Available online at www.sciencedirect.com
SCIENCE DIRECT®
Organisms, Diversity & Evolution 4 (2004) 35–49
ORGANISMS DIVERSITY & EVOLUTION
www.elsevier.de/ode

The information content of an ambiguously alignable region, a case study of the *trnL* intron from the Rhamnaceae

Lone Aagesen*

Instituto de Botánica Darwinion, Casilla de Correo 22, 1642 San Isidro, Argentina

Received 2 September 2003; accepted 11 November 2003

Abstract

An earlier analysis of the *trnL* intron in the Colletieae (Rhamnaceae) showed polyphyly of the genus *Discaria*. Polyphyly of *Discaria* is supported only by an AT-rich region of ambiguous alignment within the *trnL* intron. Polyphyly of the genus relies on extracting the information of the AT-rich region correctly. Ambiguously aligned regions are commonly excluded from phylogenetic analysis. In the present study the question was raised whether random or noisy data could generate a pattern like the one found in the AT-rich region of ambiguous alignment. The original pattern was resistant to changes in alignment parameter cost when submitted to a sensitivity analysis using direct optimization. Artificially generated random or noisy data gave well-resolved trees but these were found to be extremely sensitive to changes in parameter costs. However, information from additional data, such as conserved regions, restricts the influence of random data. It is here suggested that the information in ambiguously aligned regions need not be dismissed, provided that an appropriate method that finds all possible optimal alignments is used to extract the information. In addition to commonly used support measures, some information of robustness to changes in alignment parameter costs is needed in order to make the most reliable conclusions.

Keywords: Ambiguously alignable region; Phylogenetic information; Sensitivity analysis; Random data

Introduction

Phylogenetic analysis at the species level is often problematic when it comes to finding an appropriate molecular marker, one that contains sufficient variation to obtain a resolved and well supported phylogeny. Botanists often use segments of noncoding cpDNA regions to determine interspecific relationships, since these zones tend to evolve more rapidly than do coding sequences—partly by the accumulation of insertions and deletions (Clegg and Zurawski 1992; Gielly and Taberlet 1994). However, fast-evolving sites may be difficult to align when the sequences vary considerably in length.

How to treat a 'difficult-to-align' or 'ambiguously alignable' region is not generally agreed upon. In fact, even the notion of what exactly constitutes a region too variable to be aligned is unclear. In the literature several approaches have been suggested. Swofford et al. (1996, p. 453) recommended the exclusion of all areas that include a substantial numbers of gaps, "as positional homology is too uncertain for reliable estimates to be made from these regions". Indeed, excluding areas that seem too difficult to align is a very common practice in present-day analysis. Gatesy et al. (1993) called attention to the fact that regions of excessive variation are often identified and excluded only on the basis of subjective criteria, and provided an objective method for data exclusion later termed "culling" (Wheeler et al. 1995). Lutzoni et al. (2000, p. 631) favored the inclusion of regions of ambiguous alignment, "except the ones

*Corresponding author. Permanent address: Botanical Institute, Gothersgade 140, DK-1123 Copenhagen K, Denmark.
E-mail address: aagesen@amnh.org (L. Aagesen).

1439-6092/$ - see front matter © 2004 Elsevier GmbH. All rights reserved.
doi:10.1016/j.ode.2003.11.003

where saturation caused by multiple changes has most likely resulted in the complete loss of phylogenetic signal.'' Wheeler et al. (1995), rather than excluding any region, explored the effect of different alignments on phylogenetic analysis and provided a method for the purpose called ''elision''. Exploring the effects of different alignments was also recommended by Doyle and Davis (1998).

The present study explores a typical ambiguously alignable region, an AT-rich region (hereafter abbreviated throughout as AT-RR) of the *trnL* intron. The region has appeared in a phylogenetic analysis of the tribe Colletieae in the family Rhamnaceae (L. Aagesen, D. Medan, J. Kellermann, and H. Hilger; unpublished), but the same region has been independently identified in, and excluded from, an analysis of the family Rhamnaceae itself (Richardson et al. 2000).

When the phylogenetic analysis of the Colletieae was carried out, it became evident that the areas flanking the AT-RR contain very little phylogenetic information. It was therefore important to include the phylogenetic information present in the AT-RR. The highly variable region was of interest due to its supposedly higher rate of evolution; hence, it may include phylogenetic information for resolving relationships at the species level within this tribe. That a region is of ambiguous alignment simply means that there is more than one way to align it. Therefore, when including such a region in a phylogenetic analysis the existence and influence of multiple (equally optimal) alignments has to be explored.

One possible way to explore different optimal alignments is sensitivity analysis (Wheeler 1995) using direct optimization (Wheeler 1996). Direct optimization constructs most-parsimonious phylogenetic hypotheses directly, without the intervening step of multiple sequence alignment. The shortest tree is searched for directly on the basis of the unaligned sequences. The alignment of the sequences prior to phylogenetic analysis is avoided, because insertion and deletion events are incorporated in the character optimization procedure in addition to base substitutions (Wheeler 1996). Thus, the input to an analysis using direct optimization is the unaligned sequences, whereas the output is one or more optimal trees, each implying its own unique ''optimal alignment'' (Wheeler 2003). As any alignment is affected by insertion–deletion costs and transversion–transition costs, the sensitivity approach explores the outcome of varying the cost of these parameters.

The analysis of tribe Colletieae was carried out within a parameter space of 20 different parameter sets. Within the entire parameter space the genus *Discaria* was polyphyletic with *Discaria americana*, *D. articulata*, *D. chacaye*, *D. nitida*, *D. pubescens*, and *D. toumatou* always grouping with the monotypic genus *Adolphia* (hereafter the *Adolphia-Discaria* p.p. clade), whereas *D. nana* and *D. trinervis* were placed in other clades. The topology is supported by the AT-RR of the *trnL* intron, but not by the flanking regions (hereafter abbreviated throughout as F-R), nor by morphology.

The genus *Discaria* is of considerable biogeographic interest because of its Gondwanan distribution. Concluding that at least the *trnL* intron supports polyphyly of this genus relies on the information content found in the AT-RR. This raises questions about the quality of the information found in the AT-RR. The *Adolphia-Discaria* p.p. clade is stable in the entire parameter space explored, but does this mean that an hierarchical pattern caused by phylogeny has been extracted, or could random noise also create a pattern that would be robust to parameter variation? These questions are the main concerns of the present study.

Lutzoni et al. (2000), who analyzed the inclusion versus exclusion of ambiguous regions in phylogenetic analysis and generally favored the inclusion of such areas, raised special concerns about one region rich in As and Ts (Lutzoni et al. 2000, p. 642): ''… this region is extremely AT-rich (A = 52% and T = 37%) and requires the inclusion of many gaps. Therefore, we feel justified to at least question the phylogenetic quality of the signal it provides. If this region were saturated by changes that led to an accumulation of As and Ts followed by multiple changes between As and Ts, relatively little (if any) phylogenetic signal would actually still be present in this region. This raises a fundamental question. Should such a region be included in a phylogenetic analysis even if we were able to include it without violating positional homology?''

If Lutzoni et al. (2000) are correct that multiple random hits of As and Ts within an AT-RR actually do blur the original signal, this may have consequences for the AT-RR of this study. In this light it is of interest how sequences with multiple changes between As and Ts as well as insertion and deletion events of As and Ts behave in a sensitivity approach. Are clades stable to variation in parameter choice found when analyzing noisy or random AT-RR? Furthermore, if multiple changes between As and Ts do occur within the region, what is the effect of noise, and is this possible effect more pronounced at higher taxonomic levels? Will the phylogenetic signal be blurred by such noise, and will that force us to exclude regions of ambiguous alignment from phylogenetic analyses?

In order to identify and discuss noise, some definition of noise is required. Wenzel and Sidall (1999) explored the effect of random noise in conventional parsimony analysis. The authors discussed a spectrum of 'noise', with one end being random data only forming a pattern by chance, and the other end being homoplasy that is noisy at a specific taxonomic level but informative at a different taxonomic level. For the purpose of their paper

only the effect of noise as random data was explored. This approach is followed in this paper.

The main concerns addressed here are resistance to changes in analytic parameters rather than more traditional support measures. Goloboff et al. (2003) pointed out that well supported groups will generally survive sensitivity analyses but this need not be so. The authors distinguished between methods that attempt to measure quantities directly related to the evidence itself — e.g., Bremer support and jackknifing — and methods that examine the implications of the background knowledge-like sensitivity analysis where the influence of prior knowledge of alignment costs is examined. As we have little basis for establishing a priori alignment costs, groups that depend critically on a specific cost set are poorly established while groups resistant to changes in alignment costs are more firmly established. The present paper is mainly concerned with this aspect when exploring the phylogenetic information found in the AT-RR of the *trnL* intron in the Colletieae. The surviving groups may or may not be well supported in terms of Bremer support and similar measures; however, this aspect is only briefly addressed.

All results are compared to analyses at a higher taxonomic level within the Rhamnaceae, in order to evaluate whether a possible effect of random noise can be discovered and whether this warrants the exclusion of the AT-RR in analyses at higher levels within the family.

All taxonomic considerations have been left out in the present study, partly because they are beyond its scope, and because the limited sampling at higher taxonomic levels within the family may have biased the results.

Material and methods

The *trnL* sequences used in this study were obtained from Aagesen et al. (in prep.) and Richardson et al. (2000) who kindly provided an aligned version of the sequences used in their study. Voucher information is included in Table 1.

One possible way to evaluate whether the signal found in the AT-RR of the *trnL* intron in the Colletieae could be obtained from random data is to generate various random AT-RR, analyzing each in combination with the F-R within a defined parameter space. If the signal found in the original AT-RR differs from a random one, some measure reflecting this difference is expected to be found. In the present study the number of groups stable in the entire parameter space versus the number of groups stable in only a part of the parameter space was used when exploring the effect of applying different alignment parameter costs.

First the AT-RR was delimited within the Colletieae in order to split the sequence matrix into three parts, two including the F-R and the third one including the AT-RR only. This step was necessary for manipulating the AT-RR. Gatesy et al. (1993) proposed a replicable data exclusion method discarding all nucleotide positions whose states vary with choice of alignment parameter. This method was initially used to delimit the AT-RR. In the original analysis 20 different parameter sets were explored. The explored transversion–transition costs were 1:1, 2:1, 4:1, and 1:0, the latter ratio being a transversion-only scheme (see Wheeler and Hayashi 1998). The cost of inserting gaps was varied from equal to the cost of transversions to twice, four, eight, and 16 times more costly. Implied alignments (Wheeler 2003) were generated for several trees across the parameter cost sets. In general no variation of the alignment was found within a given cost set. Consequently, one tree was selected from each of the 20 parameter cost sets. The implied alignment of the sequences was computed from each of these 20 trees. The different alignments were then compared to determine the 5′ and 3′ ends of the AT-RR.

The resulting three matrix fragments were analyzed both in combination and separately (the two F-R combined versus the AT-RR on its own), using the same parameter space as in Aagesen et al. (in prep.), with direct optimization and the program POY ver. 2.7 (Gladstein and Wheeler 2000). Twenty-one terminal taxa were included in the matrix (20 species from the Colletieae, including two samples of *Discaria nitida*, and one outgroup species, *Noltea africana* — see Table 1).

Analyses with direct optimization are computationally very demanding. To save time the following search strategy was used: -random 20 -spr -notbr -norandomizeoutgroup -seed -1 -maxtrees 2. This creates a Wagner tree and submits it to subtree pruning and regrafting (SPR) swapping, holding a maximum of two trees. The procedure was repeated 20 times using the same outgroup, with the time used as seed for the random number generator that defined the input order of the taxa during the replicates. After completing the 20 replicates all resulting trees were submitted to tree bisection and reconnection (TBR) swapping, storing all optimal trees found. This search strategy proved to be effective within a reasonable amount of time for matrices with 20–25 terminals.

The strict consensus of the optimal trees found under each parameter set was calculated in the program NONA (Goloboff 1998), and percentages of parameter sets yielding a specific group were computed using the program Jack2hen.exe (distributed together with the program POY). All analyses were run on a 500 MHz Pentium PC.

Random AT-RRs were generated to resemble the original AT-RR in base composition and sequence length variation. To achieve this, bases and gaps were

Table 1. Sources of plant material used

Species	Source	Analysis	GenBank acc. no.
Dirachmaceae			
Dirachma socotrana Schweinf.	Socotra (Thulin et al. 1998)	R	AJ225796
Urticaceae			
Boehmeria biloba Miq.	Java (Richardson et al. 2000)	R	AJ390371
Rhamnaceae			
Adolphia infesta Meisn.	USA, California, Rancho Santa Ana Botanic Garden (858)	C	AY460408
Alphitonia excelsa Reiss.	Australia (Richardson et al. 2000)	R, Z	AJ390352
Bathiorhamnus cryptophorus Capuron	Madagascar (Richardson et al. 2000)	R, Z	AJ390340
Berchemia discolor (Klotch) Hemsley	Saudi Arabia (Thulin et al. 1998)	R	AJ225793
Ceanothus coeruleus Lag.	USA (Thulin et al. 1998)	Z	AJ225798
Colletia hystrix Clos	Argentina, Neuquén, D. Medan 774 (BAA)	C, Z	AY460409
Colletia paradoxa (Spreng.) Escal.	Argentina, Buenos Aires, A. Mantese (BAA 22105)	C	AY460410
Colletia spinosissima Gmel.	Argentina, Buenos Aires, Hort. Bot. Fac. de Agronomía UBA (607)	C	AY460411
Colletia ulicina Gill. & Hook.	Chile, Colchagua, D. Medan 791 (BAA)	C	AY460412
Colletia ulicina Gill. & Hook.	Chile (Richardson et al. 2000)	R	AJ390364
Colubrina asiatica Brongn.	Sumatra (Richardson et al. 2000)	R, Z	AJ390350
Condalia microphylla Cav.	Argentina (Richardson et al. 2000)	R	AJ390334
Crumenaria erecta Reiss.	Brazil (Richardson et al. 2000)	R, Z	AJ390346
Cryptandra cf. *spyridioides* F. Muell.	W Australia (Richardson et al. 2000)	Z	AJ390360
Discaria americana Gill. & Hook.	Germany, Bot. Garten und Bot. Museum Berlin-Dahlem (048079210)	C, Z	AY460413
Discaria articulata (Phil.) Miers	Argentina, Río Negro, San Carlos de Bariloche, leg. E. Chaia, June 1997 (no herbarium voucher)	C	AY460414
Discaria chacaye (G. Don) Tort.	Argentina, Neuquén, D. Medan 775 (BAA)	C	AY460415
Discaria nana (Clos) Weberb.	Argentina, Mendoza, D. Medan 840 (BAA)	C	AY460416
Discaria nitida Tort.	Sample 1: Australia, Royal Botanic Gardens (Melbourne 915497)	C	AY460418
	Sample 2: Australia, N.H. Scarlett 80-47 (BAA)		AY460417
Discaria pubescens (Brong.) Druce	Australia, Royal Botanic Gardens (Melbourne), from wild-sourced plants at Bendock, eastern Victoria, leg. Neville Walsh, 1997	C	AY460419
Discaria toumatou Raoul	Denmark, Bot. Garden of the University of Copenhagen (P 1981-5496)	C	AY460420
Discaria trinervis (Hook. & Arn.) Reiche	Argentina, Buenos Aires, J.J. Valla (BAA 23793)	C	AY460421
Emmenosperma alphitonioides F. Muell.	Australia (Richardson et al. 2000)	R, Z	AJ390351
Gouania mauritiana Lam.	Mauritius: (Richardson et al. 2000)	R, Z	AJ390344
Helinus integrifolius Kuntze	South Africa: (Richardson et al. 2000)	R, Z	AJ390347
Hovenia dulcis Thunb.	South Korea: (Richardson et al. 2000)	R, Z	AJ390343
Kentrothamnus weddellianus (Miers) Johnst.	Argentina, Jujuy, D. Medan 777 (BAA)	C	AY460422
Lasiodiscus mildbraedii Engl.	Sao Tomé (Richardson et al. 2000)	R, Z	AJ390353
Maesopsis eminii Engl.	Australia (Richardson et al. 2000)	R	AJ390336
Nesiota elliptica (Roxb.) Hook. f.	St Helena (Thulin et al. 1998)	Z	AJ225803
Noltea africana (L.) Reichenb.	South Africa, Belmont Valley, R.D.A. Bayliss 6334 (M)	C	AY460407
Noltea africana (L.) Reichenb.	South Africa (Richardson et al. 2000)	R, Z	AJ390357
Paliurus spina-christi Mill.	Bulgaria (Richardson et al. 2000)	R, Z	AJ390354
Phylica nitida Lam.	Mauritius (Richardson et al. 2000)	Z	AJ390356
Pleuranthodes hillebrandii (Oliver) Weberb.	Hawaii (Richardson et al. 2000)	Z	AJ390348

Table 1 (*continued*)

Species	Source	Analysis	GenBank acc. no.
Pomaderris rugosa Cheeseman	W Australia (Richardson et al. 2000)	R, Z	AJ390363
Reissekia smilacina Endl.	Brazil (Richardson et al. 2000)	Z	AJ390345
Retanilla ephedra (Vent.) Brong.	Argentina, Buenos Aires, D. Medan (BAA 21960)	C	AY460423
Retanilla patagonica (Speg.) Tort.	Argentina, Neuquén, D. Medan 776 (BAA)	C	AY460424
Retanilla stricta Hook. & Arn.	Chile, Colchagua, D. Medan 790 (BAA)	C	AY460425
Retanilla trinervia (Gill. & Hook.) Hook. & Arn.	Chile, Quillota, D. Medan et al. (BAA 21957)	C	AY460426
Reynosia uncinata Urban	Cuba (Richardson et al. 2000)	R	AJ390339
Rhamnus lycioides L.	Spain (Richardson et al. 2000)	R	AJ225792
Sageretia thea (Osbeck) M.C. Johnston	Saudi Arabia (Thulin et al. 1998)	R	AJ225792
Schistocarpaea johnsonii F. Muell.	Australia (Richardson et al. 2000)	R, Z	AJ390349
Siegfriedia darwinioides C.A. Gardner	W Australia (Richardson et al. 2000)	Z	AJ390375
Spyridium cf. *forrestianum* F. Muell.	W Australia (Richardson et al. 2000)	R	AJ251690
Spyridium globulosum (Labill.) Benth.	W Australia (Richardson et al. 2000)	Z	AJ390358
Trevoa quinquenervia Gill. & Hook.	Chile, Quillota, D. Medan et al. (BAA 22003)	C	AY460427
Trymalium floribundum Steudel	W Australia (Richardson et al. 2000)	Z	AJ390362
Ventilago viminalis Hook.	W Australia (Richardson et al. 2000)	R	AJ390337
Ziziphus glabrata Heyne ex. Roth.	Saudi Arabia (Thulin et al. 1998)	Z	AJ225799

C, R, and Z refer to analyses of Colletieae, Rhamnaceae, and Ziziphoids, respectively.

replaced randomly in the original alignment. In this case the state of each cell was replaced using the percentages of A, C, G, T, and gaps in the original alignment as probabilities of a replacement with the corresponding state. In this way bases and gaps are redistributed in a random way roughly maintaining the original base and gap composition (T = 64.7%, A = 25.8%, G = 6.7%, C = 2.9%). Since the different sequences varied in length from 28 to 61 bp, the replacement model included gaps to redistribute the number of base pairs for each species. The gaps are deleted in the final step before analyzing the sequences with POY. To estimate base and gap composition the implied alignment of the original AT-RR was computed over 25 trees randomly selected from all trees found in the original analysis. Different gap costs lead to different numbers of gap insertions, and this again affects the base/gap composition. The percentage of gaps inserted in the alignments varied between 41% and 48%. An intermediate value was chosen (based on alignment with gap cost twice the cost of transversions and transversion cost twice the cost of transitions), and 20 random AT-RRs were generated. All gaps were subsequently deleted, and each region analyzed alone and together with the F-R. All replacements were done by a macro-file for NONA constructed for the purpose.

For comparative purposes and to explore the effect of adding random noise, matrices were generated replacing 10% or 25% of the cells with noise using the approach outlined above. When replacing with noise, matrices were generated from all 25 implied alignments described above, producing a total of 2×25 replicates. Replacement within each alignment was done using the corresponding base/gap compositions for this particular alignment. The modified AT-RRs were analyzed on their own and in combination with the F-R within the parameter space mentioned above. The stability of the individual clades to replacement with random noise was briefly explored in terms of robustness to variation in parameter choice and Bremer support (Bremer 1994). The Bremer support values are specific to each cost set. The support measures were calculated for some relevant clades under two different cost sets: gap cost 1, transversion cost 1, transition cost 1 (cost set 111); and gap cost 2, transversion cost 1, transition cost 1 (cost set 211).

The results were compared with those obtained from analyzing the same region at a higher taxonomic level. In the analysis of Richardson et al. (2000), Colletieae falls within the clade 'Ziziphoids'. From the latter 25 species were sampled (see Table 1), selecting species representing major clades within the Ziziphoids. The sequences obtained were cut into three fragments corresponding to the fragments used in the Colletieae. The fragments were analyzed within the present parameter space both as original data and with the AT-RR replaced with a randomly generated AT-RR. Twenty random AT-RRs were generated as outlined above.

Ultimately, 24 taxa were sampled within the entire family (Table 1), and the AT-RR was analyzed on its own and in combination with the F-R.

Results

Defining the AT-RR

When the different alignments of the AT-RR were compared, little variation of the 3′ end was found. Fig. 1 shows an alignment of the region using gap costs twice the cost of transversions, and transversions four times as costly as transitions. A gray arrow marks the 3′ end of the region of ambiguous alignment as defined according to the method of Gatesy et al. (1993). Three bp downstream from this limit were included in the ambiguously aligned region to resemble the region excluded in previous studies (Richardson et al. 2000; Aagesen et al. in prep.). More variation was found at the 5′ end of the AT-RR. Under gap costs from equal to transversions to twice and four times as costly, the same limit of the 5′ end was found when using transversion–transition costs 1:1, 2:1, and 4:1 (marked by a white arrow in Fig. 1). This limit of the AT-RR corresponds to the one used by Richardson et al. (2000) obtained by manual alignment. When transversion–transition costs of 1:0 are used, the two last base pairs of the F-R as defined in Fig. 1 (the bases TC) interfere in the alignment of the AT-RR. At gap costs higher than four the alignments furthermore differ by moving six base pairs of the outgroup (*Noltea africana*) into the AT-RR. This is due to a six-base-pair indel found at a distance of 43 bp upstream from the 3′ end of the F-R as delimited by the white arrow in Fig. 1.

The 5′ end of the AT-RR was settled on the basis of the implied alignments using gap costs one to four times the cost of transversions, including the two base pairs interfering with the alignment when using transversion–transition costs of 1:0. Expanding the area with 43 base pairs, as implied by gap costs higher than four, was not considered. The alignment implied by gap costs eight and 16 times as costly as transversions seems to be an artifact of using extreme gap costs.

Cutting sequences into fragments

Splitting the original sequences into unambiguously recognizable fragments should ideally result in better homology statements but also has the effect of introducing some subjectivity in the analysis (Giribet 2001). When constraining the alignment by cutting the sequences into fragments some ambiguities concerning primary homologies (sensu De Pinna 1991) were obviously avoided. Not surprisingly, the alignment was slightly more stable to parameter variation (Table 2). In the Colletieae four groups were found constant in the entire parameter space when the sequences were cut into fragments, as opposed to three stable groups when the sequences were analyzed as a unit. Also, the total number of groups (the sum of the groups found in each of the parameter sets-hereafter referred to as ‘total groups’) was lower (23 groups as opposed to 28 groups) when the sequences were cut into fragments (Table 2).

Noise and random data

When analyzing the sequences by replacing the AT-RR with a random AT-RR there was a very pronounced effect of destabilizing the alignment with respect to parameter variation (Table 2 and Fig. 2). Since the Colletieae matrix and the Ziziphoids matrix included a different number of terminal taxa, the numbers were scaled through dividing by the number of taxa in the matrices in order to facilitate comparisons. In the Colletieae no groups were stable in 100% or 95% of the entire parameter space, and only a single group was stable in 90% of the 20-parameter sets in one of the 20 replicates. On the contrary, there were about 100 times more groups appearing in only a single parameter set. Replacement with 10% or 25% noise resulted in intermediate values (Table 2). The Ziziphoids clade showed the same tendencies, although less pronounced (Table 2 and Fig. 2).

Replacement with random noise quickly destabilized the clades to variation in parameter choice (Table 3). Only one of the original five groups stable under all 20 parameter sets (the *Adolphia-Discaria* p.p. clade) was nearly consistently recovered when adding 10% noise (in 23 of the 25 replicates-Tables 3 and 4). When adding 25% noise the same group appeared most frequently (as a mean the clade was stable in about 77% of the parameter space), but only in four of the 25 replicates was it recovered under all 20 cost sets (Tables 3 and 5). The genus *Colletia*, which is supported in all 20 parameter sets, both by the AT-RR and by the F-R, was surprisingly sensitive to the addition of random noise (Tables 3–5). The clade was recovered with lower frequency than some of the clades with lower support in terms of robustness to variation in parameter cost. For example, the (*D. americana*, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*) clade appears stable in 3 or 2 replicates, respectively, when 10% or 25% noise is added (Tables 4 and 5), whereas the *Colletia* clade is only stable in one of the 25 replicates when 10% noise is added (Table 4). Similarly, the *Colletia* clade has a Bremer support of 3 when gaps are weighted 2 and transversion and transition are each weighted 1, whereas the (*Adolphia*, *D. americana*, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*) clade-only supported under this cost set-has a Bremer support of 1. However, the latter clade appears as stable in one of the 25 replicates when adding 25% noise, whereas the *Colletia* clade never appears as stable when adding 25% noise (Tables 4 and 5).

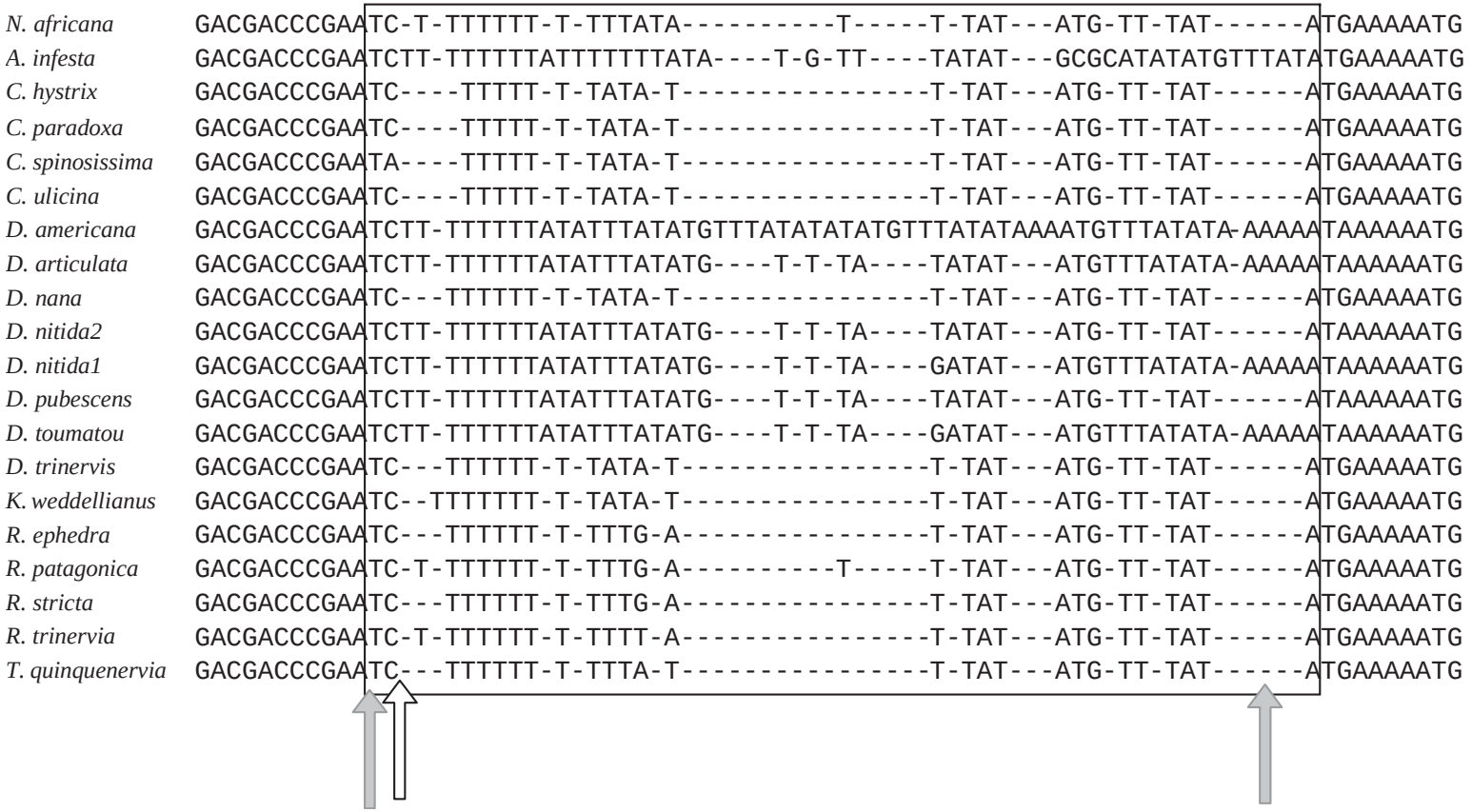

Fig. 1. Alignment of the AT-RR of the *trnL* intron in the Colletieae, using gap costs twice the cost of transversions, and transversions four times as costly as transitions. A gray arrow marks the 3′ end of the AT-RR as defined according to the method of Gatesy et al. (1993). A white arrow marks the 5′ end of the AT-RR as defined by using gap costs equal to transversions, twice, and four times as costly, and transversion–transition cost ratios 1:1, 2:1, and 4:1. A gray arrow marks the 5′ end of the AT-RR as defined by using gap equal to transversions, twice, and four times as costly, but a transversion–transition cost ratio of 1:0. The black square marks the AT-RR as delimited in this study.

Table 2. Numbers of groups stable in 20, 19, 18 or a single of the 20 parameter sets (number of parameter sets expressed in percent of total parameter sets) when the sequences are analyzed with the original AT-RR or with AT-RR, having 10% (mean of 25 replicates), 25% (mean of 25 replicates) or 100% (mean of 20 replicates) noise replacement

Matrix	No. of taxa	Percent of parameter sets				Total number of groups
		100	95	90	5	
Colletieae, sequences analyzed as a unit	21	3/0.14	1/0.05	3/0.14	4/0.19	28/1.33
Colletieae, sequences cut into three fragments	21	4/0.19	1/0.05	2/0.01	1/0.05	23/1.10
Colletieae, 10% noise	21	1.4/0.07	0.8/0.04	1.04/0.05	12.5/0.57	49.3/2.35
Colletieae, 25% noise	21	0.44/0.02	0.72/0.03	1.12/0.05	23.52/1.12	61.04/2.91
Colletieae, random AT-RR	21	0/0	0/0	0.05/0.002	99.3/4.73	145.25/6.917
Ziziphoids, sequences cut into three fragments	25	1/0.04	0/0	1/0.04	27/1.08	90/3.6
Ziziphoids, random AT-RR	25	0.15/0.006	0.15/0.006	0.35/0.014	84.35/3.374	153.2/6.128
Rhamnaceae, sequences cut into three fragments	24	0/0	1/0.04	0/0	44/1.83	123/5.13

Values at left of slashes: number of groups; values at right of slashes: number of groups divided by number of taxa.

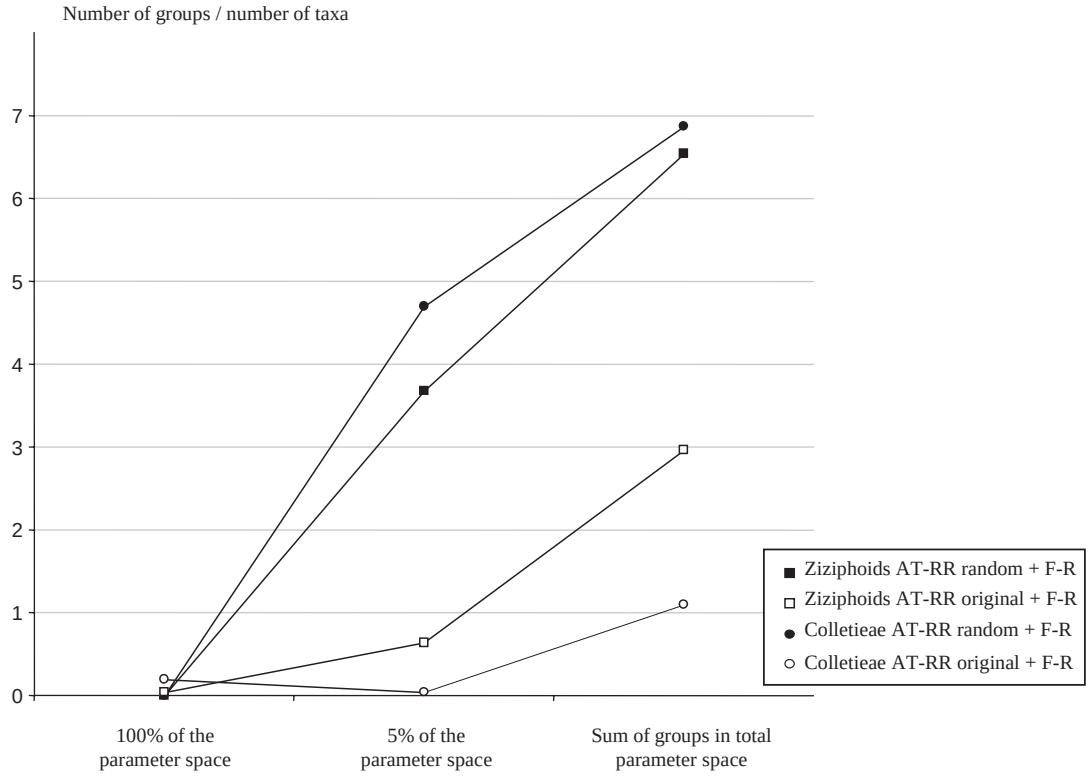

Fig. 2. The effect of replacing the AT-RR with a randomly generated one analyzed in combination with the F-R. The number of groups found in 20 (100%) or in a single (5%) of the 20 parameter sets is shown, as well as the total number of groups found in the entire parameter space when the original *trnL* sequence is analyzed cut into three fragments (white circles: Colletieae; white squares: Ziziphoids), and when the AT-RR is replaced with a randomly generated one analyzed in combination with the F-R (black circles: Colletieae; black squares: Ziziphoids).

Flanking regions

To compare the information content in the F-R and in the AT-RR as well as their possible interaction, the areas were analyzed separately, i.e., the F-R in combination versus the AT-RR on its own (Tables 6 and 7; Fig. 3). Within the Colletieae the F-R contained approximately the same amount of information as the AT-RR in terms of stable and unstable groups (Table 6). However, the F-R of the Colletieae do not have any influence on the outcome of the analysis when analyzed together with the AT-RR, the same results are obtained when the AT-RRs are analyzed alone (Fig. 3A and Table 7-*p*-values correspond to a randomized, paired *t*-test (Manly 1997) executed to evaluate the differences in number of groups using 5000 replications per analysis).

Table 3. Presence of clades stable in 20 (100%), 19 (95%), or 18 (90%) of the 20 parameter sets, in the original analysis or when 10% or 25% noise is added to the AT-RR

Clade	Original analysis (%)	25% noise mean	10% noise mean
Adolphia-Discaria p.p.	100	100	77.2
*D. nitida*1, *D. toumatou*	100	20.5	5
Colletia	100	38.75	16.8
C. hystrix, *C. paradoxa*, *C. spinosissima*	100	30.25	13.8
C. paradoxa, *C. spinosissima*	100	17	13.8
*D. nitida*2, *D. pubescens*	95	24.5	19.2
Adolphia infesta, *D. americana*, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*	90	85.5	60.4
Colletia, *D. nana*, *D. trinervis*, *Kentrothamnus weddellianus*, *Retanilla*, *Trevoa quinquenervia*	90	20.25	7.6

The mean values (mean of 25 replicates) are the presence of a given clade in number of parameter sets (expressed as percentage of total parameter space).

Table 4. Clades stable in one or more of the 25 replicates where 10% of the AT-RR is replaced with random noise, and presence of the same clade in the original analysis when analyzing the entire *trnL* intron or the F-R or AT-RR separately

Clade	Bremer support		Stable in number of replicates adding 10% random noise	Presence in parameter space of the original analysis		
	Cost set 111	Cost set 211		Complete *trnL* intron (%)	F-R (%)	AT-RR (%)
Adolphia-Discaria p.p.	7	17	23	100	—	100
D. americana, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*	—	6	3	20	—	25
*D. nitida*2, *D. pubescens*	1	3	1	95	70	—
Colletia	4	3	1	100	100	100
Colletia, *D. nana*, *D. trinervis*, *K.. weddellianus*, *Retanilla*, *T. quinquenervia*	2	1	1	90	—	95
Adolphia-Discaria p.p., *R. patagonica*	—	—	1	—	—	—
Adolphia-Discaria p.p., *R. trinervia*	—	—	1	—	—	—
D. americana, *D. articulata*	—	—	2	—	—	—
D. americana, *D. chacaye*	—	—	1	—	—	—
R. stricta, *R. trinervia*	—	—	1	—	—	—

Bremer support given for clades found under cost set 111 and cost set 211.

In the Ziziphoids and at the family level (Rhamnaceae) the AT-RRs were more stable to variation in parameter choice than were the F-Rs (Table 6). However, in the Ziziphoids adding the F-R to the AT-RR does change the outcome of the analysis at least in number of groups found in only one of the 20 cost sets (Fig. 3B and Table 7).

Discussion

Random data

One very conspicuous effect of replacing the original AT-RR with a randomly generated one was a pronounced sensitivity to variation in alignment parameters. In the Colletieae, when analyzing the original sequences, out of a total of 23 groups a single group (about 4%) was found in only one cost set (Table 2). When analyzing a random AT-RR in combination with the F-R about 68% of the total groups were found in only a single parameter set (99 groups out of 145 total groups). Groups stable in the entire parameter space were not found when analyzing a random AT-RR in combination with the F-R. Clearly, the type of signal found in the original AT-RR differs from the one found in a random AT-RR, at least under the random model used in this study with random changes, insertions and deletions of mainly As and Ts. The changes during the course of phylogeny within this particular AT-RR do not seem to conform to random hits, insertion, and deletion events of As and Ts, at least not at this specific taxonomic level.

Table 5. Clades stable in one or more of the 25 replicates where 25% of the AT-RR is replaced with random noise, and presence of the same clade in the original analysis when analyzing the entire *trnL* intron or the F-R or AT-RR separately

Clade	Bremer support		Presence in number of replicates adding 25% random noise	Presence in parameter space of the original analysis		
	Cost set 111	Cost set 211		Complete *trnL* intron (%)	F-R	AT-RR (%)
Adolphia-Discaria p.p.	7	17	4	100	—	100
D. americana, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*	6	1	2	20	—	25
Adolphia, *D. americana*, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*	—	1	1	90	—	90
C. hystrix, *C. paradoxa*, *C. ulicina*	—	—	1	—	—	—
C. hystrix, *C. paradoxa*	—	—	1	—	—	—
C. hystrix, *C. spinosissima*	—	—	1	—	—	—
D. americana, *D. chacaye*, *D. nitida*1, *D. toumatou*	—	—	1	—	—	—

Bremer support given for clades present in cost set 111 and cost set 211.

Table 6. Number of groups stable in all or a single of the 20 parameter sets (number of parameter sets expressed in percent of total parameter sets), and total number of groups found in the entire parameter space when the F-R and the AT-RR are analyzed separately

Matrix	No. of taxa	Percent of parameter sets		Total number of groups
		100%	5%	
Colletieae AT-RR	21	2/0.01	1/0.05	15/0.71
Colletieae F-R	21	3/0.14	3/0.14	12/0.57
Ziziphoids AT-RR	25	4/0.16	0/0	15/0.6
Ziziphoids F-R	25	1/0.04	21/0.84	71/2.8
Rhamnaceae AT-RR	24	3/0.13	10/0.42	25/1.04
Rhamnaceae F-R	24	0/0	50/2.08	109/4.54

Values at left of slashes: number of groups; values at right of slashes: number of groups divided by number of taxa.

The pattern obtained from analyzing the random AT-RRs is consistent with what is already known about random data. Informal inspection of the results showed that the number of trees generated from the data sets including a random AT-RR generally were lower than the number of trees generated from the original sequences. That random data generate only a few well resolved trees (but a poorly supported topology) is well known in conventional phylogenetic analysis (Hillis and Huelsenbeck 1992). Within the context of a sensitivity analysis where the costs of the alignment parameters are varied, the random AT-RRs generated in this study produced highly resolved consensus trees, but different topologies appeared when the parameter costs were changed. This is what leads to the high number of groups appearing in only a single parameter which is associated with a high number of total groups found in the entire parameter space, in addition to the lack of groups being stable in the entire parameter space. Intolerance to variation in parameter choice, with few or no groups surviving changes in the parameter cost set and many groups appearing in only a single parameter set, will hereafter be refered to as the 'noise effect'. This noise effect is obviously strictly related to this specific random model.

Flanking regions

The effect of replacing the original AT-RR with a random AT-RR was more pronounced within the Colletieae than within the Ziziphoids (Fig. 2). While the F-R contained only very little information for phylogenetic inference within the Colletieae (Table 6), the F-R within the more inclusive Ziziphoids clade may contain more hierarchical information. This phylogenetic information may exert some restrictive influence on the outcome when analyzing a random AT-RR in combination with the F-R.

Table 7. Number of groups stable in all or a single of the 20 parameter sets (number of parameter sets expressed in percent of total parameter sets), and total number of groups found in the entire parameter space when 10%, 25% or 100% noise is added to the AT-RR

Matrix	Number of replicates	100% of parameter sets		5% of parameter sets		Total number of groups	
		Mean	*p*-value	Mean	*p*-value	Mean	*p*-value
Colletieae							
AT-RR 10% noise + F-R	15	1.53		12.52		50.07	
			0.2244		0.1870		0.6428
AT-RR 10% noise	15	1.33		14.27		47.33	
AT-RR 25% noise + F-R	15	0.53		22.73		59.6	
			0.2224		0.0784		0.2114
AT-RR 25% noise	15	0.87		26.33		62	
AT-RR random + F-R	20	0	—	99.3		145.25	
					0.0625		0.2174
AT-RR random	20	0		107.2		149.4	
Ziziphoids							
AT-RR random + F-R	20	0.15		84.35		153.2	
			0.4160		0.0116		0.3698
AT-RR random	20	0.15		102.55		150.5	

The AT-RR is analyzed on its own or in combination with F-R. For noise treatments the number of replicates and mean value are given. The *p*-values correspond to a randomized, paired *t*-test comparing number of groups obtained when analyzing the AT-RR on its own or in combination with the F-R.

However, when comparing the two groups, the F-R within the Ziziphoids are more sensitive to parameter choice than are the F-R of the Colletieae (Table 6). At the family level the F-R are even more sensitive to variation in parameter choice, with no groups stable in more than 75% of the parameter space (data not shown). This corresponds to a 'noise effect', but it is not possible to judge whether this effect is produced by a pattern similar to random noise or by some other means. Inadequate sampling at the higher taxonomic levels in this study could be one of the factors causing the increased 'noise effect' in the F-R of the Ziziphoids and the Rhamnaceae. It is possible that more stable groups would appear if more thorough sampling were used within these groups.

When comparing the graphs in Fig. 3A it becomes evident that although the F-R of the Colletieae contain some amount of hierarchical information stable to variation in parameter choice the F-R have no influence on the outcome when analyzed in combination with a randomly generated AT-RR. No significant difference was found between number of groups stable in the entire parameter space, number of groups found in a single parameter set, or total number of groups when analyzing the randomly generated AT-RRs on their own or in combination with the F-R (Table 7). The signal found in the F-R is apparently too weak to have any influence on the outcome when analyzed in combination with a random AT-RR. This is not entirely the case when the F-R of the Ziziphoids are analyzed in combination with a randomly generated AT-RR (Fig. 3B). In this case numbers of groups found in a single parameter set were significantly different when analyzing the randomly generated AT-RRs on their own or in combination with the F-R (Table 7). Apparently, the signal found in the F-R of the Ziziphoids is sufficiently strong to exert some influence on the outcome when analyzed in combination with a random AT-RR, although the F-R of the Ziziphoids contain less information stable to variation in parameter choice than the F-R of the Colletieae.

This leads to the notion that robustness to variation in parameter choice on its own is not a satisfactory measure of phylogenetic signal quality, at least not when robustness is expressed as in this study. Some other measure has to be incorporated into the analysis to explain why and how the signal found in the F-R region of the Ziziphoids is stronger than the one found in the Colletieae clade.

AT-rich regions

The AT-RR and the F-R in the Colletieae contained approximately the same number of groups stable in the entire parameter space or present in only a single parameter set (Table 6). Number of total groups was higher in the AT-RR. Increased sensibility to variation in parameter choice within the AT-RR is in accordance with the presumed occurrence of a higher noise level within this region. However, when comparing this to the results obtained from analyzing a higher taxonomic level

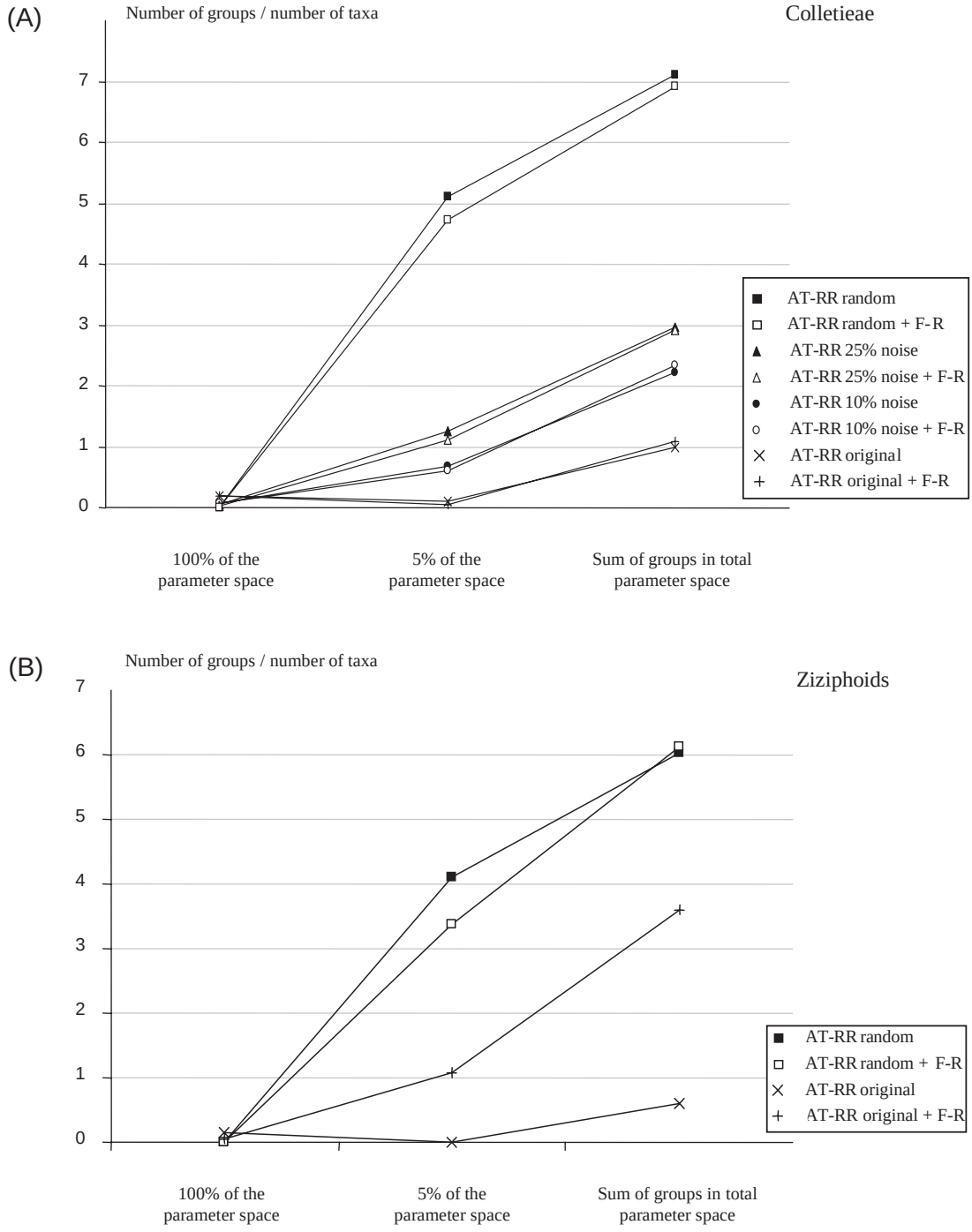

Fig. 3. Numbers of groups found in 20 (100%) or in a single (5%) of the 20 parameter sets, and total numbers of groups found in the entire parameter space when the AT-RR is analyzed on its own or in combination with the F-R. (A) Colletieae: The original AT-RR analyzed in combination with the F-R (vertical crosses), and on its own (slanted crosses). The AT-RR with 10% noise added analyzed on its own (black circles), and in combination with the F-R (white circles). The AT-RR with 25% noise added analyzed on its own (black triangles), and in combination with the F-R (white triangles). The AT-RR with 100% noise added analyzed on its own (black squares), and in combination with the F-R (white squares). (B) Ziziphoids: The original AT-RR analyzed in combination with the F-R (vertical crosses), and on its own (slanted crosses). The AT-RR with 100% noise added analyzed on its own (black squares), and in combination with the F-R (white squares).

the prediction does not hold. Within the Ziziphoids the F-R were considerably more sensitive to variation in parameter choice than the AT-RR was (Table 6). Moreover, the AT-RR at this taxonomic level was more stable to variation in parameter choice than the AT-RR within the Colletieae. This finding is enigmatic and

reinforced by the fact that also at family level the AT-RR was far more stable to variation in parameter choice than the F-R (Table 6). Clearly, random hits, insertions, or deletions of As and Ts do not seem to be prevailing within this AT-RR.

The reason for this marked robustness to variation in parameter choice seems to be the difference in sequence length within the AT-RR. Groups that are stable in the entire parameter space all correspond to a specific number (or range) of base pairs. Grouping on the basis of sequence length is reasonable as indels are products of evolution. However, concern may be raised whether the length information has received the right treatment in the present study. The approach used by POY ver. 2.7 is to count each position separately, for example five adjacent gaps will be treated as five individual gaps. This procedure reinforces the robustness of the groups found in the present analysis. However, exploring different gap treatments is not an issue of this paper. It should be mentioned that the use of a lower cost for the extension of a gap may be worth exploring when analyzing areas with much length variation. The use of a separate extension gap cost has recently been incorporated in newer versions of the program POY. When lower costs for extension gaps are applied to the AT-RR of the Colletieae, the *Colletia* clade and the *Adolphia-Discaria* p.p. clade remain stable, corroborating the robustness of these two clades. The use of extension gaps has not been explored any further in this study which was based on an older version of POY and carried out within a parameter cost space prevailing in published sensitivity analysis. In the present case where extension gaps have not been applied, it is sufficient to mention that if sequence length is strictly correlated to the phylogenetic history, and if the grouping occurs exclusively according to sequence length, erroneously placed taxa should be most frequent in higher level analyses with sparse taxon sampling and less frequent in species-level analyses with dense taxon sampling. Note that this should also hold true in conventional analyses where gaps are coded as a fifth character. Fortunately, it is in species-level analyses that the information contained in the ambiguously aligned regions is most acutely needed. In species-level analyses the phylogenetic information contained in the F-R, or other unambiguously aligned sequences, is sparse or absent. At higher taxonomic levels additional information is often present. This information may constrain any possible erroneous pattern, including those that may come from grouping on the basis of sequence length.

Noise

The most conspicuous effect of adding random noise was a gradual increase in sensitivity to the variation of parameter cost (Table 2).

When 10% of random noise is added in the Colletieae data set, a mean of 1.4 groups are stable in all 20 parameter sets, as opposed to 3 or 4 groups found in the original data set (Table 2). Only 2 clades are stable in a mean of 80% or more of the parameter space when 10% noise is added (Table 3). This loss of robustness to the addition of noise is markedly more pronounced than what was found by Wenzel and Sidall (1999). The differences are probably related to the alignment approach. Wenzel and Sidall (1999) tested robustness to noise within matrices using a static homology scheme, whereas the direct optimization used here tests robustness to noise within a dynamic homology scheme (Wheeler 2001) where the primary homologies may vary with parameter choice. The multiple hits, insertions, and deletions of mainly As and Ts may destabilize the alignment faster within the AT-RR where multiple gaps are inserted than would a more equiprobable model of replacement within the F-R of the *trnL* intron. This possibility, however, was not explored.

When adding 10% of random noise only the *Adolphia-Discaria* p.p. clade group appeared under all 20 parameter sets in almost all replicates. This clade includes all species with long sequences within the AT-RR (41–61 bp versus 28–31 bp in all other species). Nine additional groups were found constant in one or more of the replicates (Table 4). Of these nine groups only the *Colletia* clade was constant under all parameter sets when analyzing the original AT-RR or the F-R on their own or in combination. Nevertheless, this group only appeared stable under all parameter sets in a single of the 25 replicates. In contrast the (*D. americana*, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*) clade appeared stable in the entire parameter space in three of the 25 replicates, although this clade only appeared in about 20% of the parameter sets when analyzing the original matrix, and was not present when analyzing the F-R on their own (Table 4). When adding 25% noise the same tendencies are more pronounced (Table 5).

The *Adolphia-Discaria* p.p. group that appears to be most robust to noise addition is also the group with highest Bremer support. Bremer supports have been found to co-vary with robustness to variation in parameter choice (Wheeler and Hayashi 1998), but this need not always be the case (Giribet 2001, 2003). The *Colletia* clade has higher Bremer support than other clades less sensible to the addition of noise, e.g. the (*Adolphia*, *D. americana*, *D. articulata*, *D. chacaye*, *D. nitida*1, *D. toumatou*) clade found under cost set 211 (Table 5). Again, sequence length within the AT-RR may cause the different behavior of the clades, with the *Colletia* clade having the shortest sequences (28 bp).

Giribet (2003) found that results from resampling techniques such as jackknifing in many cases were more representative of node stability than Bremer support values, but not even the jackknife frequencies were well

correlated with node stability in all cases. Character conflict, e.g. as measured by relative supports (Goloboff and Farris 2001), as well as number and kind of character changes supporting a particular group, are all properties that may affect the stability of a clade.

Final remarks

Regions of ambiguous alignment are frequently excluded from phylogenetic analyses due to the questionable primary homologies established during the alignment of such regions. At all three taxonomic levels the area of ambiguous alignment studied here supported several groups stable in the entire parameter space, hence they yield the same final conclusion within a broad spectrum of possible analytic parameters. Furthermore, the AT-RR proved to be no more equivocal in supporting groups than the supposedly non-ambiguous F-R when analyzed within a sensitivity analysis approach (Wheeler 1995). Neither ambiguous alignment nor blurred phylogenetic signals seem to be a problem in this AT-RR at any of the three taxonomic levels, and thus the region should not be excluded from the analyses (note that limited sampling at higher taxonomic levels may give erroneous results in terms of grouping, but the stability of the groups found at the higher-level analyses still suggests non-random structure of the data).

This study has found no reason for excluding the AT-RR provided that an adequate method is used to find all optimal alignments/trees under a specific parameter cost set. If on the other hand the AT-RR had been found to conform to a signal also found by random data—e.g., lacking robustness to variation in parameter costs—the question would remain whether or not to exclude the area. Random data are known to provide well resolved phylogenetic trees (Hillis and Huelsenbeck 1992), and this also seems to be the case in sensitivity analyses, if only a single parameter cost set is taken into account. Therefore, including very noisy to random data could give a well resolved but spurious tree. However, if data have to be removed from the analysis some objective criterion is required to define exactly what to retain and what to exclude (Gatesy et al. 1993; Giribet and Wheeler 1999). In real cases ambiguous regions are probably neither noise-free nor completely random but rather somewhere in between, depending on the data set. It seems questionable whether we will be able to decide what kind of information a highly variable site may contain just by the mere inspection of the sequences. A more sound approach following the logic of total evidence (Kluge and Wolf 1993; Nixon and Carpenter 1996) would be to include all data. If some regions happen to be of 'random/noisy' structure, their effect on the analysis may be constrained by the information in the remainder of the data, depending on the amount and quality of the latter.

Few methods are suitable for including ambiguously alignable regions. A hand alignment will present one possible alignment of the area though not necessarily an optimal one. Lutzoni et al. (2000) proposed a method that relies on aligning the sequences prior to analysis, defining and re-coding the ambiguous areas separately in step matrices using a single cost set. Direct optimization (Wheeler 1996) will find all optimal alignments under a specific cost set, and does not require any manipulation of the sequences prior to analysis. However, the drawback of using direct optimization is the extensive computing time that presently makes sensitivity analyses of more than about 50 taxa difficult if only a single processor is used. POY can, however, run in parallel (the program and documentation can be found at ftp://ftp.amnh.org/pub/molecular/poy/), and the use of computer clusters is becoming increasingly common in large-scale phylogenetic analysis based on direct optimization.

Whichever method is used, a better resolution of the consensus tree is not a satisfactory result on its own. Noisy/random data can also result in well-resolved trees. In the present context groups defined by randomly generated data disappeared when parameter costs were varied. What exactly makes a group resistant to changes in parameter costs is unclear. Bremer support may co-vary with robustness to variation in parameter choice, but this need not be the case. Some measure of robustness to variation in parameter costs, as sometimes seen in sensitivity analysis, is useful. Obviously, even if a group is found to be robust to changes in the parameter costs, we will never know whether this pattern arises from phylogeny or something else. But this will always be the case even in conventional analysis. On the other hand, if a sensitivity analysis is made and optimal costs are defined, one would surely like to know whether a group resists a slight change of the alignment, no matter how well supported that node may be under the optimal costs. This is because parameter costs will always be approximate. Costs are held constant over the entire data set (Hickson et al. 2000), our means for picking the optimal costs are approximate (see, e.g., Dowton and Austin 2002), and although some parameters are varied (like gap costs and transversion/transistion costs), several other factors not accounted for are also expected to have an influence on the development of the sequences (Wheeler 1995).

Finally, what exactly makes a group resistant to changes in parameter costs is still unclear. Traditional support measures such as Bremer support and jackknife frequencies may or may not concur with node stability (Giribet 2003). Additional information, such as numbers of characters in conflict with a given clade, may increase

the predictive power of the final cladograms (Goloboff and Farris 2001). However, it seems clear that nodal support measures and node stability show different aspects of the data set, and that both are needed when evaluating the firmness of the final conclusions.

Acknowledgements

I wish to thank Martín Ramírez for patient help during the construction of the macro-files. Norberto Giannini, Pablo Goloboff, Gonzalo Giribet, Martín Ramírez, Ole Seberg, and one anonymous reviewer provided useful comments on the manuscript. Lauren Aaronsen improved the spelling and grammar. This work was supported by a University of Copenhagen Ph.D. grant.

References

Bremer, K., 1994. Branch support and tree stability. Cladistics 10, 295–304.

Clegg, M.T., Zurawski, G., 1992. Chloroplast DNA and the study of plant phylogeny: present status and future prospects. In: Soltis, P.S., Soltis, D.E., Doyle, J.J. (Eds.), Molecular Systematics of Plants. Chapman and Hall, New York, pp. 1–13.

De Pinna, M.C.C., 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics 7, 367–394.

Dowton, M., Austin, A.D., 2002. Increased congruence does not necessarily indicate increased phylogenetic accuracy — the behavior of the incongruence length difference test in mixed-model analyses. Syst. Biol. 51, 19–31.

Doyle, J.J., Davis, J.I., 1998. Homology in molecular phylogenetics: a parsimony perspective. In: Solis, D.E., Soltis, P.S., Doyle, J.J. (Eds.), Molecular Systematics of Plants. II. DNA Sequencing. Kluwer Academic Publishers, Boston, pp. 101–131.

Gatesy, J., DeSalle, R., Wheeler, W.C., 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol. Phylogenet. Evol. 2, 152–157.

Gielly, L., Taberlet, P., 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus *rbc*L sequences. Mol. Biol. Evol. 11, 769–777.

Giribet, G., 2001. Exploring the behavior of POY, a program for direct optimization of molecular data. Cladistics 17, S60–S70.

Giribet, G., 2003. Stability in phylogenetic formulations and its relationship to nodal support. Syst. Biol. 52, 554–564.

Giribet, G., Wheeler, W.C., 1999. On gaps. Mol. Phylogenet. Evol. 13, 132–143.

Gladstein, D., Wheeler, W.C., 2000. POY: the optimization of alignment characters. Program and documentation. American Museum of Natural History, New York. Available at ftp://ftp.amnh.org/pub/molecular/poy/.

Goloboff, P.A., 1998. NONA ver. 1.9, program and documentation, Willi Hennig Society. Web site http://www.cladistics.org/.

Goloboff, P.A., Farris, J.S., 2001. Methods for quick consensus estimation. Cladistics 17, S26–S34.

Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J., Szumik, C.A., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Hickson, R.E., Simon, C., Perry, S.W., 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. Mol. Biol. Evol. 17, 530–539.

Hillis, D.M., Huelsenbeck, J.P., 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. 83, 189–195.

Kluge, A.G., Wolf, A.J., 1993. Cladistics: what's in a word? Cladistics 9, 183–199.

Lutzoni, F., Wagner, P., Reeb, V., Zoller, S., 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violation positional homology. Syst. Biol. 49, 628–651.

Manly, B.F.J., 1997. Randomization, Bootstrap and Monte-Carlo Methods in Biology 2nd Edition. Chapman & Hall, London.

Nixon, C.K., Carpenter, J.M., 1996. On simultaneous analysis. Cladistics 12, 221–241.

Richardson, J.E., Fay, M.F., Cronk, Q.C.B., Bowman, D., Chase, M.W., 2000. A phylogenetic analysis of Rhamnaceae using *rbcL* and *trnL-F* plastid DNA sequences. Am. J. Bot. 87, 1309–1324.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics. Sinauer, Sunderland, MA, pp. 407–514.

Thulin, M., Bremer, B., Richardson, J.E., Niklasson, J., Fay, M.F., Chase, M.W., 1998. Family relationships of the enigmatic rosid genera *Barbeya* and *Dirachma* from the Horn of Africa region. Plant Syst. Evol. 213, 103–119.

Wenzel, J.W., Sidall, M.E., 1999. Noise. Cladistics. 15, 51–64.

Wheeler, W.C., 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331.

Wheeler, W.C., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.

Wheeler, W.C., 2001. Homology and DNA sequence data. In: Wagner, G.P. (Ed.), The Character Concept in Evolutionary Biology. Academic Press, New York, pp. 303–318.

Wheeler, W.C., 2003. Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. Cladistics 19, 261–268.

Wheeler, W.C., Gatesy, J., DeSalle, R., 1995. Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. Mol. Phylogenet. Evol. 4, 1–9.

Wheeler, W.C., Hayashi, C.Y., 1998. The phylogeny of the extant chelicerate orders. Cladistics 14, 173–192.