Formalizing reliability in the taxonomic congruence approach

Blaise Li & Guillaume Lecointre

Submitted: 22 February 2008
Accepted: 25 July 2008
doi:10.1111/j.1463-6409.2008.00361.x

Li, B. & Lecointre, G. (2009). Formalizing reliability in the taxonomic congruence approach. — *Zoologica Scripta*, *38*, 101–112.

In the 'total evidence' approach to phylogenetics, the reliability of a clade is implicitly measured by its degree of support, often embodied in a robustness index such as a bootstrap proportion. In the taxonomic congruence approach, the measurement of reliability has been implemented by various consensus or supertree methods, but was seldom explicitly discussed as such. We explore a reliability index for clades using their repetition across independent data sets. All possible combinations of the elementary data sets are used to compose the sets of independent data sets, across which the repetitions are counted. The more a clade occurs across such independent combinations, the higher its index. However, if other repeated clades occur that are incompatible with that clade, its index is decreased to take into account the uncertainty resulting from conflicting hypotheses. Results can be summarized through a greedy consensus tree in which clades appear according to their repetition indices. This index is tested on a 73 acanthomorph taxa data set composed of five independent molecular markers and multiple combinations of them. On this particular application, we confirm that reliability as defined here and robustness (estimated by bootstrap proportions obtained from a 'total evidence' approach) should be clearly distinguished.

Corresponding author: *Guillaume Lecointre, Équipe 'Phylogénie', UMR 7138 'Systématique, Adaptation, Évolution', Museum National d'Histoire Naturelle, Département Systématique et Évolution, case postale 26, 57 rue Cuvier, 75231 Paris cedex 05, France. E-mail: lecointr@mnhn.fr*
Blaise Li, Université Paris VI — Pierre et Marie Curie, UMR 7138, 43, rue Cuvier, Paris, France, 75005. E-mail address: blaise.li@normalesup.org

Introduction

With the increasing amount of molecular data available for phylogenetics comes an increasing hope for more extensive and well-resolved phylogenies. Indeed, the more diverse the sources of evidence, the better the expected quality of the results (Hempel 1965; Mahner & Bunge 1997), provided that the evidence is relevant to the problem under focus (Carnap 1950; Lecointre & Deleporte 2005). Quality is usually measured by some support values attached to the nodes of a phylogenetic tree. Support may be robustness (resistance to data perturbation), sensitivity (resistance to variations in the analysis method) or other kinds of measures such as decay indices or, in a supertree context, the measures proposed by Bininda-Edmonds (2003) or Cotton *et al.* (2006) (see also Wilkinson *et al.* 2003). The better the support values, the more the phylogeneticist will consider that the relationships are reliable. However, not all support measures are equally relevant to reliability assessment. The purpose of the present article is to propose a support value, the repetition index, which is designed to provide an appropriate measure of reliability for clades.

Materials and methods

Approach

In line with Carnap's degree of confirmation, reliability is the degree of credit we give to a statement at a given time, ideally taking into account all available data and knowledge relevant to this statement. In phylogenetics, the reliability issue cannot be addressed without considering how multiple data sets are handled: are all available data combined in a single matrix, or not? What are the criteria for considering a given clade reliable in each approach?

In the approach consisting in combining all the data, often called 'total evidence', one tends to trust the clades obtained inasmuch as they are based on the 'coherence' (see Rieppel 2004a,b; Kearney & Rieppel 2006) of all available characters. In the most common 'total evidence' practice, the reliability of a clade is implicitly (or even explicitly; Douady *et al.* 2003) associated with its degree of support, often measured with a Bremer support, a bootstrap proportion or a Bayesian posterior probability. Indeed, as all the available data have been gathered into a single matrix, reliability cannot be obtained otherwise.

In the 'taxonomic congruence' approach, the naturalness of data partitions is justified by positive biological knowledge (Miyamoto & Fitch 1995). The biologist fully recognizes the background knowledge justifying why a given gene can be considered independent of another one.[1] After the separate analyses, the results do not obligatorily end up with a strict consensus tree: actually, there are many ways to summarize the results (Bryant 2003). The issue about how to assess the reliability of a clade in a taxonomic congruence approach has received rather poor explicit interest until recently. In a way, most consensus techniques are implicitly extracting those clades we might have good reasons to trust. However, the term 'reliability' has never been used for that, except in rare cases (Lockhart *et al*. 1995, p. 673; Bryant 2003, p. 5; Brinkmann *et al*. 2005; Lecointre & Deleporte 2005). To have access to reliability, one must take into account other criteria than the pure global 'coherence' among individual characters used in the 'total evidence' approach. Reliable results are results that are supported by congruence among multiple independent relevant sources of information (Rodrigo *et al*. 1993; Rieppel 2004a; Lecointre & Deleporte 2005; see also Grande 1994). One should for instance consider corroboration among trees produced by the analyses of genes that are hypothesized to evolve independently.[2] This approach has been explicitly used by Chen *et al*. (2003) and Dettaï & Lecointre (2004, 2005) but without full formalization. Others (Bininda-Edmonds 2003; Seo *et al*. 2005; Wilkinson *et al*. 2005; Cotton *et al*. 2006; Moore *et al*. 2006) have devised procedures that, under certain assumptions of independence of source trees, could include some reliability evaluation, but without explicitly using this word, using the more general term 'support' instead. We will now examine how the concept of reliability we described here could be formalized so as to be computerized into a repetition index for clades.

Taxonomic congruence from independent data sets

Among the scientific community, credit is given to phylogenetic hypotheses that have been obtained from independent data sets and teams. The more a clade is recovered by the analyses of independent data, the more it is reliable (for a similar idea developed in a supertree perspective, see Pisani & Wilkinson 2002, p. 154). Independent *elementary data sets* have thus to be delineated, the set of which defining what we call the *elementary partitioning scheme* of the available data.

Independence of the data sets is important because there are many reasons why the tree obtained by the analysis of a particular data set might not represent accurately the species interrelationships. Each data set analysis might yield a tree that somewhat differs from the species tree (Maddison 1997). However, the hope is that the trees do not all differ in the same way if they are built from independent data sets. By '*independent*' we mean 'unlikely to be subject to the same causes of incongruence with respect to the true species tree'. The decision whether two data sets can be kept separate or not is based on biological background knowledge, for instance, knowledge pertaining to the functions of the genes used as evolutionary markers, or about strong differences in evolutionary pressures (differences in free mutational space, composition bias, etc., suggesting that resulting tree reconstruction artefacts will not be the same). The analyses of two genes will unlikely yield similar results by pure chance. And since the genes are supposed to be independent, similar results should not be caused by the same artefact, but by the shared feature of the genes: the common ancestry of the taxa bearing them. For some molecular phylogenetic markers, little information may be available. In such cases, when there is *a priori* no reason to suspect that two markers are not independent, the practitioner might want to take the risk to suppose independence. The elementary data sets are the data sets that cannot be further split into independent data sets.

Once the independent data sets have been defined, they should be analysed separately with the appropriate method. Then, the number of occurrences of a clade among the obtained trees is a first indication of its reliability.[3] Starting from this basic indicator, the repetition index may now be refined as follows.

Improving reliability by considering partial data combinations

One of the criticisms against separate analyses is that partitioning data favours stochastic errors. Indeed, trees from smaller data sets are usually more sensitive to stochastic effects of homoplasy than trees from larger ones. As a result, some clades could fail to be repeated because of this 'size effect' (see how Pisani & Wilkinson 2002, p. 153, discuss weak and strong phylogenetic signal). The problem can be partly avoided by examining trees obtained by partial combinations of the elementary independent data sets (Dettaï & Lecointre 2004). If these elementary data sets are A, B and C, and are analysed separately, they constitute the *elementary partitioning scheme*. Their *partial combinations* are $A \cup B$, $A \cup C$ and $B \cup C$, each of which is a data set that can be analysed and compared with the results of the analyses of other independent data sets (C, B and A, respectively). Using these partial combinations, other *partitioning schemes* can be defined, where the

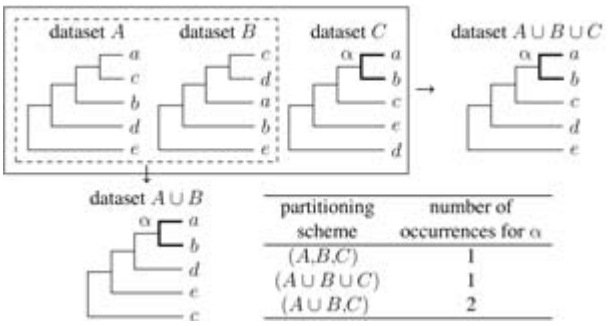
[1] *The use of biological knowledge is not restricted to taxonomic congruence approaches; knowledge in molecular evolution, for instance, is used in sophisticated model-based 'total evidence' practices.*

[2] *This implies the use of some external knowledge: the knowledge justifying the independence of the data partitions.*

[3] *Note that counting the occurrences of a clade is easy when all data sets have the same set of taxa, but otherwise not. This will be discussed later.*

elementary data sets are associated into sets of independent data sets. Here, those partitioning schemes would be ($A \cup B$,C) ($A \cup C$,B) and (A,$B \cup C$). Note that within a partitioning scheme, the results of the analyses can be compared to evaluate reliability because the constituting data sets are assembled so as to be independent: no elementary data set is present twice in a partitioning scheme. ($A \cup B \cup C$) is also a partitioning scheme to take into account, but since it has only one data set, it cannot provide numbers of occurrences higher than 1.

The elementary partitioning scheme, (A,B,C), is the one with the maximum number of independent data sets, but those data sets are the most prone to stochastic effects. The potential interest of the partial combination approach is to limit the stochastic effects usually impairing taxonomic congruence, as illustrated by the following example. Suppose that in the real species tree, there is a clade α that the analysis of data set C recovers, but that the analyses of data sets A and B separately fail to recover. Combining A and B might overcome the biases and stochastic effects that prevented α from being recovered in the separate analyses. In that particular case, using the ($A \cup B$,C) partitioning scheme provides 2 occurrences for clade α, whereas it occurred only once in the elementary partitioning scheme (see Fig. 1). For a given clade, the maximum number of independent occurrences will be obtained for a particular partitioning scheme. This scheme probably achieves the optimal way of combining the elementary data sets regarding the signal supporting the clade under focus. Indeed, a high number of occurrences is something improbable if no signal is supposed. The fact that the clade is repeated is better explained if one supposes that the combinations present in the partitioning scheme allowed common signal to emerge above noise. Therefore, the reliability of this clade should be derived from this partitioning scheme. It involves sufficient combination to overcome stochastic effects, but not too much; this allows to 'test' the clade across independent trees. Indeed, in too big a combination, there are fewer possibilities of independent occurrences, and there is even a risk for the clade to be lost because of a strong bias in one of the elementary data sets. The partial combination approach can be seen as a means of extracting more information from the data when some of the elementary data sets have weak phylogenetic signals. However, this is a computationally intensive procedure when the elementary data sets are numerous (with n elementary data sets, there are 2^n–1 analyses to do, including the total combination). In such a case, the method described in the rest of this article could also be applied using only the elementary partitioning scheme, hoping that a majority of the data sets express their historical signal.



partitioning scheme	number of occurrences for α
(A,B,C)	1
(A ∪ B ∪ C)	1
(A ∪ B,C)	2

Fig. 1 Simplified example illustrating the potential usefulness of the partial combination approach. Considering only the elementary partitioning scheme (A,B,C), clade α is only recovered by one data set. With the partial combination approach, all possible partitioning schemes are explored (though we show only three of them here). Among them, there is one in which two independent data sets recover clade α. The combination of data sets A and B has overcome stochastic effects that prevented clade α from being recovered when these data sets were kept separate. The repetition index for α will thus be based on partitioning scheme ($A \cup B$,C).

To summarize, a provisional repetition index can be computed the following way:

1 separate the data into independent elementary data sets;

2 analyse every elementary data set and every possible partial combination of them;

3 for each partitioning scheme (i.e., for each set of independent data sets), count the number of occurrences of each clade that appeared in at least one of the analyses (this number cannot be higher than the number of data sets in the considered partitioning scheme, thus, its maximal possible value is the number of elementary data sets);

4 for each clade recorded in step 3, retain as repetition index the best number of occurrences obtained among all possible partitioning schemes (the partitioning scheme providing the highest repetition for a clade may be different from the one providing the highest repetition for another one).

This provisional repetition index can be expressed by the following formula:

$R(\alpha) = \max_{D \in PSc}(\Sigma_{d \in D}\ \delta_{\alpha,d})$, where PSc is the set of the partitioning schemes, D is the set of the data sets constituting a partitioning scheme in PSc, d is a data set in D, and δ is 1 if α is produced by the analysis of the data set d, 0 otherwise.

It is basically a number of occurrences of a clade in a set of independent analyses, hence the sum over data sets. The more independent data sets there are, the highest the reliability may be. This justifies the use of a sum. What is to be summed however, is subject to discussion: bootstrap proportions, percentages in majority-rule consensuses of equally optimal trees, Bayesian posterior probabilities, raw all-or-nothing occurrences? All these possibilities lead to a repetition index, which has the dimension of a number of occurrences. Here, we simply use occurrences (the 1 or 0 represented by δ), but the methodology presented here could be equally applied using the other options.

This provisional repetition index can be computed for bipartitions in unrooted trees instead of clades. It can also be used in cases where some taxa are missing from some of the elementary data sets. However, in that case, the values of the repetition indices for the clades containing taxa missing from some data sets could be lower because of the lack of taxonomic overlap; some data sets would be unable to produce those clades. For the sake of simplicity, we will now suppose that there is a total taxonomic overlap between the elementary data sets.

Dealing with contradiction among clades

For some reason (a mistake in the delineation of the elementary data sets because of a lack of biological background knowledge, for instance) two clades can be incompatible but both repeated. In such a case, at least one of the two clades certainly does not reflect the history of the taxa (we neglect here reticulate evolution); it should not be considered reliable. Without any other assumption, one cannot tell which one of the two clades is not 'correct' — and perhaps both are incorrect. Therefore, the reliabilities of both clades should be decreased. We suggest decreasing the value of the repetition index of a clade by subtracting the repetition index of its contradictor. This would lead to an index that is basically a difference between numbers of occurrences. Actually, a clade is likely to have multiple contradictors among all clades occurring at least once through the analyses of all elementary data sets and their partial combinations. It seems meaningful to consider only the most reliable of them, that is, the one with the highest repetition index. Note that this requires that the repetition indices of all the clades contradicting the clade under focus are known, which necessitates successive approximations because the indices of the contradictors depend on the indices of their own contradictors. This can be formalized as follows.

The formerly defined repetition index, $R(\alpha)$, will be called the *first order repetition index* for clade α and noted $R_1(\alpha)$.

A clade β is said to be contradicting α if the three following conditions are true (see Berry & Gascuel 2000, p. 275):

1 $\alpha \cup \beta \neq \alpha$ (β contains at least one taxon that α has not);

2 $\alpha \cup \beta \neq \beta$ (β does not contain α);

3 $\alpha \cap \beta \neq \varnothing$ (β contains at least one taxon that α has).

Contradiction is a reciprocal relationship. It could be reformulated this way: α and β contradict one another if and only if they have at least one shared taxon and at least each a specific taxon. Contradiction is incompatibility. Clades that contradict one another are clades that are not compatible and reciprocally; they cannot be both in the same tree. Such a relationship is straightforward when the clades come from trees having the same leaves, but it is problematic when there are missing taxa (see Bininda-Edmonds 2003). We don't know where the missing taxa would be placed if they were present, and one could imagine cases where the addition of taxa to clades can make them switch from contradiction to compatibility or vice versa [see Wilkinson *et al*. (2005) for a description of the different possible situations]. This explains why the present article is restricted to cases with fully overlapping taxonomic samplings.

Assuming that α has some contradictors, let β_1 be its *'best' contradictor* according to the R_1 index (the one with the highest R_1). A second order repetition index for α can now be defined:

$$R_2(\alpha) = R_1(\alpha) - R_1(\beta_1)$$

However, according to R_2, β_1 might not be the best contradictor of α any more because it is also contradicted. It is possible that there is another contradictor of α, β_2, that is not so much contradicted as β_1 is, so that $R_2(\beta_2) > R_2(\beta_1)$. Thus, the calculation of the repetition index for α should be reconsidered by defining a third order repetition index:

$R_3(\alpha) = R_1(\alpha) - R_1(\beta_2)$, where β_2 is the best contradictor of α according to R_2. If there is more than one best contradictor according to R_2, the highest R_1 value found among these contradictors is used to calculate R_3.

This can be repeated, calculating a fourth order repetition index considering R_3 to find the next best contradictor, and so on. At each step of the process, a contradiction network in which each clade has a provisional best contradictor is implicitly established. In the most simple case, each clade will have a stable unique best contradictor, allowing the calculation of a *final repetition index* $R_f(\alpha) = R_1(\alpha) - R_1(\beta_f)$, where β_f is the best contradictor of α according to R_f (see Table 1).

In the other cases, the process of calculating the next order repetition indices will be periodic: as the number of clades is finite, the contradiction network between them has a finite number of possible configurations, so if none of these networks is stable, the state of the system will change until it comes back to a state already reached before. In such cases, since we cannot tell for each clade which contradictor is the best, we consider that each configuration of the contradiction network in a period ought to be taken into account with the same weight in the determination of the final repetition index. The mean repetition index, $\bar{R}$, over a period will thus be taken as final repetition index. This amounts to decreasing $R_1(\alpha)$ by the mean value of R_1 over the successive best contradictors of α.

Application to acanthomorph phylogeny

Data sets. Five independent elementary data sets with a common taxonomic sampling of 73 taxa have been gathered as a case study (Dettaï 2004). The data sets are the following molecular markers:

1 a mitochondrial data set comprising partial 12S and 16S rDNA for a total length of 828 base pairs (bp). They are kept together because both are elements of the mitochondrial ribosome, thus potentially subject to common evolutionary

Table 1 Example illustrating the computation of the repetition index taking into account contradiction among clades. The taxa involved in this example are designed by the letters a through *j*. It is assumed that there are only five clades and that their first order repetition indices (R_1) have already been calculated. The contradictors of α are β_1 and β_2, the best one being β_1 ($R_1(\beta_1) = 4$). The contradictors of β_1 are α and γ_1. The contradictors of β_2 are α and γ_{12}. After calculating the second order repetition indices (see the procedure in the text), the best contradictor for α has changed: it is now β_2 ($R_2(\beta_2) = 0$). After calculating the third order repetition index, no best contradictor has changed. This index can thus be taken as the final repetition index.

Clades	$\alpha=(a,b,c,d)$	$\beta_1=(a,b,e)$	$\beta_2=(c,d,f)$	$\gamma_1=(e,g,h)$	$\gamma_2=(f,i,j)$
R_1	3	**4**	3	5	3
Best contradictor	$\boldsymbol{\beta_1}$	γ_1	γ_2	β_1	β_2
R_2	3 – 4 = –1	4 – 5 = –1	3 – 3 = **0**	5 – 4 = 1	3 – 3 = 0
Best contradictor	$\boldsymbol{\beta_2}$	γ_1	γ_2	β_1	β_2
R_3	3 – 3 = 0	4 – 5 = –1	3 – 3 = 0	5 – 4 = 1	3 – 3 = 0
Best contradictor	β_2	γ_1	γ_2	β_1	β_2

constraints, and physically linked, being both on the mitochondrial chromosome;
2 partial sequences of 28S rDNA (C1-C2, D3, D6 and D12 domains). The concatenated length is 801 bp;
3 partial Rhodopsin gene (759 bp);
4 partial Mixed Lineage Leukaemia-like exon 26 (MLL, 552 bp);
5 partial Interphotoreceptor Retinoid Binding Protein module 1 (IRBP, 713 bp).

Bathypterois (Chlorophtalmoidei) was the only outgroup taxon for which we could gather the sequences for the five elementary data sets.

Tree construction method. The alignment was done by hand with SeAl (Rambaut 2002) v2.0a11 carbon. Ambiguous zones in the rDNA data sets were removed. The alignment was submitted to TreeBase (study accession number: S2152, matrix accession number: M4084).

There are 31 possible combinations of the five elementary data sets (including the 'total evidence' combination).

They were successively analysed under maximum parsimony using PAUP* (Swofford 2002) version 4.0b10 for Macintosh (PPC), each with 1000 RAS + TBR rounds, using a nexus batch file (see the nexus file on TreeBase).

The 50% majority-rule consensus trees were used to count the occurrences of the clades. The full combination was bootstrapped (1000 pseudosamples each submitted to 50 replications of RAS + TBR, 'multrees' option turned off) to compare robustness and reliability as defined here. Further data processing was done on a GNU/Linux system with the help of shell scripts and Python 2.3.4 (http://www.python.org/) scripts.

Results

Three clades occurred five times, and obtained the maximal repetition index ($R_f = 5$). These are the clades that occur for each elementary data set. Six clades occurred four times and were only slightly contradicted ($R_f = 3$). 10 other clades reached a repetition index of 2, and a total of 35 clades had a repetition index equal to or higher than 1. All these clades could be provisionally considered reliable (until new data is available) because they occur at least once more than their 'best' contradictor.

The 'total evidence' tree is presented in Fig. 3. The majority-rule consensus of the trees resulting from the separate analyses of the 5 elementary data sets stands for the result of a typical taxonomic congruence study (see Fig. 4). The raw numbers of occurrences of the clades, their repetition indices and their bootstrap supports from the full data combination are written on the trees.

The repetition indices of the clades were plotted against their bootstrap supports. Both the Spearman and Kendall rank correlation coefficients between repetition indices and total evidence bootstrap supports were –0.33.

To synthesize the results concerning acanthomorphs relationships, several methods are possible to build summary trees based on clades reliabilities. One could build a clade-taxon matrix, where each clade recorded from the phylogenetic analyses of the elementary data sets and their partial combinations is weighted according to its repetition index, and each taxon is coded 1 when present in the clade and 0 when not. This matrix could then be analysed under maximum parsimony or compatibility, leading to trees akin to MRP or MRC supertrees. We propose another summary tree, explicitly devised to include reliable clades. It is akin to the greedy consensus method (see Bryant 2003 and Bandelt & Dress 1992, p. 244):
1 group clades having the same repetition index, arrange these groups in descending order of repetition index. Within each group, group clades according to the maximum number of occurrences and order these groups according to this criterion;
2 for each group of clades having equal repetition index and equal maximum number of occurrences, beginning with the

'best' one (the most reliable), eliminate the clades within the group that are not compatible with the clades already retained. Then, retain the remaining clades of the group if they are mutually compatible and repeat this step with the next group. If the remaining clades in a group are not mutually compatible, they are all discarded. This process should not discard clades with high repetition indices because a clade having contradictors with the same repetition index and same number of occurrences has its best contradictor being at least as reliable as it is. This ensures that it does not have a high repetition index (proof in the case where all clades have a stable best contradictor available as supplementary material); **3** assemble the 'greedy summary tree' by combining the clades that have been retained.

The resulting tree is presented in Fig. 2. In this synthesis tree, one clade was not present in the tree obtained from the 'total evidence' combination, and two clades were absent from all five separate analyses. The elementary partitioning scheme provided the highest number of occurrences for 18 out of the 35 clades present in this tree. The other partitioning schemes contributing the most to the greedy summary tree are ones implying three elementary data sets and one partial combination involving 28S; with 12S and 16S (highest number of occurrences for 12 clades), with IRBP (11 clades) and with MLL (11 clades).

Discussion

A few properties of the repetition index

The partial combination approach leads us to having different sets of independent data sets over which to sum occurrences: the partitioning schemes. The repetition index is based on occurrences within a partitioning scheme. The maximum value potentially can be achieved with the elementary partitioning scheme because it is the one with the highest number of independent data sets. Thus, the maximum value of the repetition index is the number of elementary data sets. For instance, in the present study, the maximum repetition index cannot be 31, but five, as not all possible combinations can be in the same partitioning scheme (MLL + IRBP is not independent from MLL + Rhodopsin). The 'best' clades are those that are recovered by the analyses of every elementary data set.

The minimum possible value of the repetition index is the opposite of the maximum value. That would be the case for a clade that never occurs, and that is contradicted by one of the best clades.

The more elementary data sets are combined within a partitioning scheme, the better the chances to overcome stochastic errors, but the less there are independent data sets in the partitioning scheme. So if a clade fails to appear in the analyses of each elementary data set, the partial combination approach has no chance to give it the maximum value. However, this maximum can increase by the addition of new independent data sets. This raises the question of how to compare that index among different studies. One could think that it is necessary to divide the index by the number of elementary data sets included in the study to allow comparison among studies dealing with different numbers of data sets. The maximum possible value of the repetition index would then be 1, whatever the quantity of data and trees at hand. This seems not appropriate because adding more data should allow an improvement of the maximum reliability. The repetition index we propose reflects the quantity of trees supporting a clade, which is a relevant information. Actually, the indices from different studies can be compared without rescaling. Suppose we have made a study comprising three data sets. If we obtain a particular clade with 2 of the data sets and a contradictor with the other data set, the repetition index of that clade will be 1 (2 occurrences minus 1 contradiction). The reliability of this clade is low, and it cannot be higher as long as we do not analyse new data. But meanwhile, it is still more reliable than any of its contradictors. Now, suppose we add 10 new data sets to the study. In case most new data recover the clade, its reliability should increase, which will be reflected by an increase in its repetition index. However, if five of the new data sets support the clade and the five other support its contradictor, the reliability should not be improved. This will be reflected by the fact that the repetition index of the clade would still be 1 (7 occurrences minus 6 contradictions). The reliability will not be decreased either. A repetition index of 1 indicates the same level of reliability (rather low, but positive nonetheless) whatever the number of data sets used in the study. What changes is that with such a persistent low reliability, we may now hypothesize that there is some conflict between two true historical signals; this could be a sign of reticulate evolution. Monitoring the evolution of repetition indices when the number of data sets grows could be done following a procedure similar to the one devised by Struck *et al*. (2006) in a 'total evidence' context.

Robustness and reliability

The results show that, assuming that reliability and robustness can be assessed by our repetition index and by bootstrap proportions from the full combination, respectively, those two pieces of information about the results are poorly correlated. Indeed, if we consider reliable clades that have a repetition index equal or higher than 1 and robust those with a 70% or higher bootstrap support (value chosen according to Hillis & Bull 1993 and Lecointre *et al*. 1994), 30 reliable clades are not robust and 1 robust clade is not reliable. Thus, although it may not always be easy to justify the independence of the data sets, we are inclined to think that using only a total evidence approach is not suitable, because it dismisses interesting information; it does not allow us to determine which clades

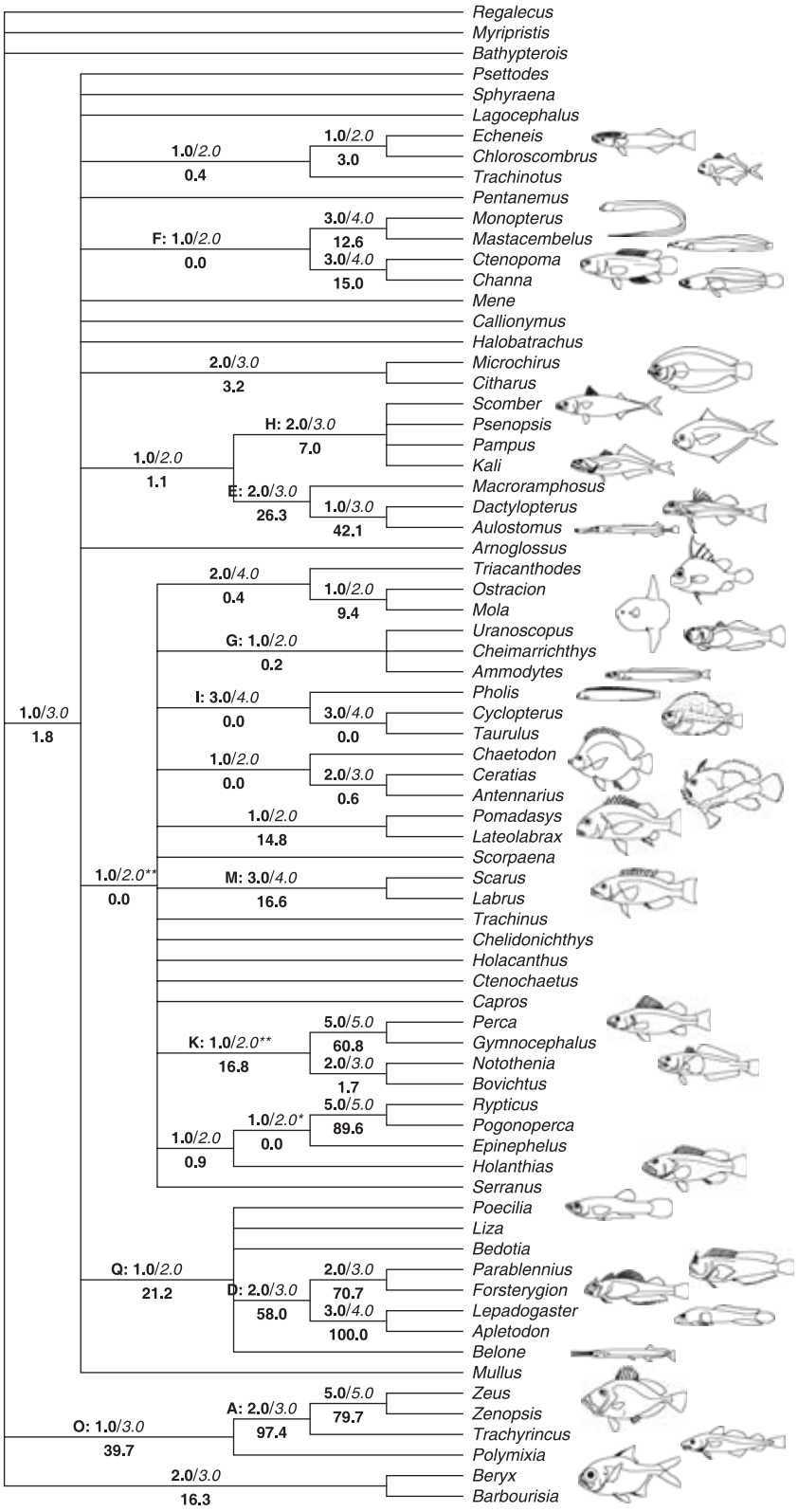


Fig. 2 Tree summarizing the most repeated clades with the greedy summary method described in the text. This tree is composed of the 'best' intercompatible clades according to the repetition index. The bootstrap supports in the 'total evidence' analysis are in bold, below the branches. The repetition indices (in bold) and the maximum number of occurrences of the clades are above the branches. *: the simultaneous analysis of all data sets did not recover this clade. **: these clades could not be recovered without combining at least two elementary data sets. Some reliable clades are lettered according to Dettaï & Lecointre (2005).

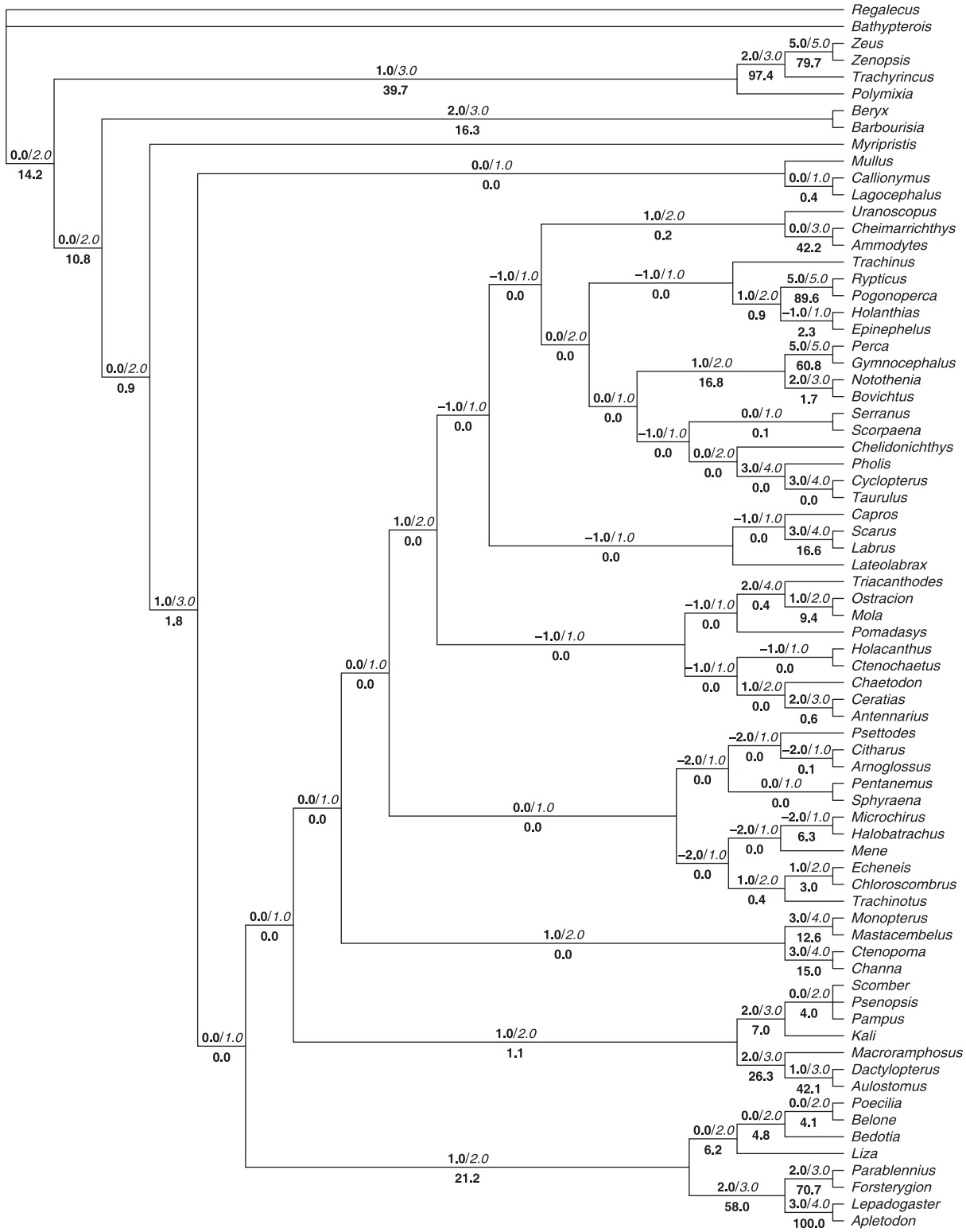


Fig. 3 50% majority-rule consensus of the equally most parsimonious trees obtained by the analysis of the combination of all five elementary data sets. The bootstrap supports are in bold, below the branches. Above the branches are the repetition indices (in bold) and the maximum number of occurrences of the clades.

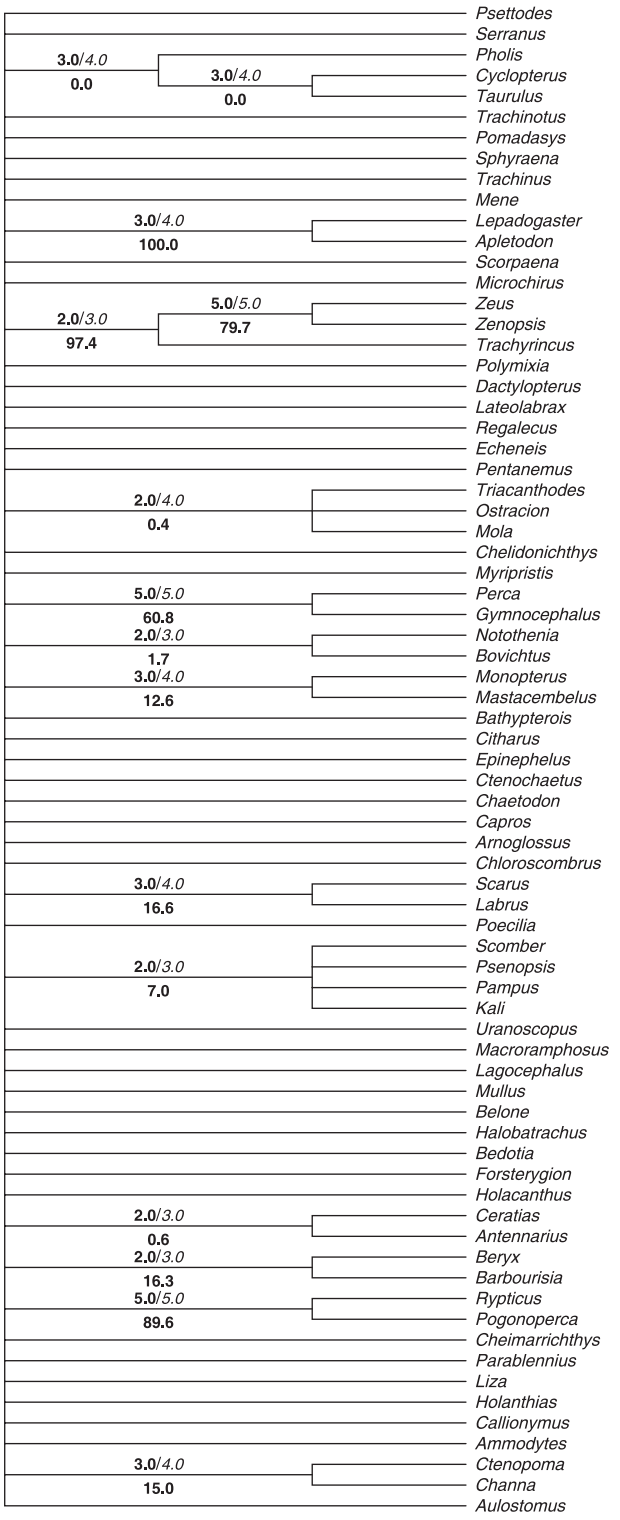


Fig. 4 Majority-rule consensus of the consensus trees obtained by the analyses of the five elementary data sets. The bootstrap supports in the 'total evidence' analysis are in bold, below the branches. The repetition indices (in bold) and the maximum number of occurrences of the clades are above the branches.

may be trusted to represent patterns resulting from species history and which may not. Sometimes, robust clades may be misleading.

Many authors, by using bootstrap proportions obtained on the full data combination, exploit the advantage of positive effects of sequence length on bootstrap proportions and may thus be inclined to trust more clades than they should. Indeed, strong bootstrap proportions are not necessarily linked to species common ancestry, even in the tree based on all available data. For instance, if the genes history is not the same as the species history (horizontal transfer, paralogy ...), or when compositional biases or long-branch attraction artefacts in a single data set are strong enough to impose the wrong topology to the tree based on the full combination (Chen *et al*. 2003; Phillips *et al*. 2004; Brinkmann *et al*. 2005). In that sense, the present method is more conservative than the 'total evidence' approach. Only the clades more repeated than their contradictors are considered reliable.

Relation to previous works

It should be noted that in Dettaï & Lecointre (2004) there was no explicit notion of contradiction among clades as in the present article. Instead, these authors used a concept of 'intruders' and 'escapees'. When a clade is repeated in several data sets, another data set could exhibit the 'same' clade minus one taxon (an 'escapee'), or plus one taxon (an 'intruder'). This allowed them to take into account clades that were 'almost the same' as a repeated clade. However, this approach seemed difficult to formalize in a way that could be implemented into a computer program. A way to circumvent this intruder and escapee issue would be to consider reduced components (also called n-taxon statements by Wilkinson (1994, 1996), or partial splits). In our present approach, clades that are almost the same will be detected by the fact that they are not strongly contradicted, and not by the fact that they are 'almost repeated' (i.e., having 'intruders' or 'escapees') as in Dettaï & Lecointre (2004). When a taxon escapes from a reference clade, the result is a clade that is compatible with the reference clade, but the escapee will be part of a clade that contradicts it. If the same taxon escapes several times from the reference clade and participates in the same contradictory clade [a repeated position in Dettaï & Lecointre (2005)], there will be a 'good' contradictor for the reference clade. Otherwise, the clade will not be really contradicted, it will only be less repeated.

Bininda-Edmonds (2003) and Wilkinson *et al*. (2005) have devised support measures for clades in supertrees that include developed considerations about compatibility and contradiction. Their measures are based on support of supertree relationships from source tree relationships. These measures can thus be interpreted as reliability measures when the source trees of the supertree analysis are based on independent data sets

(which is often the case since one usually aims for accurate supertrees). They present the interest of being applicable even when some taxa are missing. Their works, however, are focused on the clades already present in the supertree under study whereas in the present article, reliability assessment precedes summary tree building. Another idea that produces some sort of numerical reliability assessment is the bootsrap-derived procedures used by Seo *et al*. (2005) or Moore *et al*. (2006). Their approaches consist in resampling genes or source trees, respectively. This can provide an implicit reliability aspect to bootstrap proportions if the genes or source trees from which the resampling is made can be hypothesized to be independent from one another.

How to synthesize the results?

Synthesizing — in the taxonomic congruence framework — the results concerning the reliability of clades into a tree can be seen as a form of consensus construction. Simply using the majority rule consensus of the results of the separate analyses is, however, too conservative (see how Fig. 4 is poorly resolved) and does not take into account the information added by the partial combination approach. The tree from the 'total evidence' analysis (see Fig. 3) seems to be more useful in our present test case; most reliable clades are recovered. However, the principles underlying its construction do not guarantee this: a strong bias from one data set could mislead the whole reconstruction and prevent a reliable clade from being present. That is one of the reasons for using consensus methods based on the repetition indices. Instead of simply mapping these indices on trees obtained from usual methods, we propose the greedy summary tree method described earlier, because it is designed for selecting the most reliable clades (see Bandelt & Dress 1992, p. 244 for a similar approach, but based on another support value). In addition, two methods based on matrix representation were used, to compare with that greedy summary tree. These methods weight the bipartitions in the matrix according to their repetition indices. One is derived from matrix representation with parsimony (MRP, Baum & Ragan 2004) and the other from matrix representation with compatibility (MRC, Rodrigo 1996; Ross & Rodrigo 2004). The second, also akin to an asymmetric median tree (AMT, Phillips & Warnow 1996) was unfortunately too time-consuming, using the clique program from Felsenstein (2004), to be applied successfully to our 73 taxa data set. Note that for that reason, Bryant (2003, p. 6) recommends to use the greedy consensus method. We resolved the problem the same way, using the greedy summary tree (Fig. 2). The difference between the greedy summary tree and the classical greedy consensus is that the former was constructed from the reliability indices of the clades, not their raw number of occurrences. The greedy summary tree was the same tree as the MRC-derived tree on a test with 16 taxa and six elementary data sets but was dramatically faster to construct. Our MRP-derived method differs from standard MRP mainly by the fact that contradiction among clades is taken into account before writing the clade-taxon matrix. A more MRP-like approach would have consisted in taking every clade occurrence as a column in the matrix and letting the parsimony analysis manage the contradictions. Here, we weight the clades according to their final repetition indices. This also makes the difference between our MRC-derived approach and standard MRC or AMT. The implications of such a difference remain to be studied.

Acanthomorph phylogeny

Some groups identified with letters by Dettaï & Lecointre (2005) are found here (see Fig. 2). The monophyly of the group 'A', comprising gadiforms (cods) and zeioids (dories) is recovered as reliable (R_f = 2.0), also confirming the findings of Chen *et al*. (2000, 2003) and Miya *et al*. (2003). The monophyly of the group 'O', uniting *Polymixia* (beardfish) to the previous groups, also found in the same previous studies, is considered here as reliable (R_f = 1.0). Clade 'F' is found again (R_f = 1.0), uniting channids (snakeheads), anabantoids (climbing gouramies), mastacembeloids (swamp eels) and synbranchiforms (spiny eels). Clade 'E' (R_f = 2.0) contains part of the syngnathiforms (pipefishes and horsefishes) and dactylopteriforms (flying gurnards). Clade 'H' (R_f = 2.0) groups parts of the trachinoids (*Kali*) and parts of the Scombroidei (here, the mackerel), with Stromateoidei (butterfishes). Clades 'E' and 'H' are sister-taxa (R_f = 1.0). Clade 'M' (R_f = 3.0) shows a sister-group relationship between labrids (wrasses) and scarids (parrotfishes). Dettaï & Lecointre (2005) showed that this component of the ex-labroids actually was not related to other labroids like cichlids. Clade 'K' (R_f = 1.0) is showing a sister-group relationship between Antarctic fishes (the Notothenioidei) and percids (perches). Clade 'I' (R_f = 3.0) groups cottoids (sculpins) with zoarcoids (eelpouts). Clade 'G' (R_f = 1.0) groups components of the Trachinoidei (stargazer, sandlances) and Cheimarrichthydae (torrentfishes). Clade 'Q' (R_f = 1.0) groups atherinomorphs (guppies), mugiloids (mullets), blennioids (blennies) and gobiesociforms (clingfishes), the latter two forming clade 'D' (R_f = 2.0). Some of those groups appeal the polyphyly of traditional taxa, most of them poorly defined (Perciformes, Scorpaeniformes, Trachinoidei, Labroidei, Paracanthopterygii).

Some clades from Dettaï & Lecointre (2005) are not recovered. Clade 'N', grouping lophiiforms (anglerfishes), tetraodontiforms (pufferfishes), chaetodontids (butterfly-fishes) and *Capros*, is not recovered, because one of the tetraodontiforms, *Lagocephalus*, is subject to recurrent long branch attraction (LBA). The same can be said of clade 'L' (carangids, menids, flatfishes, echeneids, centropomids, sphyraenids, polynemids), from which *Arnoglossus* — a bothid flatfish — 'escapes' because of its high mutation rate in several genes.

Recurrent LBA tends to influence the summary tree: in Fig. 2, *Arnoglossus*, *Callionymus*, *Mullus* and *Lagocephalus*, four taxa showing recurrent long branch attraction, are placed in an unresolved position within a large clade. Other clades cannot be recovered here probably because the present data set is reduced compared with that of the study of Dettaï & Lecointre (2005). Only the genera present in all data sets have been taken into account here because — for the moment — the method does not deal with missing taxa.

On the other hand, some clades not previously labelled with letters are reliable here, for example, the node splitting 'basal' acanthomorphs (*Regalecus*, clade 'O' and beryciforms) from the rest of the sampling (R_f = 1.0). Nodes of medium depth exhibit poor values (see Fig. 2).

Conclusion

The present index confirms many of the new acanthomorph clades repeatedly found by the recent molecular phylogenies. Work still needs to be done to extend the methodology to cases without perfect taxonomical overlap between data sets. It should also be noted that the repetition index can be misled by recurrent long branch attraction. This can probably be softened by using more elaborate reconstruction methods as the simple maximum parsimony that was used for the present article.

Acknowledgements

Thanks to Olaf Bininda-Edmonds, Nathanael Cao, Dan Faith, Jean-François Flot, François-Joseph Lapointe, Jérôme Murienne, Davide Pisani and Mark Wilkinson for excellent comments. Thanks to Mahendra Mariadassou for help with mathematics. Thanks to Wei-Jen Chen and Agnès Dettaï for most of the data used in this study.

B.L. was supported by a PhD fellowship 'Allocation couplée' (Ministère de l'Éducation Nationale, de la Recherche et de la Technologie).

The figures were produced using TikZ (http://sourceforge.net/project/pgf/) and TreeGraph (http://www.math.uni-bonn.de/people/jmueller/extra/treegraph/, Müller & Müller 2004). The pictures on Fig. 2 are drawn from Fishbase (Froese & Pauly 2006).

References

Bandelt, H.-J. & Dress, A. (1992). Split decomposition: a new useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, *3*, 242–252.

Baum, B. & Ragan, M. (2004). The MRP method. In O. Bininda-Edmonds (Ed.) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (pp. 17–34). Dordrecht, the Netherlands: Kluver Academic.

Berry, V. & Gascuel, O. (2000). Inferring evolutionary trees with strong combinatorial evidence. *Theoretical Computer Science*, *240*, 271–298.

Bininda-Edmonds, O. (2003). Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. *Systematic Biology*, *52*(6), 839–848.

Brinkmann, H., Van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G. & Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic Biology*, *54*(5), 743–757.

Bryant, D. (2003). A classification of consensus methods for phylogenetics. In M. Janowitz, F. Lapointe, F. McMorris, B. Mirkin & F. Roberts (Eds) *Bioconsensus* (pp. 1–21). Piscataway, NJ: American Mathematical Society Publications-DIMACS.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago press.

Chen, W.-J., Bonillo, C. & Lecointre, G. (2000). Taxonomic congruence as a tool to discover new clades in the acanthomorph (Teleostei) radiation. In *Program Book and Abstracts, 80th Annual Meeting ASIH, La Paz, México, June 14–20, 2000* (p. 369). American Society of Ichthyologists and Herpetologists, American Society of Ichthyologists and Herpetologists.

Chen, W.-J., Bonillo, C. & Lecointre, G. (2003). Repeatability of clades as a criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution*, *26*, 262–288.

Cotton, J., Slater, C. & Wilkinson, M. (2006). Discriminating supported and unsupported relationships in supertrees using triplets. *Systematic Biology*, *55*(2), 345–350.

Dettaï, A. (2004). *La phylogénie des Acanthomorpha (Teleostei) inférée par l'étude de la congruence taxinomique*. PhD Thesis. Université Paris VI Pierre et Marie Curie.

Dettaï, A. & Lecointre, G. (2004). In search of nothothenioid (Teleostei) relatives. *Antarctic Science*, *16*(1), 71–85.

Dettaï, A. & Lecointre, G. (2005). Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. *Comptes Rendus Biologies*, *328*, 674–689.

Douady, C., Delsuc, F., Boucher, Y., Doolittle, F. & Douzery, E. (2003). Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, *20*(2), 248–254.

Felsenstein, J. (2004). *PHYLIP. Phylogeny Inference Package*, Version 3.6. Seattle: Department of Genome Sciences and Departement of Biology, University of Washington.

Froese, R. & Pauly, D. (2006). Fishbase. World Wide Web electronic publication. Available Via http://www.fishbase.org.

Grande, L. (1994). Repeating patterns in nature, predictability, and 'impact' in science. In L. Grande & O. Rieppel (Eds) *Interpreting the Hierarchy of Nature*, 1st edn (pp. 61–84). New York: Academic Press.

Hempel, C. (1965). *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*. New York: Free press.

Hillis, D. & Bull, J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, *42*(2), 182–192.

Kearney, M. & Rieppel, O. (2006). Rejecting 'the given' in systematics. *Cladistics*, *22*, 369–377.

Lecointre, G. & Deleporte, P. (2005). Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, *34*(1), 101–117.

Lecointre, G., Philippe, H., Lê, H. L. V. & Le Guyader, H. (1994). How many nucleotides are required to resolve a phylogenetic

problem? The use of a new statistical method applicable to available sequences. *Molecular Phylogenetics and Evolution*, *3*(4), 292–309.
Lockhart, P., Penny, D. & Meyer, A. (1995). Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. *Journal of Molecular Evolution*, *41*, 666–674.
Maddison, W. (1997). Gene trees in species trees. *Systematic Biology*, *46*(3), 523–536.
Mahner, M. & Bunge, M. (1997). *Foundations of Biophilosophy*. Berlin: Springer.
Miya, M., Takeshima, H., Endo, H., Ishiguro, N., Inoue, J., Mukai, T., Satoh, T., Yamaguchi, M., Kawaguchi, A., Mabuchi, K., Shirai, S. & Nishida, M. (2003). Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, *26*, 121–138.
Miyamoto, M. & Fitch, W. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology*, *44*(1), 64–75.
Moore, B., Smith, S. & Donoghue, M. (2006). Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Systematic Biology*, *55*(4), 662–676.
Müller, J. & Müller, K. (2004). TreeGraph: automated drawing of complex tree figures using an extensible tree description format. *Molecular Ecology Notes*, *4*, 786–788.
Phillips, M., Delsuc, F. & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, *21*(7), 1455–1458.
Phillips, C. & Warnow, T. (1996). The asymetric median tree: a new model for building consensus trees. *Discrete Applied Mathematics*, *71*, 311–335.
Pisani, D. & Wilkinson, M. (2002). Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology*, *51*(1), 151–155.
Rambaut, A. (2002). *Se-Al, Sequence Alignment Editor*, Version, 2.0a11. Oxford, OX1 3PS, UK: Department of Zoology, University of Oxford, South Parks Road.
Rieppel, O. (2004a). The language of systematics, and the philosophy of 'total evidence'. *Systematics and Biodiversity*, *2*(1), 9–19.
Rieppel, O. (2004b). What happens when the language of science threatens to break down in systematics: a popperian perspective. In D. Williams & P. Forey (Eds) *Milestones in Systematics* (pp. 57–100). CRC Press.
Rodrigo, A. (1996). On combining cladograms. *Taxon*, *45*(2), 267–274.
Rodrigo, A., Kelly-Borges, M., Bergquist, P. & Bergquist, P. (1993). A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zealand Journal of Botany*, *31*, 257–268.
Ross, H. & Rodrigo, A. (2004). An assessment of matrix representation with compatibility in supertree construction. In O. Bininda-Edmonds (Ed.) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (pp. 35–63). Dordrecht, the Netherlands: Kluver Academic.
Seo, T.-K., Hirohisa, K. & Thorne, J. (2005). Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(12), 4436–4441.
Struck, T., Purschke, G. & Halanych, K. (2006). Phylogeny of Eunicida (Annelida) and exploring data congruence using a partition addition bootstrap alteration (PABA) approach. *Systematic Biology*, *55*(1), 1–20.
Swofford, D. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods)*, Version 4.0b10. Sunderland, Massachusetts: Sinauer Associates.
Wilkinson, M. (1994). Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology*, *43*(3), 343–368.
Wilkinson, M. (1996). Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution*, *13*(3), 437–444.
Wilkinson, M., Lapointe, F.-J. & Gower, D. (2003). Branch lengths and support. *Systematic Biology*, *52*(1), 127–130.
Wilkinson, M., Pisani, D., Cotton, J. & Corfe, I. (2005). Measuring support and finding unsupported relationships in supertrees. *Systematic Biology*, *54*(5), 823–831.

Supporting Information

Additional supporting information may be found in the online version of this article.

Figure 1. Repetition indices of the clades plotted against their bootstrap supports in the 'total evidence' analysis.

Table 1. Taxonomic sampling (in bold; sequences not present in Dettaï, 2004 or in Dettaï & Lecointre, 2005).