The Royal Society

Rate of gene sequence evolution and species diversification in flowering plants: a re-evaluation

Vincent Savolainen[1] **and Jérôme Goudet**[2]

[1]*Institut de Botanique Systématique et de Géobotanique, University of Lausanne, CH-1015 Lausanne, and Conservatoire et Jardin botaniques, CH-1292 Geneva, Switzerland* (vincent.savolainen@ibsg.unil.ch)
[2]*Institut de Zoologie et d'Ecologie Animale, University of Lausanne, CH-1015 Lausanne, Switzerland* (jerome.goudet@izea.unil.ch)

Barraclough and co-workers (in a paper published in 1996) observed that there was a significant positive correlation between the rate of evolution of the *rbc*L chloroplast gene within families of flowering plants and the number of species in those families. We tested three additional data sets of our own (based on both plastid and nuclear genes) and used methods designed specifically for the comparison of sister families (based on random speciation and extinction). We show that, over all sister groups, the correlation between the rate of gene evolution and an increased diversity is not always present. Despite tending towards a positive association, the observation of individual probabilities presents a U-shaped distribution of association (i.e. it can be either significantly positive or negative). We discuss the influence of both phylogenetic sampling and applied taxonomies on the results.

Keywords: biodiversity; gene sequence evolution; *rbc*L; 18S nuclear gene; angiosperms

1. INTRODUCTION

Why are some groups of organisms much more species rich than others? This vivid question remains a puzzle to evolutionary biologists. Many possible causes have been tested, from intrinsic key innovations (e.g. vivipary, see Slowinski & Guyer 1993) to extrinsic events such as environmental shifts (e.g. climate change, see Sanderson & Donoghue 1996).

Going deeper down to the molecular level, several authors proposed that speciation may be closely linked with the rate of genetic change (e.g. Mayr 1954; Harrison 1991; Bousquet *et al.* 1992; Coyne 1992). However, as stated by Barraclough *et al.* (1996), there has been no comparative evidence for this claim. With the spread of automatic sequencing facilities, multiple large-scale molecular phylogenies have been produced, which should allow for the first direct evaluations of this hypothesis.

Based on the *rbc*L broad phylogenetic analysis of Chase *et al.* (1993), Barraclough *et al.* (1996) published results showing a positive correlation between the rate of gene sequence evolution (reflected by branch lengths in the cladogram) and the number of species within families of flowering plants. These results were at first surprising; indeed, a simple check of terminal branches in the Chase *et al.* (1993) tree showed that many of the longest branches connect families with few species (at the opposite, according to Barraclough *et al.* (1996), we would expect that long branches connect highly diversified families). Moreover, when working with phylogenetic trees it is very common to get 'unusual long branches' for which lengths do not seem to be at all correlated with species richness. To take just one example, a recent Celastrales survey (V. Savolainen and M. W. Chase, unpublished data) showed that Stackhousiaceae are connected to their sister Celastraceae *p.p.* by a very long branch, despite the fact that they contain only 25 species, versus up to 800 for the sister taxa.

After cross-checking the results in Barraclough *et al.* (1996) with T. Barraclough, we arrived at the revised results presented in table 1*b*. (Following strictly the method described by Baraclough *et al.* (1996), the values in table 1*b* are correct, whereas those originally published by Barraclough *et al.* (the values in table 1*a*) are wrong due to mistakes in the calculation.)

Because these results are equivocal, we present here an expanded survey testing the correlation between the rate of sequence evolution and of species diversification in flowering plants. We increased the data set of Barraclough *et al.* (1996) by adding all family pairs that could be identified from the Chase *et al.* (1993) tree (i.e. 23 additional family pairs). We also used two other independent datasets: the *rbc*L phylogeny of monocotyledons presented in Chase *et al.* (1995) and the angiosperm phylogeny published by Soltis *et al.* (1997), which is based on the 18S nuclear gene instead of the *rbc*L chloroplast gene. We used two tests specifically designed for comparisons of family pairs (i.e. a test devised by Slowinski & Guyer (1993) and a modified version of this by Goudet (1998)), instead of only using the Wilcoxon sign test (Wilcoxon 1945).

Thus, using multiple data sets and new tests, we re-evaluate whether a higher rate of gene evolution could have effectively caused an increased diversification in plants.

Received 5 November 1997 *Accepted* 19 December 1997

Table 1. *Correlations between species diversity and branch length in the* rbc*L phylogeny of Chase* et al. *(1993)*

((*a*) Original table from Barraclough *et al*. (1996) and (*b*) its corrected version (this paper). Probability values calculated using the Wilcoxon sign test.)

(*a*) Original values from Barraclough *et al*. (1996)

subset of nodes that are isolated by greater than the following number number of substitutions	number of family pairs	median value of diversity contrasts	*p* (Wilcoxon)
>0	39	0.41	0.048
>5	33	0.64	0.017
>10	19	0.72	0.038
>15	8	0.82	0.098

(*b*) Corrected values (after cross-checking the results in Barraclough *et al*. (1996) with T. Barraclough, we arrived at the values presented below)

subset of nodes that are isolated by greater than the following number of substitutions	number of family pairs	median value of diversity contrasts	*p* (Wilcoxon)
>0	39	0.13	0.255
>5	33	0.24	0.153
>10	19	0.72	0.038
>15	8	0.57	0.141

2. METHODS

(a) *Data sets*

In addition to the 39 family pairs of Barraclough *et al*. (1996), the following data sets have been used: (i) the monocotyledon *rbc*L phylogeny of Chase *et al*. (1995), which differs from Chase *et al*. (1993) as many additional monocotyledon species have been included; (ii) the angiosperm 18S nuclear-based phylogeny of Soltis *et al*. (1997); and (iii) an increased data set based on Chase *et al*. (1993) comprising the family pairs used by Barraclough *et al*. (1996) plus 23 additional pairs. Indeed, Barraclough *et al*. (T. G. Barraclough, personal communication) deleted from their analysis all sister families from orders sensu Cronquist, which included families not sampled by Chase *et al*. (1993). However, we did not want to use any other classification scheme as a criterion to delete families (as they are all questionable; see Savolainen *et al*. (1997) for an example in one order) and we decided to strictly follow the nomenclature originally published in Chase *et al*. (1993). The raw data are available upon request from the authors.

(b) *Statistical tests*

Species numbers for each family are from Mabberley (1993) and from Watson & Dallwitz (1991; available on the World Wide Web at `http://www.keil.ukans.edu/delta`). To avoid possible errors in tree topologies, we followed Barraclough *et al*. (1996) and performed the analysis on successive subsets of nodes separated from adjacent ones by an increasing number of substitutions.

Rather than using only the Wilcoxon sign test as described in Barraclough *et al*. (1996; see table 1*a*), we used an improved statistical approach based on random speciation and extinction, as described in Slowinski & Guyer (1993). Under a null model of random speciation and extinction, all group sizes are equally probable. The probability of observing differences in the sizes of the groups between those possessing the longest branch and their sisters can be calculated. The results per family are then combined using the Fisher procedure (e.g. Manly 1986). Finally, we used a modified method of the Slowinski & Guyer test (Goudet 1998), which used a randomization procedure instead of the Fisher combination of probabilities that Goudet showed to give unduly large type I and type II errors (see also Nee *et al*. 1996).

(c) *Power analysis*

The power of the Slowinski & Guyer test and the Goudet test were estimated. Under the hypothesis that the family pairs of the Chase *et al*. (1995) and Soltis *et al*. (1997) data sets are a random sample of all family pairs, we can sample them with replacement and reapply both tests to the bootstrapped data sets. The power of these tests can be estimated by the number of times in which the results are significant at the 5% level. One main application of power analysis is the estimation of the sample size required to achieve significance $(1-b)\%$ of the time, for which b is the type II error. Therefore, we repeated the bootstrap tests described above for a number of family pairs varying between two and then five up to 100 by increment of five, to obtain the power of the two tests for these different sample sizes.

3. RESULTS

Table 1*a* gives the original table published by Barraclough *et al*. (1996), compared with the corrected values (this paper) (table 1*b*): there is now only one value in four which is significant at the 5% level, whereas Barraclough *et al*. (1996) found the contrary. Thus, based on these data only (i.e. the 39 family pairs identified by Barraclough *et*

Table 2. *Probability values for positive association between species diversity and branch length in phylogenies*

(Calculations of subsets of nodes took into account the averaged branch lengths for those connecting several representatives of the same family. Probability values calculated using the Wilcoxon sign test, Slowinsky & Guyer test (Slowinsky & Guyer 1993) and its modified version, the Goudet test (Goudet 1998).)

(*a*) *rbc*L angiosperm phylogeny of Chase *et al.* (1993), family pairs from Barraclough *et al.* (1996)

subset of nodes	number of family pairs	*p* (Wilcoxon)	*p* (S. & Guyer)	*p* (Goudet)
>0	39	0.1340	0.001	0.095
>5	34	0.1260	0.001	0.083
>10	20	0.0450	0.001	0.023
>15	9	0.0310	0.000	0.031

(*b*) *rbc*L angiosperm phylogeny of Chase *et al.* (1993), increased data set

subset of nodes	number of family pairs	*p* (Wilcoxon)	*p* (S. & Guyer)	*p* (Goudet)
>0	56	0.0440	0.000	0.018
>5	46	0.0210	0.000	0.008
>10	27	0.0170	0.000	0.003
>15	12	0.0270	0.000	0.013

(*c*) *rbc*L monocotyledon phlyogeny of Chase *et al.* (1995)

subset of nodes	number of family pairs	*p* (Wilcoxon)	*p* (S. & Guyer)	*p* (Goudet)
>0	27	0.5360	0.005	0.495
>5	22	0.4200	0.003	0.344
>10	13	0.2240	0.000	0.019
>15	8	0.2380	0.000	0.026

(*d*) 18S angiosperm phylogeny of Soltis *et al.* (1997)

subset of nodes	number of family pairs	*p* (Wilcoxon)	*p* (S. & Guyer)	*p* (Goudet)
>0	39	0.2400	0.006	0.312
>5	16	0.0160	0.004	0.004

al. (1996) from the cladogram of Chase *et al.* (1993)), the species diversity does not seem to be correlated with the rate of gene sequence evolution.

Using the test devised by Slowinski & Guyer (1993) and its modified version (Goudet 1998), table 2 presents the results of tests of positive association for the four data sets (see § 2). Whereas the Slowinski & Guyer test always rejects the null hypothesis of non-association, its modified version by Goudet rejects it nine times out of fourteen. We would therefore conclude that whereas the positive correlation is always present using the original Slowinski & Guyer test, it is only present for well-supported nodes in the phylogenies using the Goudet test (i.e. nodes isolated from adjacent ones by an increasing number of substitutions).

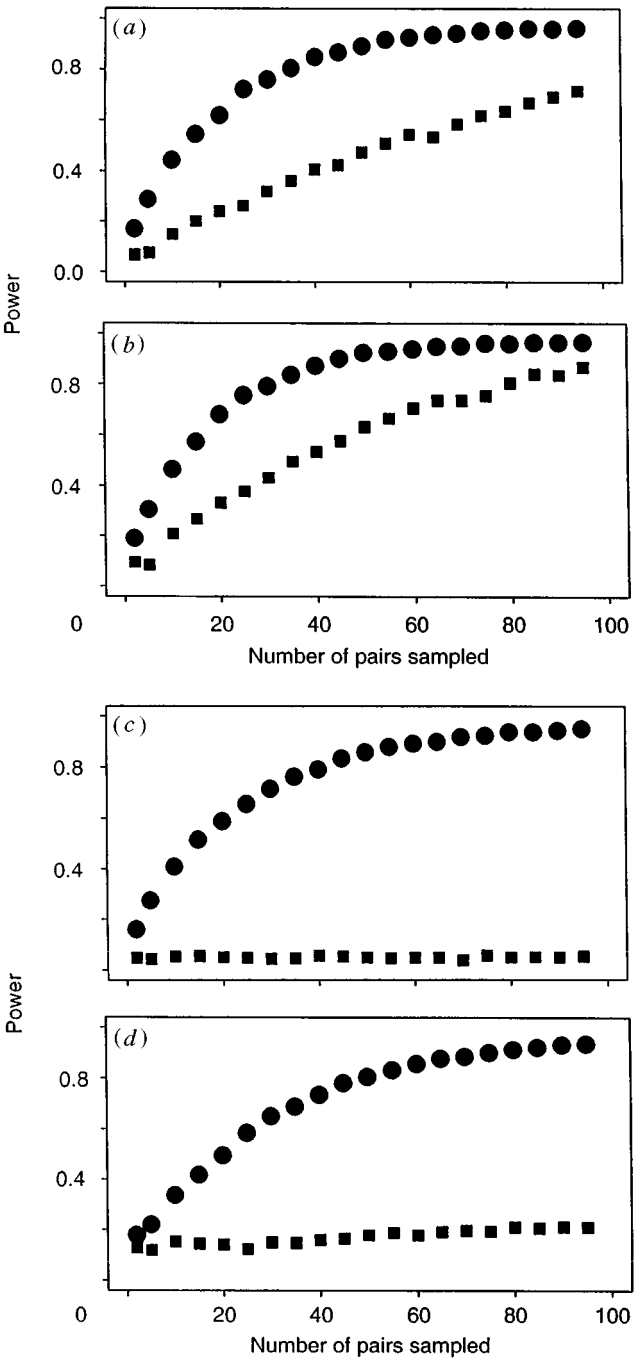

Figure 1. Power of the Slowinski & Guyer (circles) and Goudet (squares) tests (see § 1). (*a*) *rbc*L angiosperm phylogeny of Chase *et al.* (1993), family pairs from Barraclough *et al.* (1996). (*b*) *rbc*L angiosperm phylogeny of Chase *et al.* (1993), increased data set. (*c*) *rbc*L monocotyledon phylogeny of Chase *et al.* (1995). (*d*) 18S angiosperm phylogeny of Soltis *et al.* (1997). Both tests converge for large sample sizes using the *rbc*L data sets (*a*,*b*), whereas they tend towards opposite conclusions when using the monocotyledon and 18S data sets (*c*,*d*).

To discriminate between these hypotheses we carried out a power analysis (see § 2). Whereas for the *rbc*L angiosperm data the two tests converge for a large number of family pairs (figure 1*a*,*b*), these same two tests give opposite conclusions for both the monocotyledon and 18S angiosperm data sets (figure 1*c*,*d*). These paradoxical results are discussed in the following section, in the

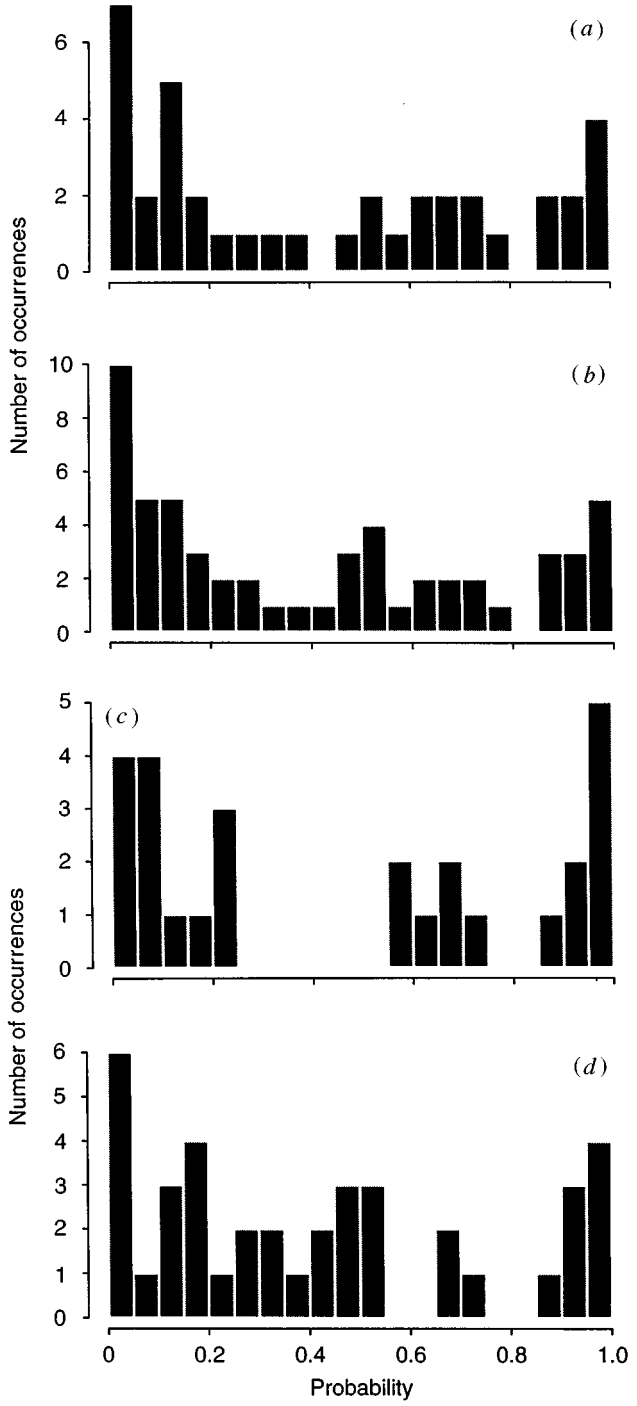

Figure 2. Distribution of sister groups p values from Slowinsky & Guyer null model. (*a*) *rbc*L angiosperm phylogeny of Chase *et al.* (1993), family pairs from Barraclough *et al.* (1996). (*b*) *rbc*L angiosperm phylogeny of Chase *et al.* (1993), increased data set. (*c*) *rbc*L monocotyledon phylogeny of Chase *et al.* (1995). (*d*) 18S angiosperm phylogeny of Soltis *et al.* (1997).

light of the sister groups' distribution of p values (figure 2).

4. DISCUSSION AND CONCLUSION

First, testing the association between the rate of sequence evolution and species diversification, and having given here the corrected values compared to those originally published by Barraclough *et al.* (1996), the overall resulting probabilities do not show a significant positive association (table 1). However, when using another test specifically designed to compare family pairs (namely the Slowinski & Guyer test, instead of the Wilcoxon sign test) and additional data sets (based on both plastid and nuclear genes), a positive association was found to be always significant (table 2).

Second, when a randomization procedure (instead of a Fisher procedure) is used to combine probabilities (the Slowinski & Guyer test modified by Goudet (1998); see also Nee *et al.* 1996), the association between branch length and the number of species tends towards a positive association using the angiosperm *rbc*L data sets, whereas it tends to the opposite conclusion using the monocotyledon and the 18S data sets (figure 1), when all nodes are considered.

Because the Fisher procedure has been shown to give unduly large type I errors when the distribution is U-shaped (Goudet 1998), we could expect that our data would follow this sort of distribution. Looking at the distribution of individual probabilities (figure 2), their distribution is indeed more U-shaped. Thus, despite the fact that the overall tendency is towards a positive association, close observation of individual probabilities shows that the association can go either way.

Why is this so? The first reason is that, in the monocotyledon and 18S data sets, there are more family pairs where one is very large and the other is very small, which leads statistically to marginal associations. A check of the smallest families (less than ten species) shows that they represent 20% in the monocotyledon and the 18S phylogenies, although this value decreases to 8% in the angiosperm *rbc*L phylogeny. However, these data sets were not published at the same time: the angiosperm *rbc*L phylogeny was published earlier (Chase *et al.* (1993) versus Chase *et al.* (1995) and Soltis *et al.* (1997)). Because members of small families are often rare and geographically restricted, it is difficult to collect them. Thus, at the time of the angiosperm paper of Chase *et al.* (1993), no specific sampling plan guided this study, and many representatives of these small families were not yet available. Later, Chase *et al.* (1995) and Soltis *et al.* (1997) acquired these samples and added them. For example, in Chase *et al.* (1993), Potamogetonaceae were sister to Alismataceae: both have approximately 100 species and the Slowinski & Guyer p value we calculated is 0.5. In 1995, Chase *et al.* added newly available families: Potamogetonaceae became sister to Zosteraceae (18 species) whereas Alismataceae had Limnocharitaceae as its sister (12 species). The p values we have calculated on these data show a marginal positive association in the former ($p=0.15$) and a negative one in the latter ($p=0.89$).

The second reason has to do with the taxonomies employed. When we reanalyse the data from the Chase *et al.* phylogeny, we stated that we followed strictly the nomenclature presented in Chase *et al.* (1993). This nomenclature is largely based on the one of Cronquist (1981), which inspired Mabberley (1993), and was in turn used by Barraclough *et al.* (1996). The Cronquist classification scheme is the widest used and taught so far, but it is also well known that Arthur Cronquist was 'reluctant to assign the rank of family to small satellite groups' (Cronquist 1981). As a result, Cronquist (1981) and

Mabberley (1993) described 383 angiosperm families. However, based on various studies (e.g. Dahlgren *et al.* 1995), the number of flowering plant families can be increased. Watson & Dallwitz (1991; updated 1997 version available on the World Wide Web, see § 1) used 567 families. Because we applied the Watson & Dallwitz nomenclature, small families popped up which, again, led to extreme values of association.

Finally, a third reason might be that parts of the phylogenies are not always well-supported, particularly for sister groups separated by short nodes. Long branches with few sampled taxa are subject to the phenomenon of 'long branch attraction', which would in turn shorten the branches and lead to spurious clusterings. Indeed, when removing these short nodes from all four data sets, the overall tendency is towards an increase in the significance of association (table 2).

To conclude, we think it would be premature to say that there is undoubtedly a cause-and-effect relationship between the rate of gene evolution and an increased diversity in plants. Despite having shown that the overall tendency is towards a positive association, many individual values go in the opposite way. When this association is negative, it depends on the lineages or subsets of nodes under consideration. These associations are better examined using unbiased tests (see Goudet 1998; Nee *et al.* 1996), and they are severely influenced by the taxonomies employed and the phylogenetic sampling. This enhances the need for (i) a new, fully integrated system of classification, and (ii) intensively sampled, multiple molecular phylogenies. Then evolutionary processes, as the fundamental relationship between micro- and macroevolution, will be more accurately studied among plants. Moreover, correlations are not explanations: is the rate of DNA sequence evolution a direct cause of the species diversity, and how closely linked are these factors? Gene evolution has been correlated many times with, among others, the well-known effect of generation time, the metabolic rate or with some mutagenic factor; how does species diversity fit with these traits? However complex is the biological network affecting speciation, we can only conclude here that there is not yet a general rule that would apply to all plant families.

We are very grateful to Timothy G. Barraclough, Mark W. Chase, Philippe Clerc, Laurent Keller, Jean-François Manen, Max Reuter and anonymous referees for useful comments on the manuscript. We also would like to thank Timothy Barraclough for sending us his raw data, Olivier Blanc who initiated this work with preliminary calculations, and Sibylle Becher and Magnus Borer for revising the English. J. G. was supported by the Swiss National Science Foundation (grant no. 31-43443.95).

REFERENCES

Barraclough, T. G., Harvey, P. H. & Nee, S. 1996 Rate of *rbc*L gene sequence evolution and species diversification in flowering plants (angiosperms). *Proc. R. Soc. Lond.* B **263**, 589–591.

Bousquet, J., Straus, S. H., Doerksen, A, H. & Price, R. A. 1992 Extensive variation in evolutionary rate of *rbc*L gene sequences among seed plants. *Proc. Natn. Acad. Sci. USA* **89**, 7844–7848.

Chase, M. W. (and 41 others) 1993 Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. *Ann. Missouri Bot. Gard.* **80**, 528–580.

Chase, M. W., Stewenson, D. W., Wilkin, P. & Rudall, P. J. 1995 Monocotyledon systematics: a combined analysis. In *Monocotyledons: systematics and evolution* (ed. P. J. Rudall, P. J. Cribb, D. F. Cutler & C. J. Humphries), pp. 685–730. Kew: Royal Botanic Gardens.

Coyne, J. A. 1992 Genetics and speciation. *Nature, Lond.* **355**, 511–515.

Cronquist, A. 1981 *An integrated system of classification of flowering plants*. New York: Columbia University Press.

Dahlgren, R. M. T., Clifford, H. T. & Yeo, P. F. 1985 *The families of monocotyledons: structure, evolution and taxonomy*. Berlin: Springer-Verlag.

Goudet, J. 1998 Tests of key innovation and Fisher procedure to combine probabilities. *Am. Nat.* (Submitted.)

Harrison, R. G. 1991 Molecular change at speciation. *A. Rev. Ecol. Syst.* **22**, 281–308.

Mabberley, D. J. 1993 *The plant-book*. Cambridge University Press.

Manly, B. F. J. 1986 *The statistics of natural selection on animal populations*. London: Chapman and Hall.

Mayr, E. 1954 Change of genetic environment and evolution. In *Evolution as a process* (ed. J. Huxley, A. C. Hardy & E. B. Ford), pp. 157–180. London: George Allen & Unwin.

Nee, S., Barraclough, T. G. & Harvey, P. H. 1996 Temporal changes in biodiversity: detecting patterns and identifying causes. In *Biodiversity, a biology of numbers and difference* (ed. K. J. Gaston), pp. 230–252. Oxford: Blackwell Scientific Publications.

Sanderson, M. J. & Donoghue, M. J. 1996 Reconstructing shifts in diversification rates on phylogenetic trees. *Trends Evol. Ecol.* **11**, 15–20.

Savolainen, V., Manen, J.-F. & Spichiger R. 1997 Polyphyletism of celastrales deduced from a chloroplast noncoding DNA region. *Molec. Phylogenet. Evol.* **7**, 145–157.

Slowinski, J. B. & Guyer, C. 1993 Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. *Am. Nat.* **14**, 1019–1024.

Soltis, D. E. (and 15 others) 1997 Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Ann. Missouri Bot. Gard.* **84**, 1–49.

Watson, L. & Dallwitz, M. J. 1991 The families of angiosperms: automated descriptions, with interactive identification and information retrieval. *Aust. Syst. Bot.* **4**, 681–95.

Wilcoxon, F. 1945 Individual comparisons by ranking methods. *Biometrics* **1**, 80–83.