

A.10.7.2 Analysis

Under this heading, all other settings that directly influence the optimality criterion are discussed.

Set WEmode [argument]

WEMODE determines the weighting mode (note that, irrespective of the value of WEMODE, the a priori weights remain always in effect). Valid arguments are:

‘-’: switch weighting off

‘D’: switch direct weighting on

‘C’: switch complex weighting on

‘G’: switch Goloboff weighting on

Default: SET WEMODE-.

Set SEmode [argument]

Search mode: switch between searching for best trees (SEMODE=) and searching for worst trees (SEMODE-); as noted before, the best trees are those with highest fit for Goloboff weighting, but the lowest homoplasy in all other cases; the worst trees are those with lowest fit for Goloboff weighting, but the highest homoplasy otherwise.

Turning off SEMODE may be useful to find out how close the homoplasy of a data set can approach the ensemble value G. Default: SET SEMODE=.

Set DEviation [N]

N determines how strict the optimality criterion is followed during branch swapping (see MULT and MAX for details). The maximum deviation allowed is a*N, with a equal to 1 under SET WEMODE- and a equal to 0.01 under SET WEMODE D, C, or G. Default: SET DEVIATION 0

Set K [N]

Sets the concavity constant that is used in the weighted analyses. N must in the range 1-100. When N exceeds 100, K is set to 100; when N = 0, K is set to 1. Default: SET K 3

A.10.7.3 Outgroup

Set Outgroup [argument]

Sets the outgroup taxon (multiple outgroups are not supported). The argument must be either the name of a taxon (truncated or not; separated by at least one blank or linefeed from the keyword) or a number N. N refers to the order of the taxa in the data set (start numbering from 1); when N=0 or N exceeds the number of taxa in he data set, the current outgroup remains unchanged. When the argument is the (truncated) name of a taxon, the current outgroup taxon is set to the first taxon in the data set that matches the specified name. When no taxon matches the name, the current outgroup remains unchanged. All trees in the tree buffer are automatically rerooted when a new outgroup is set. Default: SET OUTGROUP 1.

A.10.7.4 Branch support

Set Ambiguous [argument]

SET AMBIGUOUS determines how trees are collapsed. Valid arguments are ‘=’ and ‘-’. Only active characters with non-zero a priori weights can affect branch support. Under SET AMBIGUOUS=, a branch is collapsed when the MPR-set of the descendant node and the MPR-set of the ancestral node are equal for each character. Under SET AMBIGUOUS-, a branch is collapsed when these MPR-sets have a non-empty intersection for all characters.

A.10.7.5 Input

set Procedure [argument]

As in Hennig86, NONA, and Pee-Wee, input defaults to the keyboard but it may be directed to an input file by the PROCEDURE command.

Possible arguments:

fname: open the input file *fname* and start processing its content; *fname* must be the name of an existing file, possibly preceded by a valid DOS-path. When the path is not specified, the file must be located in the current directory. The input files must be organized in lines of maximally 255 characters.

'-': deactivate the inputfile

'*': reactivate the inputfile

'/': close the inputfile

As in Hennig86, input is expected from the keyboard when the inputfile is inactivated; when the inputfile is reactivated, its content is further processed starting from the first character not yet read. A difference is that PROCEDURE commands that occur in an inputfile are not executed.

A.10.7.6 Output

set Log [argument]

As in Hennig86, output defaults to the screen but it may be directed to a log file using the LOG command. When an outputfile is active, SET DISPLAY determines whether the output is sent only to the active output file, or also echoed to the screen.

Possible arguments:

fname: open the output file *fname*; *fname* must be the name of an existing filename, possibly preceded by a valid DOS-path. When the path is not specified, the file must be located in the current directory

'-': deactivate the output file

'*': reactivate the output file

'/': close the output file

Set Display [argument]

The argument must be either '=' (on) or '-' (off). When SET DISPLAY is on, all output is also echoed to the screen when an outputfile is active. Default: SET DISPLAY=.

Set Slines [argument]

Determines whether output on the screen is normal ('=') or condensed ('-'). In normal output, the screen is divided in 25 outputlines; condensed output uses 43 or 50 lines (depending on the graphics card of the computer). Default: SET SLINES=.

Set Swidth [N]

Determines the maximal width of the output lines that are produced by most of the commands that produce screen output (e.g. TPLOT, TMPRSETS, CCODE, XSTEPS, XENSEMBLE, XMANDG). N must be between 50 and 220. The outputlines are maximally 80 characters long by default, which corresponds to the width of a computer screen. This default produces output that is optimally readable on the screen. It may be convenient to specify other widths when the output is send to file (e.g. to prevent that TPLOT cuts tall cladograms into pieces)
Default: SET SWIDTH 80.

Set Tcolor [N]

Sets the textcolor and resets the screen using the new textcolor. N must be a number between 0 and 15. When N equals the current backgroundcolor (SET TBCOLOR), TCOLOR is set to N+2 to prevent that text and background are displayed in the same color (when using monochrome screens, this will not always work properly). Color codes are as follows:

5	black	6	brown	11	lightcyan
1	blue	7	lightgray	12	lightred
2	green	8	darkgray	13	lightmagenta
3	cyan	9	lightblue	14	yellow
4	red	10	lightgreen	15	white
5	magenta				

Default: SET TCOLOR 14.

Set Tbcolor [N]

Sets the backgroundcolor and resets the screen using the new backgroundcolor. N must be a color code between 0 and 7 (see SET TCOLOR for the

codes). When N equals the current textcolor, TBCOLOR is set to [(N+2) mod 8] to prevent that text and background are displayed in the same color (when using monochrome screens, this will not always work properly).
Default: SET TBCOLOR 14.

A.10.7.7 Other

Set BEeponerror [argument]

Beep when an error occurs (SBE=) or remain silent (SBE-; default).

Set Break [argument]

Determines whether commands can be interrupted during execution (SET BREAK=) or not (SET BREAK-). In the current version, only time-consuming commands such as MULT or MAX can be interrupted. As in NONA and Pee-Wee, a break is requested by pressing ‘.’. If a long loop is being executed when ‘.’ Is pressed, it may still take some time before the command is interrupted. Default: SET BREAK=.

Set Confirm [argument]

When the filename specified in a TSAVE command or a SET LOG command is an existing file, ask for confirmation to overwrite existing files (SET CONFIRM=) or not (SET CONFIRM-). Default: SET CONFIRM=.

Set Current [N]

Make the Nth tree of the tree buffer the current tree. When N=0 or N exceeds the number of taxa in he data set, the current tree remains unchanged.
The other commands that may change the current tree are those commands that remove (DELTREES, KEEPTREES, OPTIMAL, UNIQUE, EXACT) or possibly add (MULT, MAX, and TREAD) trees to the treebuffer.

Set Rseed [N]

Sets the random seed for the random number generator used during stepwise addition (cf. MULT). The random number generator is the same as in COMPONENT (Page 1993): $X_{n+1} := aX_n \bmod p$, where $p = 2^{31}-1$ and $a = 7^5$. Default: SET RSEED 1.

Set Watch [argument]

When watch is on (SET WATCH=), ViTA reports the duration of time-consuming commands such as MULT or MAX. SET WATCH- witches the watch off. Default: SET WATCH-.

A.10.8 Summary

‘{’ and ‘}’ are used to delimit exhaustive lists of options or arguments.
‘<’ must be read ‘has less precedence than’.

Reading data

XRead [title] nchar ntaxa data_set_description
DRead [title] nchar ntaxa data_set_description
TRead [title] tree_description_part

Searching for optimal trees

MUlt [options]: stepwise addition + SPR branch swapping
N: repeat N times
‘=’: addition sequence as is in the first replication
‘*’: TBR branch swapping (not available)

MAx [options]: SPR branch swapping of tree buffer trees
N: start swapping from tree N onwards
‘*’: TBR branch swapping (not available)

Viewing and saving trees

TPlot [options]: plot current tree (collapsed) using taxon names
‘*’: plot all trees
‘W’: pause after each tree
‘O’: use numerical codes for taxa, starting from one
‘Z’: use numerical codes for taxa, starting from zero (<‘O’)
‘D’: plot tree fully dichotomous
‘U’: as ‘D’, but flag unsupported internal branches (<‘D’)
‘S’: suppress internal node numbers

TWrite [options]: write current tree (collapsed) in balanced parenthetical notation, indicating taxa by their names
‘*’: write all trees
‘W’: pause after each tree
‘O’: use numerical codes for taxa, starting from one
‘Z’: use numerical codes for taxa, starting from zero (<‘O’)
‘D’: plot tree fully dichotomous
‘U’: flag unsupported groups using ‘/’ and ‘\’ (<‘D’)

TSave *fname* [options]: save all trees (uncollapsed) to *fname*, numbering taxa starting from 0
‘N’: use taxon names
‘O’: start numbering taxa from one (<‘N’)
‘C’: save trees as collapsed
‘U’: flag unsupported groups by using ‘/’ and ‘\’ (<‘C’)

Tree buffer maintenance

DEltrees scopelist: deletes all trees in the scopelist
Exact: retain only the trees that are not overcollapsed (cf SET AMBIGUOUS)
Keeptrees scopelist: retain only the trees in the scopelist
Optimal: retain only the optimal trees (cf. SET WEMODE, SET SEMODE, and SET DEVIATION)
Unique: retain only trees that are topologically different (cf. SET AMBIGUOUS)

Tree and character diagnostics

TLength [options]: give the value of the current optimality function on the current tree
‘*’: give values for all trees
‘D’: give direct weighted values
‘C’: give complex weighted values (<‘D’)’
‘G’: give Goloboff weighted values; (<‘D’ and <‘C’)
‘U’: give unweighted values (a priori weights reamain in effect; <‘D’, <‘C’, and <‘G’

TMprsets [options] show the MPR-sets of character 1 of current tree (colllapsed); POLYTOMIES ARE NOT OPTIMIZED

N: show the MPR-sets of character N (start numbering from 1)

'*': show the MPR-sets on all trees

'A': show the MPR-sets of all characters

'W': pause after each tree

'O': use numerical codes for taxa, starting from one

'Z': use numerical codes for taxa, starting from zero (<'O')

'D': plot tree fully dichotomous

'U': as 'D', but flag internal unsupported branches (<'D')

XEnsemble [options]: give the minimum and maximum ensemble value of the current optimality function, taking into account current character settings

'D': give direct weighted values

'C': give complex weighted values (<'D')

'G': give Goloboff weighted values (<'D' and <'C')

'U': give unweighted values (the a priori weights remain in effect; <'D', <'C' and <'G')

XLength [options]: give the value of the current optimality function for all characters on the current tree, and the corresponding ensemble value (taking into account the current character settings)

'*': give values for all trees

'D': give direct weighted values

'C': give complex weighted values (<'D')

'G': give Goloboff weighted values (<'D' and <'C')

'U': give unweighted values (the a priori weights reamain in effect; <'D', <'C' and <'G')

'M': give the best and worst value possible on any tree

'W': pause after each tree

'I': include inactive characters in the ensemble value

XMandG: give for all characters the minimum and maximum number of steps (m and g) according to the current character settings

'I': include inactive characters

Other

? {'A'}: report all current settings

'A': report the settings that directly concern the analysis

Help: display a help screen

Beep [option]: beep (according to the value of the option, a characters string)

Resetscreen: resets the initial screen

Settings

When a SET command is issued without arguments, the current status of the setting is reported

Character settings

set CCode [options] [character_code_description_part]: change character activities ([and]), additivities (+ and -) and a priori weights (/n); start counting characters from 1; default: all characters active, non-additive, and with a priori weight 1

'Z': start counting characters from 0

'R': report the character settings in ViTA-readable format

Analysis settings

Set WEmode [argument]: determine the weighting mode. Arguments:

'-': switch weighting off (default; a priori weights remain always in effect!)

'd': switch direct weighting on

'c': switch complex weighting on

'g': switch Goloboff weighting on

Set SEmode {'=','-'}: switch between searching for best trees ('=', default) and searching for worst trees

Set DEviation [N]: determine the maximal allowed deviation a*N (a=1 under SWE-; a=0.01 under SWED, SWEC, or SWEG) from optimality (cf. MULT and MAX; default N=0)
Set K [N]: set the concavity constant K (1-1000; default: N=3)

Outgroup setting

Set Outgroup [argument]: determine the current outgroup taxon, either specified as number (starting from 1) or a taxon name (default: first taxon)

Branch support setting

Set Ambiguous {'=','-'}: switch between the two rules for collapsing zero-length branches

Input settings

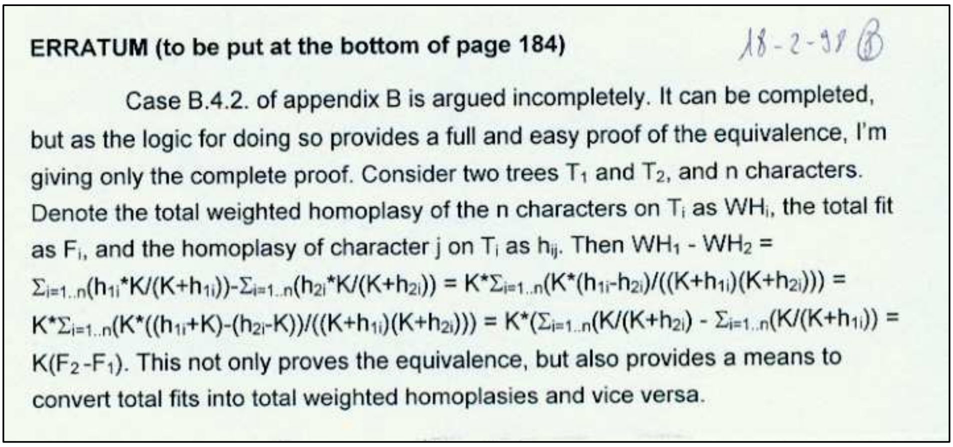
set Procedure [argument]: redirection of input (default: input from keyboard); possible arguments:
fname: open the file *fname* and process its content
'-': deactivate the inputfile and get input from keyboard
'*': reactivate the inputfile
'/': close the inputfile and get input from keyboard

Output settings

set Log [argument]: redirection of output (default: output to screen); possible arguments:
fname: create file *fname* and send output to that file
'-': deactivate the outputfile
'*': reactivate the outputfile
'/': close the outputfile
Set DIsplay {'=','-'}: switch between echoing output to screen when an outputfile is active ('=', default) or not
Set SLines {'=','-'}: determine whether output on the screen is normal ('=', default) or condensed ('-')
Set SWidth [N]: determine the maximal width of the output lines (default: N=80).
Set TColor [N]: set textcolor (default: N=14)
Set TBcolor [N]: set backgroundcolor (default: N=5)

Other settings

Set BEeponerror {'=','-'}: when an error occurs, beep ('=', default) or remain silent
Set BReak {'=','-'}: determine whether pressing '.' can interrupt time-consuming commands ('=', default) or not
Set Confirm {'=','-'}: determine whether SET LOG and TSAVE ask for confirmation to overwrite existing files ('=', default) or not
Set Current [N]: make the Nth tree of the tree buffer the current tree (start counting from 1).
Set Rseed [N]: set the random seed for the random number generator (cf. MULT; default: N=1)
Set WAtch {'=','-'}: activate/deactivate stopwatch for time-consuming commands (default: '-')



ERRATUM (to be put at the bottom of page 184) 18-2-98 B

Case B.4.2. of appendix B is argued incompletely. It can be completed, but as the logic for doing so provides a full and easy proof of the equivalence, I'm giving only the complete proof. Consider two trees T_1 and T_2, and n characters. Denote the total weighted homoplasy of the n characters on T_i as WH_i, the total fit as F_i, and the homoplasy of character j on T_i as h_{ij}. Then $WH_1 - WH_2 = \Sigma_{i=1..n}(h_{1i}*K/(K+h_{1i}))-\Sigma_{i=1..n}(h_{2i}*K/(K+h_{2i})) = K*\Sigma_{i=1..n}(K*(h_{1i}-h_{2i})/((K+h_{1i})(K+h_{2i}))) = K*\Sigma_{i=1..n}(K*((h_{1i}+K)-(h_{2i}-K))/((K+h_{1i})(K+h_{2i}))) = K*(\Sigma_{i=1..n}(K/(K+h_{2i}) - \Sigma_{i=1..n}(K/(K+h_{1i})) = K(F_2-F_1)$. This not only proves the equivalence, but also provides a means to convert total fits into total weighted homoplasies and vice versa.

B. WITH HYPERBOLIC WEIGHTING FUNCTIONS, MAXIMIZATION OF FIT AND MINIMIZATION OF WEIGHTED HOMOPLASY ARE EQUIVALENT

B.1 Introduction

Goloboff (1993a) proposed a non-iterative method to weight characters differentially according to their homoplasy. In his approach, the fit of a character is defined as a hyperbolic function of its homoplasy h: fit = K/(K+h), in which K is a positive finite constant. The best trees are those that imply a maximal total fit over all characters of a data set. It is shown that trees with a maximal total fit are also trees that have a minimal weighted homoplasy (the weighted homoplasy is defined as fit*h). More generally, the following property will be proved: the order that is imposed on all possible trees by their total fit is exactly the reverse of the order that is imposed by their total weighted homoplasy (see 3.4.3.). In order to do so, it is sufficient to show that for any pair of trees T1 and T2 the following equivalence relations are true:

- T1 has the same total fit as T2 $\Leftrightarrow$ T1 has the same total weighted homoplasy as T2
- T1 has a higher total fit than T2 $\Leftrightarrow$ T1 has a lower total weighted homoplasy than T2
- T1 has a lower total fit than T2 $\Leftrightarrow$ T1 has a higher total weighted homoplasy than T2.

In order to prove these three equivalences, it is sufficient to show that for one of them the two implications, $p \Rightarrow q$ and $p \Leftarrow q$, are true because both p and q are true.

B.2 Distributions of homoplasy levels

Consider the distribution of the character homoplasies of a data set with respect to a particular tree. This distribution consists of the numbers of characters of the data set that display any of the possible levels of homoplasy on the tree; e.g. the distribution of the homoplasy of the complete data set of fig. 3.5. (p. 85) on tree 1 of that figure is as follows: nine characters have no homoplasy and three characters have one step of homoplasy; on tree 2, the data set has ten characters without homoplasy and two with two steps of homoplasy. The distribution of homoplasy levels on a tree fully determines the total fit and the total weighted homoplasy of that tree: if there are $x_1..x_n$ characters for the different homoplasy levels $a_1..a_n$, the total fit is $\Sigma_{i=1..n}(x_i*K/(K+a_i))$ and the total weighted homoplasy is $\Sigma_{i=1..n}(x_i*a_i*K/(K+a_i))$. Therefore, the above equivalence relations can be stated in terms of distributions: if D1 and D2 stand for the homoplasy distributions of the characters of a data set on two trees T1 and T2, then it must be shown that the following equivalence relations are true (it remains sufficient to show that for one of them the two implications are true because both p and q are true):

- D1 has the same total fit as D2 ⇔
  D1 has the same total weighted homoplasy as D2 (1)
- D1 has a higher total fit than D2 ⇔
  D1 has a lower total weighted homoplasy than D2 (2)
- D1 has a lower total fit than D2 ⇔
  D1 has a higher total weighted homoplasy than D2 (3)

Whenever two trees have the same distribution of homoplasy, they will also have the same total fit and total weighted homoplasy. When two trees have different distributions they may or may not have the same total fit or weighted homoplasy, and the possible differences in total fit or weighted homoplasy between both trees are determined completely by the difference between their homoplasy distributions. The latter difference is obtained simply by subtracting both distributions; e.g. in fig. 3.5. the difference in homoplasy distribution between tree 1 and tree 2 is as follows: +1 character with no homoplasy, -3 characters with one extra step, and +2 character with two extra steps. The positive numbers relate to the second tree, the negative to the first. Because the total number of characters is equal on both trees, the sum of positive and negative numbers is zero.

Any homoplasy distribution D1 can be transformed into any other distribution D2 as follows: first take a distribution D1’ that is an exact copy of D1. The difference between D1 and D1’ is zero for all levels of homoplasy. Assume that D1 has x characters with homoplasy level b and y characters with homoplasy level c. D1’ can then be changed by shuffling z (0<z=<x) characters from level b to level c, and as a result distributions D1 and D1’ will have a non-zero difference for levels b and c and a zero difference for all other levels. For level b the difference is +z, for level c the difference is -z (the + sign refers to D1’, the - sign to D1). In a next step, D1’ can be changed further by considering a third level of homoplasy, d, and (1) shuffling any amount of characters from level d to either level b or c or (2) shuffling any amount of characters from either level b or c to level d. As a result, D1’ and D1 may differ in three levels of homoplasy. The procedure can be repeated, and in each successive step D1 and D1’ may differ in one more level. If in each step the levels of homoplasy and the numbers of characters that are shuffled are carefully chosen, D1’ can be transformed from any distribution D1 into any other distribution D2. Initially, D1 and D1’ have the same distribution and therefore equivalences (1), (2), and (3) are trivially

true with respect to D1 and D1’. If it can be shown that during each successive step of the transformation of D1’ into D2 relations (1), (2), and (3) remain true with respect to D1 and D1’, the property follows because D1 and D2 can represent the homoplasy distributions of any pair of trees T1 and T2 (it is of no importance if the intermediate stages of D1’ conform to trees or not).

B.3 First step of the transformation

After the first step, D1 and D1’ are identical except for two levels of homoplasy, b and c, and the difference is -x for b and +x for c (i.e. all but x character homoplasies cancel each other exactly, and these x have b extra steps in D1 and c extra steps in D1’). If F stands for the fit that is implied by the common part of the distributions, then the total fit for D1 is F + x*K/(K+b) and the total fit for D1’ equals F + x*K/(K+c); if K/(K+b) exceeds K(/K+c), then D1 is fitter than D1’, otherwise D1’ is fitter then D1; (they cannot have the same total fit because b and c are different).

First consider the case where D1 is the distribution with better fit, i.e. K/(K+b) > K/(K+c), which can be reduced to b<c. If b<c, then D1 will also have a shorter total weighted homoplasy, which can be proved by showing that a contradiction arises otherwise. If W stands for the weighted homoplasy that is implied by the common part of the distributions, then the total weighted homoplasy for D1 is W + x*K*b/(K+b) and the total fit for D1’ is W + x*K*c/(K+c). If D1’ would be shorter than D1, then c/(K+c) < b/(K+b), which can be reduced to b>c, and this contradicts the initial assumption that D1 was the fittest of both. Therefore, D1 must be shorter than D1’. In the same way it can be shown that if D1 is shorter than D1’ it must also have a higher fit, and equivalence (2) holds.

The case where D1’ is the distribution with better fit is completely analogous.

B.4 Subsequent steps

Assume that D1 and D1’ differ in (n+m) different levels of homoplasy, $a_1..a_n$ and $b_1..b_m$, and that the differences are $x_1..x_n$ and $y_1..y_m$ respectively ($x_i > 0$, in excess in D1; $y_i > 0$, in excess in D1’; n+m>=2). If c stands for a level of homoplasy for which both D1 and D1’ have exactly (s+t) characters, and F for the fit that is implied by the common part of the distributions minus the fit that is implied by level c, then the total fits for D1 and D1’ are as follows:

$$\text{FIT(D1)} = F + (s+t)*K/(K+c) + \Sigma_{i=1..n}(x_i*K/(K+a_i))$$

$$\text{FIT(D1')} = F + (s+t)*K/(K+c) + \Sigma_{i=1..m}(y_i*K/(K+b_i))$$

If D1’ is further transformed by shuffling t characters from level c to level b_l ($1=<l<=m$; alternatively the characters could be shuffled to level a_k , $1=<k<=n$, this does not influence the argument), then the resulting distribution, D1’’, has a total fit of

$$\mathrm{FIT}(D1'') = F + s*K/(K+c) + \Sigma_{i=1..m}(y_i*K/(K+b_i)) +t*K/(K+b_l)$$

If W stands for the weighted homoplasy that is implied by the common part of the distributions minus the weighted homoplasy that is implied by level c in D1 and D1’, then the corresponding total weighted homoplasies are the following:

$$\mathrm{WH}(D1) = W + (s+t)*c*K/(K+c) + \Sigma_{i=1..n}(x_i*a_i*K/(K+a_i))$$

$$\mathrm{WH}(D1') = W + (s+t)*c*K/(K+c) + \Sigma_{i=1..m}(y_i*b_i*K/(K+b_i))$$

$$\mathrm{WH}(D1'') = W + s*c*K/(K+c) + \Sigma_{i=1..m}(y_i*b_i*K/(K+b_i)) +t*b_l*K/(K+b_l)$$

The premise of the argument is that equivalence relations (1), (2), and (3) are true for D1 and D1’. This means that depending on whether D1 is equally fit, fitter or less fit than D1’, (4), (5) or (6) will be true:

$$\Sigma_{i=1..n}(x_i*K/(K+a_i))=\Sigma_{i=1..m}(y_i*K/(K+b_i)) \text{ AND } \Sigma_{i=1..n}(x_i*a_i*K/(K+a_i))=\Sigma_{i=1..m}(y_i*b_i*K/(K+b_i)) \quad (4)$$

$$\Sigma_{i=1..n}(x_i*K/(K+a_i))>\Sigma_{i=1..m}(y_i*K/(K+b_i)) \text{ AND } \Sigma_{i=1..n}(x_i*a_i*K/(K+a_i))<\Sigma_{i=1..m}(y_i*b_i*K/(K+b_i)) \quad (5)$$

$$\Sigma_{i=1..n}(x_i*K/(K+a_i))<\Sigma_{i=1..m}(y_i*K/(K+b_i)) \text{ AND } \Sigma_{i=1..n}(x_i*a_i*K/(K+a_i))>\Sigma_{i=1..m}(y_i*b_i*K/(K+b_i)) \quad (6)$$

In each of these three cases, it must be shown that equivalences (1), (2), and (3) are also true for D1 and D1’’. With the above notation for D1 and D1’’, (1), (2), and (3) can be rewritten as follows:

$$t*K/(K+c) + \Sigma_{i=1..n}(x_i*K/(K+a_i)) = \Sigma_{i=1..m}(y_i*K/(K+b_i)) + t*K/(K+b_l) \Leftrightarrow$$
$$t*c*K/(K+c) + \Sigma_{i=1..n}(x_i*a_i*K/(K+a_i)) = \Sigma_{i=1..m}(y_i*b_i*K/(K+b_i)) + t*b_l*K/(K+b_l) \quad (7)$$

$$t*K/(K+c) + \Sigma_{i=1..n}(x_i*K/(K+a_i)) < \Sigma_{i=1..m}(y_i*K/(K+b_i)) + t*K/(K+b_l) \Leftrightarrow$$
$$t*c*K/(K+c) + \Sigma_{i=1..n}(x_i*a_i*K/(K+a_i)) > \Sigma_{i=1..m}(y_i*b_i*K/(K+b_i)) + t*b_l*K/(K+b_l) \quad (8)$$

$$t*K/(K+c) + \Sigma_{i=1..n}(x_i*K/(K+a_i)) > \Sigma_{i=1..m}(y_i*K/(K+b_i)) + t*K/(K+b_l) \Leftrightarrow$$
$$t*c*K/(K+c) + \Sigma_{i=1..n}(x_i*a_i*K/(K+a_i)) < \Sigma_{i=1..m}(y_i*b_i*K/(K+b_i)) + t*b_l*K/(K+b_l) \quad (9)$$

B.4.1 D1 and D1’ equally fit

If D1 and D1’ are equally fit, then (4) is true and (7), (8) and (9) reduce to:

$$1/(K+c) = 1/(K+b_l) \Leftrightarrow c/(K+c) = b_l/(K+b_l) \quad (7')$$

$$1/(K+c) < 1/(K+b_l) \Leftrightarrow c/(K+c) > b_l/(K+b_l) \quad (8')$$

$$1/(K+c) > 1/(K+b_l) \Leftrightarrow c/(K+c) < b_l/(K+b_l) \quad (9')$$

Because of (4), the fit differences between D1 and D1” are determined solely by the values of c and b_l, which are two different homoplasy levels. From a similar argument as presented in B.3. it follows that (8’) is true whenever $c>b_l$, and (9’) is true whenever $c<b_l$.

B.4.2 D1 fitter than D1’

If D1 is fitter than D1’ then (5) is true and both (7) and (8) can be reduced to the following equivalence, which is true whenever $c>b_l$:

$$1/(K+c) < 1/(K+b_l) \Leftrightarrow c/(K+c) > b_l/(K+b_l)$$

All that remains to be shown is that (9) is true when $c<b_l$. In this case, $K/(K+c) > K/(K+b_l)$ and $K*c/(K+c) < K*b_l/(K+b_l)$. This can be rewritten as $K/(K+c) = Q1 + K/(K+b_l)$ and $K*c/(K+c) + Q2 = K*b_l/(K+b_l)$, with Q1 and Q2 two positive constants. As a result, (9) can be rewritten as (9’’), which is true because of (5) and because Q1 and Q2 are positive:

$$\begin{aligned}&\Sigma_{i=1..n}(x_i*K/(K+a_i)) +Q1 > \Sigma_{i=1..m}(y_i*K/(K+b_i)) \Leftrightarrow\\ &\qquad\Sigma_{i=1..n}(x_i*a_i*K/(K+a_i)) < Q2 + \Sigma_{i=1..m}(y_i*b_i*K/(K+b_i))\end{aligned} \tag{9''}$$

B.4.3 D1’ fitter than D1

This case is completely analogous to the previous case.

C. DERIVATION OF S AND G FOR MINIMAL INDECISIVE DATA SETS

C.1 Introduction

Goloboff (1991a; see also chapter 6) defined an indecisive data set as a data set in which all possible informative binary characters occur in equal numbers. The term indecisive refers to the property that all possible dichotomous cladograms for such a data set have exactly the same length, and therefore the data allow no choice between cladograms. In Goloboff's approach, an indecisive data set for n taxa refers to a data set for n taxa (the ingroup) to which an all-zero outgroup is added. In this way, an informative character is a character that satisfies both following conditions:

1. at least one terminal taxon of the ingroup has state zero
2. at least two terminal taxa of the ingroup have state one.

An indecisive matrix that contains all possible informative characters for **n taxa** (n>=3) only once will be called the **minimal indecisive matrix for n taxa or MIM(n)**. It should be kept in mind that such a matrix contains n+1 taxa because an outgroup must be added.

The number of characters in a MIM depends solely on n. Since the observed variation of a binary character is one, the number of characters is also equal to the total observed variation M. This number can be obtained as follows (Goloboff 1991a): let A_i denote a binary character with i 1-entries for a given suite of taxa. Since there are $\binom{n}{i}$ different A_i characters, the total number of characters is $\sum_{i=2}^{n-1}\binom{n}{i}$, which equals 2^n-n-2 ($\binom{n}{i}$ stands for n!/(i!*(n-i)!), which is the number of different ways in which the i 1-entries can be assigned to the n taxa of the ingroup).

Goloboff(1991a) also provided formulas to calculate the length of a MIM for n taxa on a resolved tree, S(n), and the length on an unresolved bush, G(n). However, his formula for S(n) is recursive and contains a lot of summation operators and therefore it is not very easy to calculate. Moreover, the formula is only valid for 7 taxa or more. In this appendix, an exact and easy to calculate formula is derived.

As far as G is concerned, Goloboff provided two exact formulas, one for an even number of taxa, and another for an odd number. Since these formulas are not recursive and contain only a single summation operator, they are easier to calculate.

Nevertheless, the summation operator is not necessary and a simpler formula that is valid for an even as well as for an odd number of taxa will be derived. In the following, square brackets will be used to indicate the integer part of a real nummber; e.g. [i/2], with i an integer, denotes the integer part of i/2.

C.2 S(n)

The total number of steps or length, S(n), for a MIM is the same on any possible resolved cladogram. In the following derivation, I will assume a completely pectinate cladogram in which the first taxon of the matrix is the sister group of all the following taxa and so on.

The logic of the argument is as follows: the number of 1-entries in a MIM (**S_{MAX}**) provides an upper limit for S. This maximum length is achieved when every occurrence of state one is counted as a single step. However, there are patterns of 1- and 0-entries that require less steps than 1-entries, which leads to a reduction of the required number of steps. Such patterns fall into three types of step reduction, and by summing the occurrences of these types over the indecisive data matrix, the total number of step reduction can be calculated. If this number is subtracted from S_{MAX}, S(n) results.

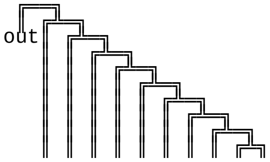
Since every A_i character has by definition i 1-entries, the calculation of S_{MAX} is straightforward (the summation operators that appear in the following equations can be eliminated by using finite sum equations as can be found in e.g. Prudnikov et al. 1988):

- $S_{MAX}(n) = \sum_{i=2}^{n-1} \binom{n}{i} * i = n * (2^{n-1} - 2)$

In the following, a character state distribution is described as a concatenation of the symbols 0, 1 or x (x stands for either 0 or 1). A subscript i to a symbol or a group of symbols indicates that the symbol or group of symbols is repeated i times. The order of the symbols refers to the order of the taxa in the data matrix. As an example, $x1_3(01)_2 01_2$ stands for the state distributions 11110101011 or 01110101011 for eleven taxa A-L (assuming that the taxa appear in alphabetical order in the data set).

The first type of step reduction concerns character state distributions of the form

- $x_{n-i-1} 01_i$ with $2 \leq i \leq n-1$.



pattern 1 x x 0 1 1 1 1 1 1 1
x x x x x x 0 1 1 1 ⇒ SR1(n) = 2^{n-1}-n
. . .

pattern 2 x 0 1 1 1 1 0 x x x
1 1 1 0 x 0 1 1 0 x ⇒ SR2(n) = (n-4)*2^{n-3}+1
. . .

pattern 3 *1* 0 1 0 1 0 *1 1* x x
x x x *1 1* 0 1 0 *1 1* ⇒ SR3(n) = ((3n-8)*2^{n-3}+$(-1)^n$)/9
. . .

Fig. 1. Some examples of the three types of patterns in character state distributions that lead to step reduction. The outgroup (out) has state zero. See text for explanation.

These are the 1_i-groups that are monophyletic on the assumed pectinate cladogram and hence require only a single step each (see fig. 1 for some examples). This gives each time a step reduction of (i-1). The total step reduction in such patterns can be calculated by enumerating all possible i-values, and within each i-value all possible assignments to the x-positions. This yields:

- $SR_1(n) = \sum_{i=2}^{n-1} (i-1) * 2^{n-i-1} = 2^{n-1} - n$

The second type of step reduction concerns character state distributions of the form

- $x_{n-i-j-2}01_i0x_j$ with 2≤i≤n-2 and 0≤j≤n-i-2, or of the form
- 1_i0x_{n-i-1} with 2≤i≤n-1

In such distributions, the 1_i-groups form paraphyletic groups that (1) are delimited by taxa having state zero and (2) whose members all have state one (see fig. c.1. for some examples). Each such group with i members gives a step reduction of (i-2). For the calculation of the total step reduction in these patterns it has to be taken into account that the 01_i0 group involved in the first subpattern can appear in different positions within the string of the state distribution. This yields the following total:

- $SR_2(n) = \sum_{i=2}^{n-2} (i-2) * (n-i-1) * 2^{n-i-2} + \sum_{i=2}^{n-1} (i-2) * 2^{n-i-1} = (n-4) * 2^{n-3} + 1$

In the third type of step reduction, two groups of the first or the second pattern are separated by a $0(10)_i$-group (i≥0). Every such case gives a single supplementary step of reduction in addition to the reduction that is present in the two other patterns that are involved. Character state distributions that satisfy these conditions are of the following form:

- $x_{n-2*i-j-5}110(10)_i11x_j$ with $0 \leq 2*i \leq n-5$ and $0 \leq j \leq n-2*i-5$

As a limiting case, the monophyletic group involved can consist of only a single taxon:

- $x_{n-2*i-4}110(10)_i1$with $0 \leq 2*i < n-4$

Summed over all characters in the MIM this yields:

- $$SR_3(n) = \sum_{i=0}^{[(n-5)/2]} (n-2i-4)*2^{n-2i-5} + \sum_{i=0}^{[(n-4)/2]} 2^{n-2i-4} = \frac{1}{9}\left((3n-8)*2^{n-3} + (-1)^n\right)$$

All other patterns of 0- and 1-entries are patterns in which any two 1-entries are separated by at least two neighbouring zero-entries. As a result these patterns will never lead to step reduction, and S(n) is obtained as $S_{MAX}(n)$ - $SR_1(n)$ - $SR_2(n)$ - $SR_3(n)$:

- $$\boxed{S(n) = \frac{1}{9}\left(2^n*(3n+1) - (-1)^n\right) - (n+1)}$$

C.3 G(n)

For every character A_i with i ≤ [n/2], the maximal number of steps equals i. For every character A_i with i > [n/2], the maximal number of steps equals n-i+1 (since only n-i zeros are present, there are maximally n-i+1 clusters of ones that can be separated by these zeros). Summation over all possible informative characters gives:

- $$G(n) = \sum_{i=2}^{[n/2]} \binom{n}{i}*i + \sum_{i=[n/2]+1}^{n-1} \binom{n}{i}*(n-i+1)$$

Since $\binom{n}{i} = \binom{n}{n-i}$ this can be expressed as:

- $$G(n) = \sum_{i=2}^{[n/2]} \binom{n}{i}*i + \sum_{i=n-[n/2]-1}^{n-(n-1)} \binom{n}{i}*(n-(n-i)+1)$$ or

- $$G(n) = \sum_{i=2}^{[n/2]} \binom{n}{i}*i + \sum_{i=1}^{n-[n/2]-1} \binom{n}{i}*(i+1)$$

This equation is equivalent to the following pair of equations:

- $$\begin{cases} G(n_{even}) = \sum_{i=2}^{n/2} \binom{n+1}{i}*i \\ G(n_{odd}) = \sum_{i=2}^{(n-1)/2} \binom{n}{i}*(2i+1) + 2n \end{cases}$$

which can be expressed as:

- $$\begin{cases} G(n_{even}) = (n+1) * (2^{n-1} - 1) - \frac{n+1}{2} * \binom{n}{n/2} \\ G(n_{odd}) = (n+1) * (2^{n-1} - 1) - n * \binom{n-1}{(n-1)/2} \end{cases}$$

Combining both equations ultimately yields the following expression:

- $$\boxed{G(n) = (n+1) * (2^{n-1} - 1) - \frac{n+1}{2} * \binom{n}{[(n+1)/2]}}$$

SAMENVATTING

Inleiding

Het centrale thema van dit proefschrift is cladisme of spaarzaamheidsanalyse, een methode voor fylogenetische analyse die haar wortels heeft in het theoretisch werk van de Duitse entomoloog Willi Hennig (1913-1976). Na een bescheiden start in de jaren zestig en een periode van exponentiële groei in de jaren zeventig en tachtig is spaarzaamheidsanalyse momenteel een basisprocedure voor de fylogenetische interpretatie van systematische gegevens. Naast de ontwikkeling van de numerische technieken voor fylogenetische analyse, is een andere belangrijke verdienste van het cladisme dat het de groei gestimuleerd heeft van een theoretisch kader en een terminologie die toelaten om op een nauwkeurige wijze te denken en te praten over fylogenetische verwantschappen. Voor een Nederlandse inleiding tot de voornaamste begrippen en termen verwijs ik naar het eerste hoofdstuk.

Momenteel bestaat er over de basisprincipes van het cladisme nagenoeg eensgezindheid (zie bv. Stewart 1993), maar dit betekent geenszins dat het theoretisch werk beëindigd is. Integendeel, oude ideeën worden constant verfijnd en nieuwe ideeën blijven opduiken. Twee hiervan, drie-item analyse (Nelson & Platnick 1991) en het wegen van kenmerken met behulp van geïmpliceerde gewichten (Goloboff 1993a), worden gedetailleerd besproken en verder uitgediept in hoofdstukken twee en drie.

De twee volgende hoofdstukken handelen over de fylogenie van de angiospermenfamilie Gentianaceae, één van de grotere families van de orde Gentianales. Recente inzichten in de fylogenie van deze orde worden in hun historische context besproken en de fylogenetische structuur van de Gentianaceae wordt behandeld aan de hand van een cladistische analyse van een morfologische gegevensmatrix.

In het laatste hoofdstuk wordt het concept van besluiteloze gegevens (*indecisive data*; Goloboff 1991a) uitgebreid tot gegevensmatrices met ontbrekende gegevens. Deze resultaten worden gebruikt om aan te tonen dat het concept van *cladistic data decisiveness* moeilijk te vatten is in eenvoudige indices.

Drie-item analyse

Drie-item analyse is een methode die enkele jaren geleden geïntroduceerd werd als een nieuwe en verbeterde vorm van spaarzaamheidsanalyse, zowel in systematiek (Nelson & Platnick 1991) als in biogeografie (Nelson & Ladiges 1991a, b). De naam van de methode verwijst naar het feit dat elke uitspraak over verwantschappen van meer dan drie items (homologe structuren in systematiek, arealen in biogeografie) ontbonden wordt in een reeks van fundamentele uitspraken die elk slechts over drie items handelen. Een dergelijke fundamentele uitspraak zegt welke twee van de drie items nauwer met elkaar verwant zijn dan elk van beide met de derde (zie fig. 1 voor enkele voorbeelden).

	STANDAARD SPAARZAAMHEIDSANALYSE		
kenmerken		a	b
taxa	A	0	0
	B	0	0
	C	0	1
	D	1	1
	E	1	1

	DRIE-ITEM ANALYSE	
	a	b
O	000	000000
A	0??	000???
B	?0?	???000
C	??0	11?11?
D	111	1?11?1
E	111	?11?11

Fig. 1. De voorstelling van de verspreiding van de kenmerktoestanden van twee kenmerken, a en b, over vijf taxa, A-E. Links: voorstelling in standaard spaarzaamheidsanalyse. Rechts: voorstelling in drie-item analyse; elke kolom staat voor één drie-item uitspraak; kenmerk a impliceert drie drie-item uitspraken, kenmerk b zes; de hypothetische buitengroep (O) is toegevoegd om duidelijk te maken dat toestand 0 de plesiomorfe toestand is; taxa die niet betrokken zijn in een drie-item uitspraak worden aangeduid met een vraagteken.

Volgens drie-item analyse zijn de beste cladogrammen voor een bepaalde gegevensmatrix niet de cladogrammen met de kleinste lengte (zoals in standaard spaarzaamheidsanalyse), maar de cladogrammen die het grootst aantal fundamentele uitspraken herbergen ("*accommodate*"; een cladogram herbergt een drie-item uitspraak wanneer de uitspraak geen homoplasie vereist op het cladogram). Zonder echt duidelijk te maken waarom, hoopten Nelson & Platnick (1991) dat deze cladogrammen een preciezere weergave zouden geven van de hiërarchische structuur in de gegevens dan de kortste cladogrammen.

De methode werd al snel sterk bekritiseerd vanwege meerdere tekortkomingen (Harvey 1992, Kluge 1993, 1994, Wilkinson 1994b, De Laet & Smets 1995, Farris et al. 1995) die uiteindelijk allemaal te herleiden zijn tot de volgende drie: (1) drie-item analyse veronderstelt dat de evolutie van kenmerken irreversibel verloopt; (2) drie-item analyse negeert het feit dat binnen één kenmerk niet alle uitspraken over drie taxa logisch onafhankelijk zijn; (3) drie-item analyse aanvaardt paren van fundamentele uitspraken die elkaar uitsluiten op een bepaald cladogram toch als elkaar versterkende ondersteuning voor dat cladogram.

In het tweede hoofdstuk wordt aangetoond dat geen enkele van deze drie tekortkomingen op een adequate wijze weerlegd werd door voorstanders van drie-item analyse (Nelson 1992, 1993, 1994, 1996, Nelson & Ladiges 1992, 1993, 1994, Platnick 1993). Tegelijkertijd wordt voor elk van deze tekortkomingen een oplossing voorgesteld.

Irreversibele kenmerkevolutie

In de voorbeelden van fig. 1 staat 0 voor de plesiomorfe toestand van een kenmerk en 1 voor de apomorfe. Uit die veronderstelling volgt dat enkel 0-1-1 drie-item uitspraken beschouwd moeten worden (in een 0-1-1 uitspraak bezit één taxon toestand 0 en twee taxa toestand 1). Inderdaad, met 0 als plesiomorfe toestand zegt een 0-0-1 uitspraak enkel dat twee taxa de plesiomorfe toestand behouden hebben ten opzichte van een derde taxon dat een afgeleide toestand ontwikkeld heeft. Een dergelijke uitspraak is niet-informatief met betrekking tot de cladistische verwantschappen tussen deze taxa en moet bijgevolg niet verder in overweging genomen worden (1-1-1 en 0-0-0 uitspraken zijn evenmin informatief en mogen dus eveneens weggelaten worden).

Op het eerste gezicht impliceert de beperking tot 0-1-1 uitspraken enkel dat van tevoren moet uitgemaakt worden welke toestand apomorf is en welke plesiomorf, maar bij nadere beschouwing impliceert dit ook dat de evolutie van een kenmerk irreversibel verloopt. Het is enkel onder deze sterke veronderstelling dat 0-1-1 uitspraken steeds informatief zijn en 0-0-1 uitspraken nooit. Immers, wanneer er reversie optreedt, dan zijn er per definitie gebieden in een cladogram waar een gereverteerde toestand 0 afgeleid is ten opzichte van toestand 1. Bijgevolg worden sommige 0-0-1 uitspraken informatief, terwijl sommige 0-1-1 uitspraken niet langer informatief zijn.

Beide veronderstellingen, a priori polarisatie en irreversibele kenmerkevolutie, zijn overbodig wanneer men ervan uitgaat dat een fundamentele uitspraak niet over drie, maar over vier taxa handelt: een 0-0-1-1 vier-item uitspraak waarin twee taxa de ene toestand bezitten, en twee andere de andere. Een dergelijke uitspraak is steeds cladistisch informatief, ongeacht de toestand die plesiomorf is en ongeacht het feit of de kenmerkevolutie irreversibel verloopt of niet. Alle andere types van vier-item uitspraken (0-0-0-0, 0-0-0-1, 0-1-1-1 en 1-1-1-1) zijn steeds niet-informatief en dienen dus niet beschouwd te worden. Een voorbeeld, met dezelfde taxa en kenmerken als in fig. 1 wordt gegeven in fig. 2.

	a	b
A	0	0
B	0	0
C	0	1
D	1	1
E	1	1

	a	b
A	00?	000
B	0?0	000
C	?00	11?
D	111	1?1
E	111	?11

Fig. 2. De voorstelling van de verspreiding van de kenmerktoestanden van twee kenmerken, a en b, over vijf taxa, A-E. Links: voorstelling in standaard spaarzaamheidsanalyse. Rechts: voorstelling in vier-item analyse; aangezien de kenmerken niet gepolariseerd zijn, is er geen hypothetische buitengroep aanwezig.

Kluge (1994: 408-410) presenteerde twee hypothetische gegevensmatrices die duidelijk illustreren dat drie-item analyse inderdaad in problemen komt wanneer reversies optreden. Met deze matrices wordt geïllustreerd dat een overschakeling naar vier-item uitspraken die problemen effectief verhelpt. De analyse van Kluges matrices gebeurde met behulp van het computerprogramma ViTA2 (zie appendix A), waarin een algoritme ter beschikking gesteld wordt dat het aantal geherbergde vier-item uitspraken rechtstreeks berekent vanuit de standaard voorstelling van kenmerken. Twee dergelijke algoritmes worden in detail besproken.

Logische afhankelijkheid

Het probleem van logische afhankelijkheid wordt geïllustreerd aan de hand van de vier-item matrix van fig. 2 (eenzelfde probleem doet zich voor in drie-item analyse). Zowel kenmerk a als b hebben drie verschillende vier-item uitspraken, maar in beide gevallen zijn slechts twee van de drie logisch afhankelijk. Neem bv. kenmerk a: volgens de eerste uitspraak hebben taxa A en B enerzijds en taxa D en E anderzijds een verschillende kenmerktoestand, en volgens de tweede uitspraak geldt hetzelfde voor de paren AC en DE. Uit deze beide eerste uitspraken volgt dat ook B en C enerzijds en D en E anderzijds een verschillende toestand moeten bezitten, en dit is precies de derde vier-item uitspraak van kenmerk a. In het algemeen heeft een kenmerk met zt taxa die toestand 0 hebben en ot taxa met toestand 1 (zt*(zt-1)/2)*(ot*(ot-1)/2) verschillende vier-item uitspraken, waarvan er slechts (zt-1)*(ot-1) onafhankelijk zijn. Daarnaast zijn er zt*(ot*(ot-1)/2) verschillende drie-item uitspraken waarvan er slechts zt*(ot-1) onafhankelijk zijn. Zowel voor drie- als vier-item uitspraken hangen deze aantallen dus af van de precieze aantallen taxa met toestand 0 en toestand 1.

Nelson & Platnick (1991: 363) onderkenden dit probleem en suggereerden dat een mogelijke oplossing erin bestond om aan de kenmerken verschillende gewichten toe te kennen op basis van de verhouding van hun aantal onafhankelijke uitspraken en hun totaal aantal uitspraken. Deze methode van fractioneel wegen werd verder uitgewerkt door Nelson & Ladiges (1992; zie ook 1994). Er wordt aangetoond dat deze verhouding beïnvloed wordt door de graad van de homoplasie die in een kenmerk aanwezig is en dat fractioneel wegen enkel correct werkt met gegevensmatrices die volledig vrij zijn van homoplasie; in alle andere gevallen geeft de methode een vertekend beeld. Uiteindelijk ligt de oplossing van het probleem van logische afhankelijkheid in een aanpassing van één van de algoritmes die hogerop vermeld werden.

Wederzijdse uitsluiting

Het probleem van wederzijdse uitsluiting wordt voor drie-item analyse geïllustreerd in fig. 3 (eenzelfde probleem doet zich voor in vier-item analyse), waar de verdeling van de kenmerktoestanden van één kenmerk op een cladogram aangegeven is. Dit kenmerk heeft meerdere drie-item uitspraken die op het gegeven cladogram geherbergd zijn. Neem bijvoorbeeld de uitspraak B[DE]. Zonder verwijzing naar een cladogram drukt deze de hypothese uit dat de aanwezigheid van de plesiomorfe toestand in taxon B en de apomorfe toestand in D en E een aanwijzing is dat D en E nauwer met elkaar verwant zijn dan één van hen met taxon B verwant is. Het feit dat de uitspraak geherbergd is op het cladogram van fig. 3 betekent dan dat dit cladogram deze hypothese ondersteunt. Meer bepaald is dit zo omdat het ontstaan van de afgeleide toestand kan teruggebracht worden tot knooppunt b, een voorouder van D en E die geen voorouder is van B. Taxon B heeft de plesiomorfe toestand die aanwezig was in knooppunten a en c behouden.

Dezelfde redenering kan worden toegepast op de geherbergde uitspraken A[CE], A[CD] en A[DE]. Voor A[CE] bijvoorbeeld kan de oorsprong van de afgeleide toestand teruggebracht worden tot knooppunt c. Vanuit dit knooppunt is de afgeleide toestand doorheen knooppunten a en b in taxa C en E terechtgekomen .

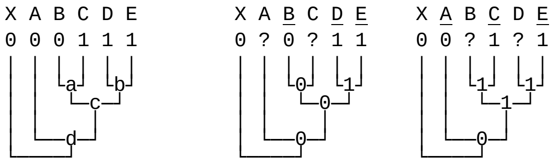
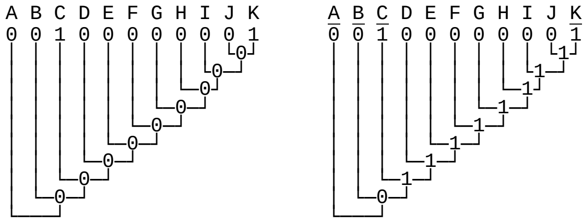


Fig. 3. Een hypothetisch kenmerk op een hypothetisch cladogram. Taxon X is de buitengroep die ter wille van drie-item analyse toegevoegd werd. Drie-item uitspraken B[DE] en A[CE] kennen verschillende toestanden toe aan knooppunten a en c (naar Farris et al. 1995).

Wanneer uitspraken B[DE] en A[CE] echter met elkaar geconfronteerd worden, doet er zich een probleem voor: om uitspraak B[DE] te verklaren door gemeenschappelijke afstamming moeten we aannemen dat knooppunten a en c de plesiomorfe toestand voor dit kenmerk behouden hebben, terwijl uitspraak A[CE] vereist dat diezelfde knooppunten voor datzelfde kenmerk de afgeleide toestand bezitten. Precies omdat beide uitspraken eenzelfde kenmerk betreffen, sluiten deze verklaringen elkaar uit: ofwel kan B[DE] verklaard worden door gemeenschappelijke afstamming, ofwel A[CE], maar niet allebei tegelijkertijd. Ook dit probleem kan opgelost worden door een verdere verfijning van de hogervermelde algoritmes.

Uiteindelijk ontstaat zo een methode waaruit de drie oorspronkelijke problemen van drie-item analyse verwijderd zijn: terwijl in drie-item analyse zoals voorgesteld door Nelson & Platnick (1991) het totaal aantal geherbergde drie-item uitspraken gemaximaliseerd wordt, betreft de maximalisatie in deze afgeleide methode enkel onafhankelijke vier-item uitspraken die mekaar niet uitsluiten. Dit kan echter tot heel tegenintuïtieve resultaten leiden zoals in de volgende voorbeelden geïllustreerd wordt.

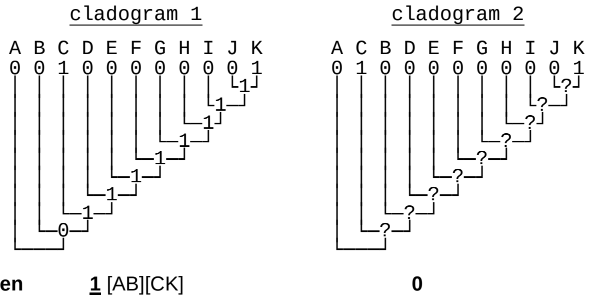


geherbergde vier-item uitspraken 0 1 [AB][CK]

Fig. 4. Een marginale verhoging van het aantal geherbergde onafhankelijke vier-item uitspraken die elkaar niet uitsluiten leidt op eenzelfde cladogram en voor eenzelfde kenmerk tot een erg tegenintuïtieve verklaring van de evolutie van het kenmerk. Zie tekst voor verdere verklaring.

In het eerste voorbeeld (fig. 4) worden voor eenzelfde cladogram en eenzelfde kenmerk twee verschillende toekenningen van toestanden aan de inwendige knooppunten vergeleken. Volgens standaard spaarzaamheidsanalyse (links) is afgeleide toestand 1 onafhankelijk ontstaan in taxa C en K; als een gevolg hiervan is geen enkele vier-item uitspraak van het kenmerk op het cladogram geherbergd. Bij maximalisatie van het aantal geherbergde vier-item uitspraken (rechts) zou de evolutie van het kenmerk op een andere wijze verklaard worden: indien C en K de afgeleide toestand overgeërfd hebben van hun meest recente gemeenschappelijke voorouder is er één uitspraak, [AB][CK], die op het cladogram geherbergd is; deze interpretatie vereist echter de veronderstelling dat toestand 0 in taxa D-J zeven maal onafhankelijk door reversie ontstaan is. Het voorbeeld kan aangepast worden door de reeks taxa tussen C en K die toestand 0 bezitten verder uit te breiden, zodat de conclusies op basis van vier-item analyse meer en meer onwaarschijnlijk worden.

In fig. 5 wordt een gelijkaardig effect geïllustreerd, maar nu voor eenzelfde kenmerk op twee verschillende cladogrammen (de twee cladogrammen verschillen in de positie van taxa B en C). Met de toekenningen aan de inwendige knooppunten zoals ze in de figuur gegeven zijn, is er één enkele vier-item uitspraak van het kenmerk geherbergd op het linker cladogram ([AB][CK]), en dit is ook de best mogelijke oplossing onder vier-item analyse: alle andere mogelijke toekenningen van toestanden aan de inwendige knooppunten leiden tot een verlies van deze geherbergde uitspraak. Ook hier leidt het herbergen van uitspraak [AB][CK] tot sterke en onwaarschijnlijke conclusies over de evolutie van het kenmerk: toestand 0 is zeven keer onafhankelijk ontstaan in de zeven evolutionaire lijnen die naar taxa D-J leiden. Op het tweede cladogram van fig. 5 kan op geen enkele wijze een vier-item uitspraak van het kenmerk geherbergd worden, en bijgevolg is de voor de hand liggende interpretatie dat toestand 1 door convergentie twee keer onafhankelijk ontstaan is in taxa C en K niet in tegenspraak met een vier-item analyse van dit cladogram. Globaal genomen is het eerste cladogram dus beter dan het tweede omdat het één vier-item uitspraak meer herbergt dan het tweede, maar tegelijkertijd impliceert het erg onrealistische veronderstellingen betreffende de evolutie van het kenmerk. Bovendien zijn dergelijke onrealistische veronderstellingen niet nodig op het tweede cladogram.



geherbergde vier-item uitspraken **1** [AB][CK] **0**

Fig. 5. Een marginale verhoging van het aantal geherbergde onafhankelijke vier-item uitspraken die elkaar niet uitsluiten beperkt de mogelijke toestanden in de inwendige knooppunten heel sterk.

Dergelijke tegenintuïtieve resultaten komen voor doordat vier-item analyse op een kunstmatige wijze een beperking legt op de maximale homoplasie die in één enkele toestand mag voorkomen, en deze beperking volgt rechtstreeks uit het ontbinden van kenmerken in fundamentele uitspraken. Het hoeft dan ook niet te verwonderen dat het een verdere aanpassing van de methode ertoe leidt dat deze gereduceerd wordt tot standaard spaarzaamheidsanalyse.

Wegen van kenmerken op basis van hun homoplasie

Het gewicht van een kenmerk kan gedefinieerd worden als “*the numerical change in the parsimony criterion produced by adding one step in that character, and weight is intended to reflect the importance of a step as evidence on phylogenetic relationships*" (Farris 1990: 92). Het gewicht van een kenmerk is dus een maat voor de betrouwbaarheid van het kenmerk bij het achterhalen van fylogenetische relaties. Spaarzaamheidsanalyse sluit het verschillend wegen van kenmerken inderdaad niet uit, en de algemeen verspreide praktijk om aan alle kenmerken in een gegevensmatrix hetzelfde gewicht toe te kennen houdt net zozeer een beslissing in betreffende het wegen van kenmerken als het toekennen van verschillende gewichten.

Het toekennen van verschillende gewichten wordt soms slechts als een middel beschouwd tot verdere selectie van cladogrammen wanneer onder gelijk wegen meer dan één cladogram bekomen wordt (zie bv. Rodrigo 1992, Sharkey 1993: 212, Sang 1995, Turner 1995, Turner & Zandee 1995). Volgens deze visie zou het toekennen van verschillende gewichten overbodig zijn wanneer met gelijke gewichten maar één enkel cladogram bekomen wordt. Dit is echter een erg onlogische houding: wanneer er inderdaad goede redenen zijn om aan te nemen dat een bepaald kenmerk een hoger gewicht verdient dan een ander, dan dienen deze verschillende gewichten van in het begin in overweging genomen te worden, ook wanneer onder gelijke gewichten slechts één enkel cladogram bekomen wordt, of wanneer met deze verschillende gewichten meer of andere cladogrammen bekomen worden (zie bv. Farris 1983: 10-11, Carpenter 1988: 293, 1994: 216, Rodrigo 1989: 101-102, Goloboff 1993a: 83, 1995).

In het verleden werden reeds heel wat verschillende methoden voorgesteld om gewichten toe te kennen aan kenmerken. Goloboff (1993a) deelde deze methoden in volgens het achterliggende principe waarmee de betrouwbaarheid van de kenmerken bepaald wordt. Op basis hiervan maakte hij een onderscheid tussen a priori wegen, wegen op basis van compatibiliteit, en wegen op basis van homoplasie.

Bij a priori wegen worden de gewichten vastgelegd vooraleer de cladistische analyse in de strikte zin begint. In het algemeen kan men stellen dat kenmerken die beter bestudeerd werden een hoger gewicht verdienen dan kenmerken die slecht gekend zijn (Neff 1986). Dit is een eenvoudige vaststelling, maar in de praktijk is het niet evident om vast te stellen hoe goed een kenmerk bestudeerd is, en nog minder om dit om te zetten in een gewicht. Vanuit een statistisch gezichtspunt bestaat er een lineaire relatie tussen het a priori gewicht van een kenmerk en de negatieve logaritme van eenvoudige functies van de waarschijnlijkheden waarmee er in het kenmerk toestandsveranderingen optreden, tenminste wanneer deze waarschijnlijkheden niet te groot zijn; de precieze relatie tussen gewichten en waarschijnlijkheden hangt af van de onderliggende evolutionaire modellen die gebruikt worden (zie bv. Farris 1978, Felsenstein 1981, Albert et al. 1993: 755-756). Dit verband lost het probleem om gewichten te bepalen echter niet op, maar verschuift het naar een ander niveau: het schatten van de waarschijnlijkheden en de keuze en verantwoording van de gehanteerde evolutionaire modellen.

In de twee andere benaderingen, wegen op basis van compatibiliteit en wegen op basis van homoplasie, worden de gewichten van de kenmerken in een gegevensmatrix bepaald aan de hand van informatie die in de matrix zelf aanwezig is. Dit kan dus probleemloos gecombineerd worden met het gebruik van verschillende a priori gewichten. De basisidee is in beide gevallen dat de betrouwbaarheid van een kenmerk bepaald wordt door de mate waarin het kenmerk overeenstemt met het hiërarchisch patroon dat in de matrix aanwezig is. Dit idee werd door Patterson (1982: 44) als volgt kernachtig uitgedrukt: “*we do not need to weight homologies: they weight themselves*”. De precieze manier waarop de overeenstemming met de globale hiërarchische structuur gemeten wordt verschilt in beide benaderingen. Bij wegen volgens compatibiliteit (zie bv. Sharkey 1994) wordt het gewicht van een kenmerk afgeleid van het aantal incompatibiliteiten (Le Quesne 1969, 1983) die het kenmerk vertoont met de andere kenmerken uit de gegevensmatrix, en een dergelijke schatting staat volledig los van concrete cladogrammen. Bij wegen volgens homoplasie gebeurt de schatting op basis van de homoplasie die het kenmerk vertoont op één of meerdere cladogrammen die tijdens het weegproces geselecteerd worden.

Aangezien het gewicht van een kenmerk aangeeft in welke mate de verspreiding van de kenmerktoestanden de fylogenetische verwantschappen weerspiegelt en aangezien een cladogram een hypothese is voor fylogenetische verwantschappen, is het wegen op basis van homoplasie te verkiezen boven het wegen op basis van compatibiliteit. Hierbij stelt zich dan wel het volgende probleem: om de gewichten van de kenmerken te schatten moet het vertakkingspatroon dat door de gegevensmatrix geïmpliceerd wordt reeds gekend zijn, maar om dat vertakkingspatroon te berekenen moet men anderzijds reeds over de gewichten beschikken. Successief wegen (Farris 1969) en wegen met geïmpliceerde gewichten

(Goloboff 1993a) zijn twee methoden die dit probleem proberen op te lossen zonder in cirkelredeneringen te vervallen.

De methode van successief wegen (Farris 1969) gaat als volgt te werk. Aan de kenmerken van een gegevensmatrix worden begingewichten toegekend en de meest spaarzame cladogrammen onder deze gewichten worden bepaald. Vervolgens wordt nagegaan in welke mate de kenmerken in overeenstemming zijn met deze cladogrammen, en de gewichten worden op basis hiervan herberekend: hoe meer homoplasie een kenmerk bezit, des te lager wordt zijn nieuw gewicht. Farris probeerde meerdere functies van de homoplasie uit als weegfunctie en op basis van simulaties verkoos hij uiteindelijk concaaf dalende functies zoals bijvoorbeeld de consistentie-index. In een volgende stap van de weegprocedure worden dan de meest spaarzame cladogrammen onder deze nieuwe gewichten bepaald. Deze cyclus van gewichten herberekenen en meest spaarzame cladogrammen bepalen wordt herhaald totdat de gewichten niet langer veranderen. De meest spaarzame cladogrammen onder deze gewichten worden dan aanvaard als de beste cladogrammen voor de gegevensmatrix.

Het probleem is echter dat enerzijds kenmerken met een hoog gewicht op sommige meest spaarzame cladogrammen toch veel homoplasie kunnen hebben en dat anderzijds kenmerken met een laag gewicht toch weinig homoplasie kunnen vertonen. Goloboff (1993a) stelde daarom voor om de achterliggende logica van de methode op een consequente wijze verder door te trekken: elk individueel cladogram staat voor een hypothese van evolutionaire verwantschapen, en om deze hypothese te evalueren mag men geen gewichten hanteren die uit andere hypothesen volgen (dit zou neerkomen op een cirkelredenering). Daarom impliceert elk cladogram zijn eigen unieke verzameling van kenmerkgewichten, en het zijn deze gewichten die moeten gebruikt worden bij de evaluatie van dat cladogram. Goloboff noemde deze gewichten geïmpliceerde gewichten (“*implied weights*”) of kenmerkgepastheden (“*character fits*”). Aangezien deze gewichten voor elk cladogram onmiddellijk kunnen berekend worden, is de methode van Goloboff in tegenstelling tot successief wegen niet-iteratief. De beste cladogrammen zijn volgens deze methode van geïmpliceerde gewichten de cladogrammen waar de kenmerken gemiddeld genomen de hoogste gewichten hebben en dus maximaal betrouwbaar zijn.

Het volgende voorbeeld illustreert deze denkwijze. Veronderstel dat er voor een bepaalde gegevensmatrix twee cladogrammen zijn waarvoor alle kenmerken op twee na dezelfde homoplasie bezitten. Het eerste kenmerk heeft één stap homoplasie op het eerste cladogram en twee stappen op het tweede, terwijl het tweede kenmerk vijftien stappen homoplasie heeft op het eerste cladogram en veertien op het tweede.

De som van de homoplasie in de twee kenmerken is dus dezelfde voor beide cladogrammen. In standaard spaarzaamheidsanalyse worden beide cladogrammen daarom als even goede verklaringen voor de gegevens beschouwd. Men kan echter argumenteren dat eenzelfde verschil in homoplasie (één stap voor beide kenmerken) belangrijker is in het eerste kenmerk omdat dit kenmerk vrij goed overeenstemt met de hiërarchische structuur die door de volledige matrix bepaald wordt. Daarom verdient dit kenmerk een hoog gewicht. Het tweede kenmerk daarentegen is op beide cladogrammen heel slecht in overeenstemming met de overige kenmerken en krijgt daarom een laag gewicht. In combinatie leidt dit tot een lichte voorkeur voor het eerste cladogram.

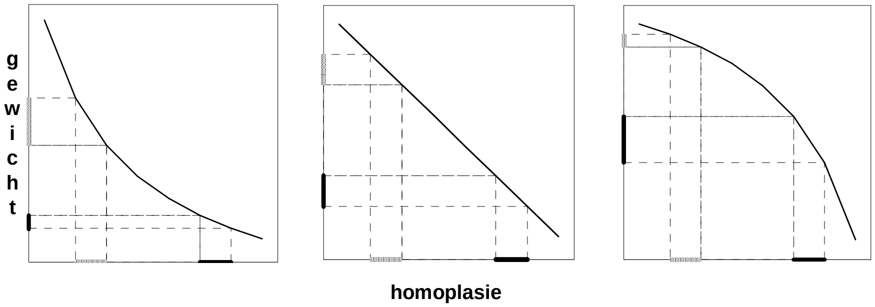


Fig. 6. Geïmpliceerde gewichten als concaaf (links), lineair (midden) of convex (rechts) dalende functies van de homoplasie (naar Goloboff 1993a, fig. 1). Zie tekst voor verdere verklaring.

Bij successief wegen verkoos Farris (1969) concaaf dalende weegfuncties omdat deze in simulaties de beste resultaten gaven. Wanneer nu de betrouwbaarheid van kenmerken gemaximaliseerd wordt, treden dergelijke functies op een natuurlijke wijze op: het zijn immers de enige functies die ertoe leiden dat bij een conflict tussen twee kenmerken de voorrang gegeven wordt aan het kenmerk met de minste homoplasie. Dit wordt geïllustreerd in fig. 6. Bij een concaaf dalende weegfunctie (fig. 6, links) zal het toevoegen van een extra stap homoplasie aan een kenmerk een des te grotere daling van het gewicht veroorzaken naarmate het kenmerk minder homoplasie heeft. Het toevoegen van een extra stap homoplasie aan een kenmerk met weinig homoplasie zal het gemiddeld gewicht van de kenmerken dus sterker doen dalen dan het toevoegen van een extra stap aan een kenmerk dat reeds veel homoplasie bezit, en op die wijze zullen extra stappen homoplasie indien mogelijk toegevoegd worden aan kenmerken die reeds veel homoplasie bezitten. Wanneer de weegfunctie lineair dalend verloopt (fig. 6, midden), is het effect van het toevoegen van een extra stap homoplasie in een kenmerk onafhankelijk van de homoplasie die reeds aanwezig is. Dit komt dus neer op standaard spaarzaamheidsanalyse. Convex dalende functies tenslotte (fig. 6, rechts) leiden tot de absurde situatie dat kenmerkconflicten opgelost worden in het voordeel van de meest onbetrouwbare kenmerken.

Eénmaal het duidelijk is dat een goede weegfunctie een concaaf dalend verloop moet hebben, stelt zich de vraag welk type concave functie de voorkeur verdient. Goloboff (1993a) verwierp de consistentie-index en de herschaalde retentie-index (het product van de retentie- en de consistentie-index) omdat deze niet alleen beïnvloed worden door de homoplasie, maar tevens door de geobserveerde variatie (m) en de informatieve variatie (g-m). Uiteindelijk verkoos hij een hyperbolische functie die uitsluitend afhangt van de homoplasie h en een constante K die de concaviteitsconstante genoemd wordt: gewicht = K/(K+h). In fig. 7 wordt deze functie getoond voor enkele verschillende waarden van deze constante. De rol van de concaviteitsconstante K is als volgt: hoe lager K, des te sterker het differentieel wegen in de zin dat (1) dezelfde hoeveelheid homoplasie een sterkere verlaging van het gewicht als gevolg heeft en (2) de totale verlaging van het gewicht van een kenmerk ten gevolge van eender welke hoeveelheid homoplasie meer en meer geconcentreerd is in de eerste stap homoplasie. Voor hoge K-waarden benadert de weegfunctie een dalende rechte, en de maximalisatie van de geïmplceerde gewichten zal bijgevolg meer en meer dezelfde resultaten geven als standaard spaarzaamheidsanalyse.

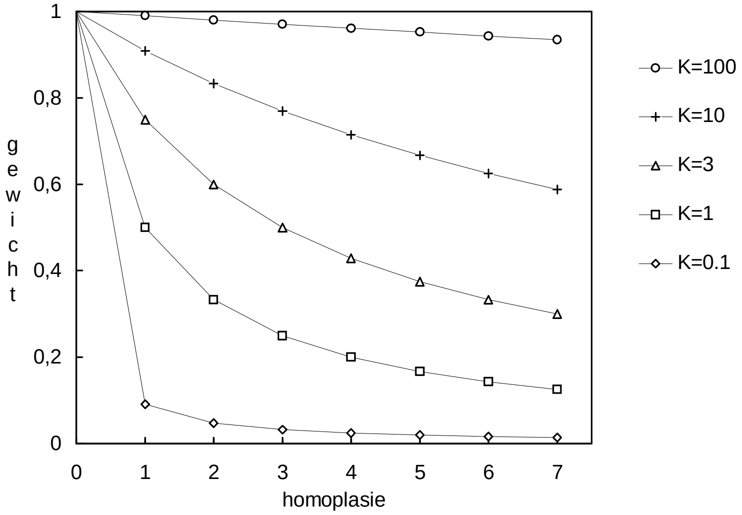


Fig. 7. De hyperbolische weegfunctie K/(K+homoplasie) voor enkele verschillende waarden van de concaviteitsconstante K.

Goloboff (1993a) ging ervan uit dat zijn benadering in overeenstemming was met de basisprincipes van cladistische analyse, maar de cladogrammen die de beste hiërarchische verklaring geven voor een gegevensmatrix zijn volgens deze basisprincipes (Farris 1983) de cladogrammen met de laagste gewogen homoplasie, en dit zijn niet noodzakelijk de cladogrammen met het hoogste gemiddeld kenmerkgewicht, zoals aangetoond wordt. Wanneer de geïmpliceerde gewichten effectief als gewichten gebruikt worden om de gewogen homoplasie te minimaliseren, zijn concaaf dalende functies niet langer de enige functies die kenmerkconflicten ten voordele van betrouwbare kenmerken oplossen. Meer zelfs, er bestaan dan zelfs concaaf dalende functies die conflicten oplossen in het voordeel van onbetrouwbare kenmerken. De enige vereiste waaraan een goede weegfunctie moet voldoen is dat de gewogen homoplasie een convex stijgend verloop heeft (fig. 8, links). Vergelijkbaar met de situatie in fig. 6 komt een lineair stijgende gewogen homoplasie (fig. 8, midden) neer op standaard spaarzaamheidsanalyse, en bij een concaaf stijgende gewogen homoplasie (fig. 8, rechts) worden conflicten opgelost ten voordele van onbetrouwbare kenmerken.

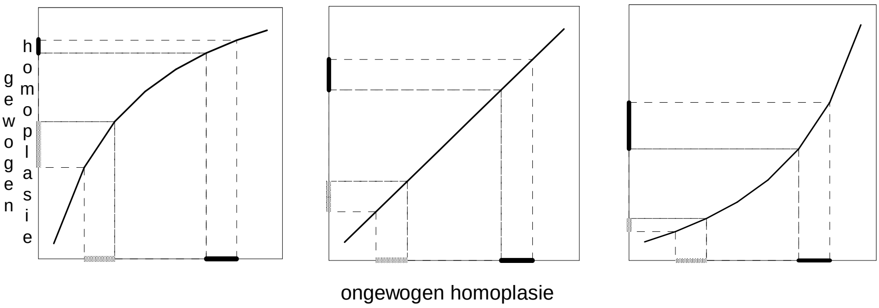


Fig. 8. Gewogen homoplasie als een convexe (links), lineaire (midden) of concave (rechts) stijgende functie van de homoplasie. Enkel weegfucties die resulteren in een convex stijgende gewogen homoplasie lossen conflicten tussen kenmerken op ten voordele van betrouwbare kenmerken.

De vraag of de weegfunctie zelf convex, lineair of concaaf dalend moet zijn is hiermee weer geopend. Een mogelijk argument voor concaaf dalende weegfuncties is reeds hogerop vermeld: meerdere evolutionaire modellen voorspellen dat het gewicht van een kenmerk zich gedraagt als de negatieve logaritme van eenvoudige functies van de waarschijnlijkheden voor de toestandsveranderingen van het kenmerk (Farris 1978, Felsenstein 1981).

Vanuit dit standpunt kan men hyperbolische functies beschouwen als benaderingen van dergelijke negatieve logaritmische functies. Dit wordt geïllustreerd in fig. 9. In de getoonde logaritmische functie, -ln((1+h)/C), wordt de homoplasie van een kenmerk gebruikt om een schatting te maken van de hogervermelde waarschijnlijkheden: hoe meer homoplasie, hoe hoger de waarschijnlijkheid dat er in het kenmerk een toestandsverandering optreedt; de constante C fungeert hierbij als een concaviteitsconstante. Om de vergelijking tussen de logaritmische en de hyperbolische weegfuncties te vergemakkelijken, werden de logaritmische waarden in fig. 9 herschaald zodat afwezigheid van homoplasie overeenkomt met gewicht 1.

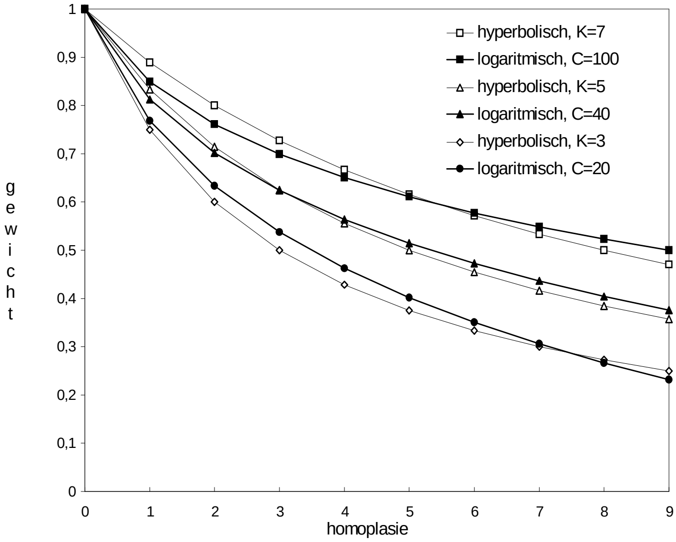


Fig. 9: Hyperbolische en logaritmische weegfuncties.

Er wordt aangetoond dat voor hyperbolische weegfuncties het maximaliseren van het gemiddelde kenmerkgewicht en het minimaliseren van de gewogen homoplasie equivalent zijn. Deze eigenschap geldt niet voor logaritmische weegfuncties.

In een volgende punt wordt besproken hoe geïmpliceerde gewichten kunnen gebruikt worden om bij ongeordende meertoestandskenmerken differentiële weging van de verschillende transformaties te bekomen. Een voor de hand liggende toepassing ligt in de analyse van nucleotidesequenties, waar transversies en transities vaak een verschillende waarschijnlijkheid bezitten.

In standaard spaarzaamheidsanalyse zijn de kortste bomen tevens de bomen met de laagste homoplasie en de bomen die de maximale hoeveelheid onafhankelijke paarsgewijze similariteiten behouden. Wanneer de homoplasie gewogen wordt met geïmpliceerde gewichten geldt deze equivalentie echter niet langer. Dit geeft aanleiding tot een aantal varianten op het gebruik van geïmpliceerde gewichten. Eén van deze varianten, complex wegen genoemd, is tevens gevoelig aan de informatieve variatie die in een kenmerk aanwezig is. Dit betekent dat bij gelijke homoplasie een kenmerk met meer informatieve variatie als meer betrouwbaar beschouwd wordt dan een kenmerk met minder informatieve variatie. Er worden twee voorbeelden van complex wegen besproken. In het eerste voorbeeld wordt aangetoond dat besluiteloze gegevens (“*indecisive data*”, Goloboff 1991a, zie verder) onder complex wegen niet langer besluiteloos zijn. In het tweede voorbeeld wordt een morfologische gegevensmatrix voor de Gentianaceae (zie verder) geanalyseerd. Deze analyse werd uitgevoerd met behulp van ViTA, een nieuw computerprogramma voor spaarzaamheidsanalyse dat in appendix A uitvoerig besproken wordt. Voor tussenliggende K-waarden wijken de resultaten onder complex wegen in enkele opvallende details af van de cladogrammen die bekomen worden met hogere en lagere K-waarden en van de resultaten die bekomen worden met standaard spaarzaamheidsanalyse en maximalisatie van de kenmerkgewichten. Enkele mogelijk oorzaken worden besproken.

Fylogenie van de Gentianaceae

De Gentianaceae zijn een cosmopolitische familie van gemiddelde grootte (76 genera volgens Brummitt 1992 en ongeveer 1200 species volgens Mabberley 1990). Recente cladistische analyses van de sequentie van het *rbc*L gen (Olmstead et al. 1993, Bremer et al. 1994), van restrictieplaats variatie in het chloroplastgenoom (Downie & Palmer 1992) en van morfologische, anatomische, embryologische en chemische gegevens (Struwe et al. 1994) wijzen er allemaal op dat deze familie één van de grotere families is van de monofyletische orde Gentianales. Hoewel er een consensus groeit dat de Gentianales inderdaad monofyletisch zijn, is er wat betreft de interne fylogenetische structuur van de orde nog veel werk te verrichten, en dit zowel wat betreft de relaties tussen de families (met inbegrip van de kleinere families Dialypetalanthaceae en Saccifoliaceae) als wat betreft de interne structuur en afbakening van de grotere families (Loganiaceae, Apocynaceae, Asclepiadaceae en Rubiaceae).

De meeste hogervermelde cladistische analyses bevestigen dat de Loganiaceae sensu Leeuwenberg & Leenhouts (1980) niet monofyletisch zijn: sommige genera zijn helemaal niet met de Gentianales verwant terwijl de resterende genera een parafyletische groep vormen aan de basis van de Gentianales. Zo werden bijvoorbeeld de genera *Potalia* Aubl., *Fagraea* Thunb. and *Anthocleista* Afzel. ex R. Br. (tribus Potalieeae van de Loganiaceae sensu Leeuwenberg & Leenhouts 1980) recent door Struwe et al. (1994) formeel getransfereerd naar de Gentianaceae.

De meest recente monografische studie die de volledige familie Gentianaceae behandelde is momenteel reeds meer dan een eeuw oud (Gilg 1895). Gilg (1895) onderscheidde twee subfamilies, Gentianoideae and Menyanthoideae. Subfamilie Menyanthoideae werd door Wagenitz (1964) op van basis anatomische, embryologische en fytochemische kenmerken als een aparte familie Menyanthaceae erkend. De algemene bloemmorfologie en vegetatieve morfologie wijzen op een verwantschap met Solanales of Gentianales (Cronquist 1981), maar zowel de sequentie van het *rbc*L gen (Chase et al. 1993, Olmstead et al. 1992, 1993) als restrictieplaats variatie van het chloroplastgenoom (Downie & Palmer 1992) associëren de Menyanthaceae met Campanulales/Asterales. De Gentianaceae zoals we ze nu kennen komen dus overeen met subfamilie Gentianoideae van Gilg. Binnen deze subfamilie onderscheidde hij vijf tribus: Gentianeae (met subtribus Exacinae, Erythraeinae, Chironiinae, Gentianinae and Tachiinae), Rusbyantheae, Helieae, Voyrieae and Leiphaimeae. De classificatie van Gilg was nagenoeg uitsluitend gebaseerd op pollenkenmerken, en het hoeft dan ook niet te verwonderen dat er op basis van andere gegevens heel wat wijzigingen voorgesteld zijn. Naast de status en de positie van de Menyanthoideae waren de voornaamste punten van kritiek de status van de (sub)tribus Rusbyantheae, Helieae, Voyrieae, Leiphaimeae and Tachiinae. Maas (1984a) merkte op dat het neotropische genus *Lisianthius* P. Browne en een aantal verwante genera (de "*lisanthoid gentians*", Sytsma 1988) verspreid zijn over Helieae, Tachiinae and Rusbyantheae. Momenteel is het duidelijk dat *Rusbyanthus cinchonifolius* Gilg, de enige soort van tribus Rusbyantheae, tot het genus *Macrocarpaea* Gilg (Tachiinae) behoort (Weaver 1974, Maas 1984b). Het genus *Voyriella* Miq. (Leiphaimeae) toont verwantschappen met de genera *Curtia* Cham. & Schltdl. and *Tapeinostemon* Benth. (Erythraeinae), terwijl *Leiphaimos* Cham. & Schltdl. (het tweede genus van Gilgs Leiphaimeae) nu in *Voyria* Aubl. opgenomen is (Weaver 1974, Maas & Ruyters 1986). De tribus Rusbyantheae en Leiphaimeae van Gilg (1895) zijn dus overbodig (Weaver 1974).

De voorgestelde cladistische analyse van de Gentianaceae is een uitbreiding van de studie van Mészáros (1994). De uitbreiding betreft zowel het aantal kenmerken (zie tabellen 5.2. en 5.3. in hoofdstuk vijf) als het aantal taxa. De 21 genera die opgenomen werden vertegenwoordigen alle (sub)tribus van Gilg behalve Leiphaimeae, Rusbyantheae and Voyrieae (een tribus met één enkel mycotroof geslacht, *Voyria* Aubl.). De genera *Anthocleista* and *Fagraea* (Loganiaceae sensu Leeuwenberg & Leenhouts 1980) werden opgenomen als buitengroep.

De gegevensmatrix werd voornamelijk aan de hand van literatuurgegevens opgesteld, en geanalyseerd onder een brede waaier van veronderstellingen betreffende a priori gewichten en ordening van kenmerken. Dit gebeurde zowel volgens standaard spaarzaamheid als volgens Goloboffs (1993a) methode van geïmpliceerde gewichten (met verschillende waarden voor de concaviteitsconstante K). Al deze verschillende analyses resulteerden in cladogrammen die congruent waren wat betreft de globale verwantschappen in de familie. Als voorbeeld wordt in fig. 10 het strikte consensuscladogram getoond van de acht meest spaarzame bomen die bekomen werden onder standaard spaarzaamheidsanalyse met alle kenmerken ongeordend (lengte 111; CI 0,51; RI 0,64).

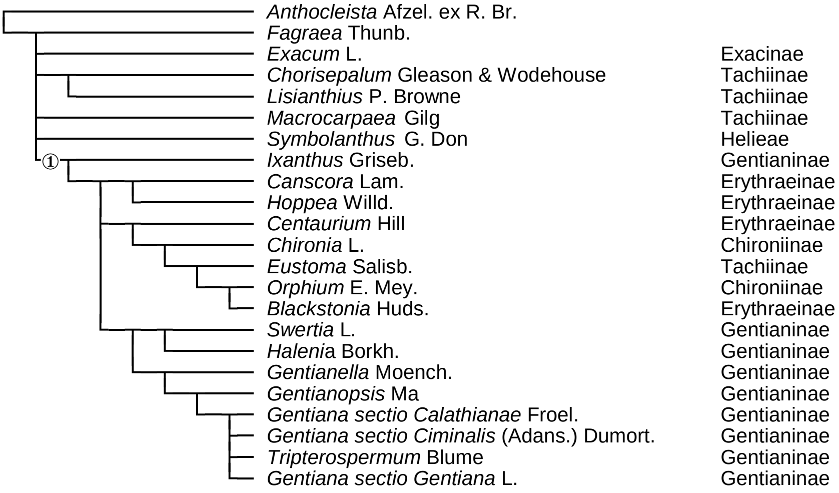

Anthocleista Afzel. ex R. Br.
Fagraea Thunb.
Exacum L. Exacinae
Chorisepalum Gleason & Wodehouse Tachiinae
Lisianthius P. Browne Tachiinae
Macrocarpaea Gilg Tachiinae
Symbolanthus G. Don Helieae
①
Ixanthus Griseb. Gentianinae
Canscora Lam. Erythraeinae
Hoppea Willd. Erythraeinae
Centaurium Hill Erythraeinae
Chironia L. Chironiinae
Eustoma Salisb. Tachiinae
Orphium E. Mey. Chironiinae
Blackstonia Huds. Erythraeinae
Swertia L. Gentianinae
Halenia Borkh. Gentianinae
Gentianella Moench. Gentianinae
Gentianopsis Ma Gentianinae
Gentiana sectio Calathianae Froel. Gentianinae
Gentiana sectio Ciminalis (Adans.) Dumort. Gentianinae
Tripterospermum Blume Gentianinae
Gentiana sectio Gentiana L. Gentianinae

Fig. 10. Strikt consensuscladogram van de acht meest spaarzame cladogrammen onder standaard spaarzaamheidsanalyse met alle kenmerken ongeordend.

De best ondersteunde groep (groep 1 in fig. 10) omvat *Eustoma* (Tachiinae) en alle opgenomen Gentianinae, Erythraeinae and Chironiinae (in sommige analyses kwam ook *Exacum* in deze groep terecht). De basale vertakking in deze groep is tussen *Ixanthus* enerzijds en alle andere genera anderzijds. Zo vormt *Ixanthus*, een endeem van de Canarische Eilanden, een overgang tussen de voornamelijk houtige en tropische genera aan de basis van het cladogram en de overige genera uit groep 1, voornamelijk kruidachtige vertegenwoordigers uit de gematigde streken. Binnen groep 1 is subtribus Gentianinae met uitsluiting van *Ixanthus* eveneens monofyletisch, maar Erythraeinae en Chironiinae niet. In de meeste analyses verschenen beide subtribus echter samen met *Eustoma* als zustergroep van subtribus Gentianinae, zodat Erythraeinae, Chironiinae and *Eustoma* mogelijks moeten samengevoegd worden. De basale verwantschappen waren in nagenoeg alle analyses onduidelijk.

Besluiteloze gegevens

Goloboff (1991a, b) definieerde de “*cladistic decisiveness*“ van een gegevensmatrix als de mate waarin alle verschillende cladogrammen voor die gegevensmatrix in lengte verschillen. Op die wijze staat de *decisiveness* van een gegevensmatrix voor de graad waarin de gegevensmatrix toelaat om een keuze tussen cladogrammen te maken. De DD-index werd voorgesteld om *data decisiveness* van een gegevensmatrix effectief te meten.

Hiernaast definieerde Goloboff eveneens besluiteloze gegevensmatrices (“*indecisive data sets*”). Dit zijn gegevensmatrices die op elk mogelijk cladogram dezelfde lengte hebben. Goloboff beperkte zijn bespreking tot gegevensmatrices waarin geen ontbrekende gegevens voorkomen en die uitsluitend binaire kenmerken bevatten. In dat geval bestaat er voor een bepaald aantal taxa in essentie maar één enkele besluiteloze gegevensmatrix. Goloboffs formule voor de lengte hiervan is moeilijk te berekenen en niet helemaal exact. Daarom werd de volgende alternatieve formule afgeleid voor de lengte van een besluiteloze matrix voor n taxa:

$$S(n) = \frac{1}{9}\left(2^{n} * (3n+1) - (-1)^{n}\right) - (n+1)$$

Ook Goloboffs formules (een eerste voor een even aantal taxa, een tweede voor een oneven aantal) voor G(n), de lengte van een besluiteloze matrix voor n taxa op een volledig onopgelost cladogram zijn nodeloos ingewikkeld. Een eenvoudiger formule voor G(n) is als volgt:

$$G(n) = (n+1) * (2^{n-1} - 1) - \frac{n+1}{2} * \binom{n}{\left[(n+1)/2\right]}$$

In deze formule staat $\binom{n}{i}$, met 0=<i=<n, voor n!/(i!*(n-i!)) en de vierkante haken voor het gehele deel van de uitdrukking tussen de haakjes.

Vervolgens wordt aangetoond hoe het concept van besluiteloze gegevens kan uitgebreid worden tot matrices waarin ook ontbrekende gegevens voorkomen. Met behulp van dergelijke matrices worden tenslotte een aantal hypothetische gegevensmatrices geconstrueerd die illustreren dat het concept van “*data decisiveness*” moeilijk te vatten is in eenvoudige indices zoals Goloboffs DD-index.