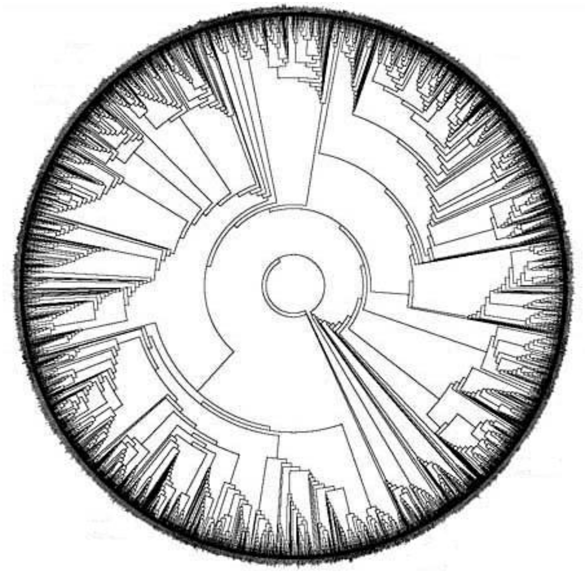
University of South Bohemia

Faculty of Science

Department of Zoology

Meta-analysis of methodological artifacts of the phylogenetic imbalance



Mgr thesis

Bc. Jana Smrčková

Supervisors: prof. RNDr. Jan Zrzavý, CSc.[1]

prof. RNDr. Tomáš Herben, CSc.[2]

České Budějovice, 2011

[1] University of South Bohemia, Faculty of Science, [2] Charles University in Prague, Faculty of Science

Smrčková, J., 2011: Meta-analysis of methodological artifacts of the phylogenetic imbalance. Mgr. Thesis, in English. – 38 p., Faculty of Science, the University of South Bohemia, České Budějovice, Czech Republic.

This work was supported by the Student Grant Agency of the Faculty of Science, University of South Bohemia (SGA 2009).

Annotation

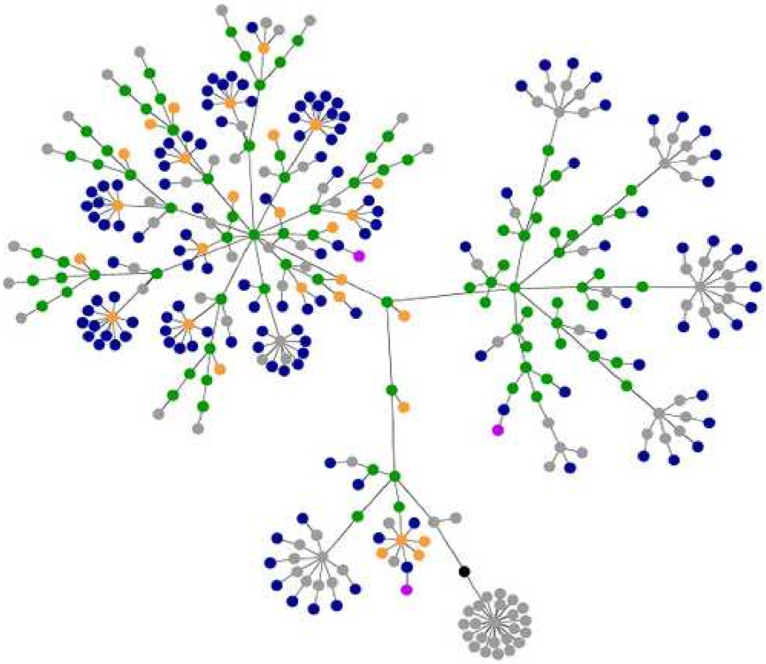
The influence of possible methodological artifacts (e.g. type of data or tree construction methods) on the tree topology was evaluated. A total of 413 phylogenetic trees was downloaded from the tree repository TreeBASE. Three indices of topology imbalance were employed, namely, Fusco & Cronk index, weighted average, and Colless index. The study reveals that methodological artifacts have probably a weak influence on the tree shape. Therefore, patterns in tree balance could reflect macroevolutionary processess, not a methodological bias.

Prohlašuji, že svoji diplomovou práci jsem vypracovala samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury. Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své diplomové práce, a to v nezkrácené podobě - v úpravě vzniklé vypuštěním vyznačených částí archivovaných Přírodovědeckou fakultou - elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

3. 1. 2011 v Českých Budějovicích

………………………………………………………………

Jana Smrčková



Acknowledgements

In particular, I would like to express my sincere gratitude to my supervisors prof. RNDr. Jan Zrzavý, CSc. and prof. RNDr. Tomáš Herben, CSc. I am grateful for an interesting and stimulating topic which I enjoyed studying and for an extensive support, advice, and for a great deal of patience with me. Moreover, I would like to thank Antonín Macháč and prof. RNDr. Václav Hypša CSc. for valuable advice. Last but not least, I am indebted to Laboratory of Molecular Phylogeny and Evolution of Parasites and its students for background and support.

Contents

Introduction 1

Materials and Methods... 4

Indices of the tree imbalance 4

The trees: download and data modifications 6

Data characteristics.. 7

Levels of qualitative variables 8

Analyses 8

Software 9

Results... 11

Data matrix and variables 11

Are imbalance indices dependent on size of a tree? 11

Is tree balance dependent on phylogenetic quality of characters? 14

Does tree balance differ between taxonomic groups?.. 16

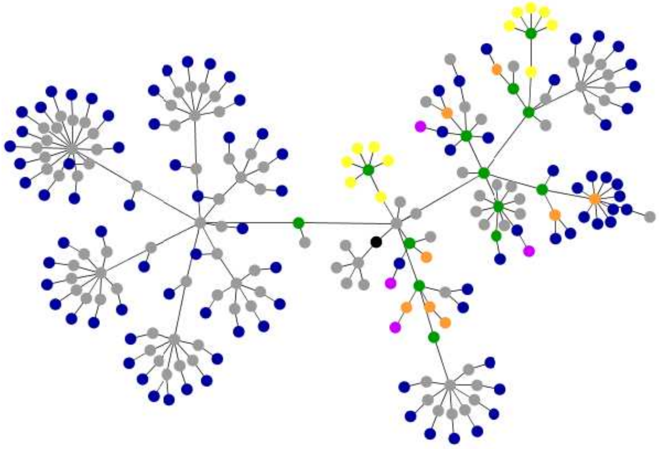
Is tree balance dependent of methodology and type of data? 17

Discussion.. 18

References 23

Online references... 27

Supplementary... 28



Introduction

The topology of phylogenetic trees is believed to reflect evolutionary history of the given organismal group (e.g. Heard 1992; Guyer & Slowinski 1993; Aldous 2001). One of the attributes of phylogenetic trees studied in analyses of macroevolutionary processes is the "tree balance", i.e. the symmetry of the tree nodes. Balance of phylogenetic trees has been hypothesized to reflect variation of the speciation and extinction rates (Heard 1996; Heath et al. 2008). If the probability of speciation and extinction did not vary substantially among phylogenetic lineages, the tree should be more or less symmetric and thus "balanced". If the speciation and extinction rates between the individual lineages differ, the tree is hence asymmetric and "imbalanced". The variation of the speciation and extinction rates is essential for the species selection / sorting (Vrba & Eldredge 1984). Many studies (Guyer & Slowinski 1991; Heard 1992; Guyer & Slowinski 1993; Mooers 1995; Heard 1996; Katzourakis et al. 2001) confirmed that real topologies are more imbalanced (asymmetric) than it would be attributable to the Equal-Rate Markov model of random branching (ERM; see Yule 1924; Cavalli-Sforza & Edwards 1967; Harding 1971); this is a null model of diversification in which all lineages have the same probability of speciation and extinction at the given time.

Though, uncertainty still exists about nature of this imbalance and extent to which artifacts influence the tree shapes (Mooers & Heard 1997). A number of studies suggest that estimated trees topologies are distorted with respect to various aspects of the phylogenetic tree construction. For instance, it was stated that topologies well corroborated in terms of consistency and retention indices (i.e. trees based on datasets including important hierarchical signal) are more balanced than trees based on poor data (Guyer & Slowinski 1991; Mooers et al. 1995; Salisbury 1999). It has also been debated whether various methods for estimating phylogenies could affect topology of the resulting phylogenetic trees (Heard 1992; Heard & Mooers 1996; Kirkpatrick & Slatkin 1996). Kirkpatrick & Slatkin (1996) ascertained that all studied construction methods (i.e. UPGMA, maximum parsimony, neighbour joining, maximum likelihood) were biased; if the real tree was symmetrical, the estimated tree was biased towards assymetry and vice versa.

Nevertheless, the available studies analyzing the possible methodological bias are usually based on limited numbers of small trees (Savolainen et al. 2002), and the results based on different indices of balance commonly disagree. For instance, Savage (1983) analysed a large collection of very small phylogenetic trees (having four to seven terminal taxa) and ascertained that the assembled trees were indistinguishable from those derived by the Equal-Rate Markov model of random branching. However, it is known that trees with less than eight tips could be, in practice, hardly distinguishable from ERM model (Rogers, 1994). In contrast, more recent studies of tree shape showed that tree topology is consistently more imbalanced than expected under the ERM model (Guyer & Slowinski 1991; Heard 1992; Guyer & Slowinski 1993; Mooers 1995). Similarly, it was argued that Colless' (1995) conclusion about difference of imbalance between cladistic and phenetic trees was distorted by high prevalence of homoplastic information in the analysed datasets (both random and empirical data matrices) (Heard & Mooers 1996). It was implied that when the data are essentially random, the cladistic and phenetic techniques produce different tree estimates, albeit these estimates were still "incorrect". Mooers et al. (1995) concluded that estimated trees based on poor data are more imbalanced, but Heard (1992) and Stam (2002) who analysed larger datasets of estimated trees using the same imbalance index as Mooers et al. (1995) stated that the correlation between tree balance and quality of data is virtually nonexistent. In the same vein, the largest study of tree imbalance (Herrada et al. 2008) was based on phylogenies rooted with outgroups, but it was stated that outgroup taxa might be a serious source of artifactual imbalance (Altaba 2009).

Even though it is computationally less demanding to obtain a topology (of an ultrametric tree) than an additive phylogenetic tree containing information about different branch lengths, more studies utilize index considering branch lengths data rather than topological data for the macroevolutionary analyses (see e.g. Pybus & Harvey 2000; Savolainen et al. 2002; Sims & McConvay 2002; Pie & Tschá 2009). In this thesis, topological index rather than a measure considering distribution of branch lengths was adopted, because some methods (including parsimony) do not consistently produce additive trees with explicit branch lengths (phylograms, chronograms), whereas a large number of tree topologies are readily accessible in the online database (Piel et al. 2001; Page 2007). More accurate methods of tree-topology analysis (such as indices independent of the size of a tree), careful evaluation of various possible artifacts, and inclusion of large number of taxa are needed to improve our understanding of macroevolutionary processes that could affect the real “shape” of phylogeny (Altaba 2009). Moreover, the exponential increase of published organismal phylogenies (Sanderson et al. 1993, Pagel 1997) allows accumulation of a large number of data; results based on a large dataset would amend our understanding of tree shape’s artifacts.

The aim of this thesis is to assess the amount of possible methodological and taxonomical bias in a large dataset of estimated phylogenetic trees, using the indices of imbalance evaluating each node individually (Fusco & Cronk 1995) as well as the frequently used whole-tree shape index (Colless 1982). The topological influence of variables such as type of data, construction methodology, and quality of data was evaluated. Moreover, differences between amounts of the imbalance between high-rank monophyletic taxa were assessed.

Materials and Methods

Indices of the tree imbalance

To measure tree imbalance, I used index proposed by Fusco & Cronk (1995). The Fusco & Cronk index was primarily designed for analyses of large trees including polytomies and is therefore well suited for large-scale studies of tree shape. As many estimated trees contain polytomies, application of the most commonly used measures such as Colless or Sackin index (Sackin 1972; Colless 1982) would be problematic, because these indices require fully resolved topologies (i.e. binary trees).

The Fusco & Cronk index is computed as follows: Let *S* denote size of a node (that is the number of tips that subtend from a given node). Then, consider a node with two descendant branches and measure the size of the bigger descendant branch (branch encompassing more tips, *B*) and minimum size of the bigger branch (*m*). *m* is computed as $m = \frac{S}{2}$ and it is rounded if *S* is odd. Let *M* be a maximum size of a bigger branch computed as *M* = *S* - 1. The imbalance of a node is then computed as:

$$I_{FC} = \frac{B - m}{M - m}$$

Essentialy, this index measures how much the bigger branch deviates from equality of branch sizes compared with possible range of values for the bigger branch. The measure is applied to all informative nodes, that is to nodes containing only two descendant branches (the index thus does not compute imbalance possibly hidden within the "multiple branching" of polytomic nodes). The size of an informative node has to be bigger than 3, because only one possible topology exists for tree size 3 (Suppl. Fig. 1). The index ranges from 0 (complete balance, i.e. symmetrical tree, Suppl. Fig. 2) to 1 (complete imbalance, i.e. comb-like tree). The index produces series of imbalance values computed for all informative nodes of the tree. It therefore represents overall imbalance of a given phylogeny more finely than indices computing a single value of imbalance. According to Fusco & Cronk (1995), all values of imbalance for given phylogenies were summarized in statistics of median (the measure of central tendency; MED), and quartile deviation (QD). Quartile deviation is sometimes called semi-interquartile range. It is computed as a half of the difference between first and third quartile and employed as a statistics of dispersion. QD ranges from 0 (homogenous values of imbalance in a tree) to 0.5 (heterogenous values of balance). MED and QD were chosen, because they represent skewed distribution of imbalance index better than arithmetic mean and variance. Essentially, median is a measure of total balance of a tree, whereas quartile deviation characterizes variability of nodal values of imbalance in a tree. QD thus shows how individual values of balance are distributed in a tree.

Since values of even nodes expected under null model are dependent on size of a node, correction has been applied (Purvis et al. 2002). Therefore, values have been weighted as follows:

$I' = I * \frac{S-1}{S}$ for even nodes,

I′ = *I* for odd nodes.

The modification enables statistical comparison between even and odd nodes. However, a limitation to this approach exists. For example, consider a completely imbalanced informative node of *S* = 4. The imbalance of the node is computed as *I* = (3-2) / (3-2) = 1. Because the node is even, it is weighted by ((*S*-1)/2) and its value is 0.75. Consequently, this approach hinders comparison between nodes of different *S*, because the maximum value of imbalance of small even nodes is smaller than 1. Because small nodes are more abundant in cladograms than the large ones, overall statistics would be ovewhelmed with imbalance values of small nodes. I have obviated this limitation by weighting every informative node by its size and summarizing these values in the new index: weighted average.

In order to thoroughly understand patterns of the tree balance, I have reanalysed all topologies without polytomies also with Colless index (Colless 1982), corrected by Rogers (1983) and Heard (1992). Colless index is computed as follows:

$$I_c = \frac{2}{(n-1)(n-2)}\sum_{i=1}^{n-1}(r_i - s_i)$$

The Colless index summarizes differences in size of descendant lineages (*r*, *s*) and divides it by largest attainable score. *n* denotes size of a tree. Colless index produces one value of balance for each phylogeny and gives more weight to larger nodes, i.e. nodes near the root. Colless index is very sensitive in detecting nonrandom variation of speciation and extinction rates (Kirkpatrick & Slatkin 1993). I have used normalized version of the index, so that the index gives value 0 for completely balanced tree and 1 for completely imbalanced tree.

Indices of Fusco & Cronk and Colless were calculated in Mesa 1.9.23.

The trees: download and data modifications

Topologies of the phylogenetic trees were downloaded from the TreeBASE database (www.treebase.org) (Piel et al. 2001) between December 2008 and July 2010. TreeBASE is currently the only extensive repository of published topologies, but it does not implement efficient searching algorithms (Page 2007). TreeBASE allows „tree surfing“, that is searching for topologies that share taxa, but this approach is both time-consuming and inefficient because it does not find all available topologies (Chen et al. 2008). Therefore, Phylofinder (Chen et al. 2008) and TreeBASE mirror (Brian O’Meara: http://www.brianomeara.info/treebase) were used to list the available trees. Phylofinder enabled sampling all trees that were descendants of a query taxon; for example, it would list topology of sunflowers, if all descendant trees of Viridiplantae were needed. To be selected, all trees had to meet following conditions:

1. The tree had to belong to one of the three organismal kingdoms: Viridiplantae, Animalia (Metazoa), and Fungi. This approach allows thorough sampling of major lineages of a Tree of life.
2. The tree’s operational taxonomic units had to be species (or, occasionally, subspecies), not supraspecific higher taxa. Recent studies caution against confounding specific and supraspecific higher taxa (Bininda-Emonds et al. 1998; Purvis & Agapow 2002). In a few

cases, trees with a small number of subspecies were allowed for the present analysis. If all subspecies of the nominal species constituted one monophyletic group, all but one of them were pruned. If pair of subspecies was not monophyletic, I did not prune them.

3. Information on the taxa that were used for the rooting of a tree had to be available. However, all outgroups were pruned from the analysis.
4. Information about tree construction method and type of data (e.g. morphological or behavioral) had to be available.
5. All trees had to be neontological. Paleontological trees tend to be more imbalanced than neontological trees because they are based on taxa from disparate time periods rather than on taxa from one time period (Harcourt-Brown et al. 2001) and because they often include species of the stem lineages, not of the crown groups exclusively (Panchen 1982).

Data characteristics

The utility and quality of data for each tree were recorded, including size of a tree (SO). I used consistency index (CI), retention index (RI), and index of consistency excluding uninformative characters (CIE) as indicators of the character fit (ratio of homoplasies / homologies and internal / terminal novelties in a tree). Tree robustness was evaluated as number of the most parsimonious trees (MPT) derived from a dataset. Numbers of parsimony-informative characters (PARSI) and number of variable characters (VAR) were recorded as well. Character-taxon ratio (C/T) is a ratio of columns / rows in a data matrix; when it is smaller than 1, then smaller number of characters than taxons exist in a data matrix (all characters and ingroup taxa were used). I used index of polytomy (PI) to assess the ratio of polytomic nodes in a phylogeny. PI was computed as follows:

PI = number of resolved internal nodes / (number of taxa - 1).

PI is dependent on size of a tree and was since primarily used to roughly evaluate number of polytomies and to sort data. All data were assorted to monophyletic groups (TAX.GROUP) and the superordinate "kingdoms" (KNG).

Various characteristics of measured data and algorithms used for reconstruction of phylogeny were employed: phylogenetic estimation method (METH), program used for reconstruction of a phylogeny (PROG), and type of data (DATA) were recorded.

Levels of qualitative variables

TAX.GROUP: Bryophyta, Polypodiopsida, Pinophyta, Magnoliophyta, Basidiomycota, Ascomycota, Glomeromycota, Zygomycota, Chytridiomycota, Lophotrochozoa, Ecdysozoa, and Deuterostomia.

KNG: Viridiplantae, Animalia (Metazoa), and Fungi.

METH: Maximum parsimony without successive weighting, Maximum likelihood, Neighbor joining, Bayesian analysis, Maximum parsimony with successive weighting, and Cluster analysis.

PROG: By hand analysis, Clados, Clustal X or W, DNAdist and NEIGHBOUR, Garli, Hennig86, MEGA, Mesquite, MrBayes, Nona, Paml, PAUP*, Phylip, Phyml, Poy, RAxML, TNT, Treefinder, and Xac.

DATA: Multilocus information (RFLP, RADP, AFLP), Nucleotides, Amino acids, Morphological characters, and Presence/absence data of chemical compounds (metabolites).

Analyses

To thoroughly evaluate all possible sources of bias, four clusters of variables have been assessed:

1. Taxonomic affiliation (TAX.GROUP, KNG)
2. Methodology (METH, PROG)
3. Type of data (DATA)
4. Quality of data (CI, CIE, RI, PARSI, VAR, MPT)

Variable PROG was not included in statistical analyses due to insufficient number of repetitions in respective factor levels (most phylogenies were analysed in PAUP* program) (Suppl. Table 1). Variables size of a tree, PARSI, VAR, MPT, and C/T ratio were log-transformed. Because of the highly disproportionate number of trees in the dataset in individual categories, only two categories of the TAX.GROUP variable were used in the analyses: Magnoliophyta and Ascomycota (177 and 77 phylogenies, respectively). From the same reason, only Viridiplantae and Fungi were used for statistical testing of the variable KNG (200 and 121 phylogenies, respectively). Likewise, I have utilised only trees constructed by Maximum parsimony and Maximum likelihood methods in analyses of METH variable (258 and 98 trees, respectively) and trees constructed from morphological and nucleotide matrices (57 and 303 phylogenies, respectively) (Suppl. Table 2-4). Bonferroni procedure was not applied due to substantial reduction of statistical test power and elevated probability of type II error (Nakagawa, 2004).

To test the relationship between quantitative variables, regression analyses were applied. Moreover, nonparametric test such as Kolmogorov-Smirnov statistics and Mann-Whitney U test were adopted because the assumption of the parametric test (the same number of observations in respective categories) was frequently violated. Due to combinatorial reasons, tree balance decreases with increasing number of taxa (Suppl. Fig. 1). Therefore, the effect of tree size needs to be controlled for. Colless index is heavily dependent on the tree size (Heard 1992); I have obviated this dependency by fitting values of Colless imbalance by the LOESS smoother (local regression) (Fig. 1). Values of bandwidth (alpha value) and degree of local polynomial were assessed with Akaike information criterion (alpha = 0.2, local polynomial = 2). Residual values were subsequently used in residual regression analysis. According to Fusco & Cronk (1995), patterns of imbalance in individual topologies were assessed for the three representative trees of Viridiplantae, Fungi, and Animalia (Suppl. Fig. 3-5).

Software

Calculations of Fusco & Cronk and Colless indices of tree imbalance were carried out using the software MeSA 1.9.23, written by Paul M. Agapow. Modifications and pruning of phylogenies was conducted in Mesquite 2.74. Calculations of indices of tree quality were carried out in PAUP 4.0. Statistical analyses were conducted in program Statistica 8.0 and R 2.10.1 (packages locfit and splom).

Results

Data matrix and variables

Altogether 413 phylogenetic trees (93 animal, 200 plant, and 121 fungal) were accumulated, containing 7,527 informative nodes. Sizes of phylogenetic trees varied from 4 to 420 terminal taxa. Distribution of the categories of qualitative variables (PROG, DATA, TAX.GROUP, and METH) among the analysed trees is summarized in Suppl. Tables 1-4. Overall distribution of indices of tree imbalance was non-normal, with exception of the Colless index (Table 1, Suppl. Fig. 6-8).

Table 1. Normality tests of distribution of tree imbalance. Shapiro-Wilk's W test and Kolmogorov-Smirnov tests were used.

Index of imbalance	Shapiro-Wilk's W test	Kolmogorov-Smirnov test	Distribution
Fusco & Cronk MED	W = 0.858, p < 0.001	d = 0.21, p < 0.01	highly non-normal
Fusco & Cronk QD	W = 0.989, p < 0.01	d = 0.046, p > 0.05	slightly non-normal
Weighted average	W = 0.945, p < 0.001	d = 0.069, p < 0.05	non-normal
Colless index	W = 0.99, p > 0.05	d = 0.052, p > 0.2	normal

Are imbalance indices dependent on size of a tree?

Overall, indices of tree imbalance are independent on size of a tree, with the exception of the quartile deviation. Regression analysis showed that dependence of Colless index on the size of a tree was successfully removed by fitting the values of Colless index with LOESS smoother (p = 0.9) (Fig. 1). Plotting values of Fusco & Cronk imbalance and size of a node (Fig. 2) showed that Purvis et al. (2002) adjustment eliminated dependence of the nodal imbalance on size of a node to a large extent. Nonetheless, regression analysis (Table 2.) showed that the dependence was not completely removed ($p < 10^{-5}$). The regression of MED with tree size was not significant ($R^2_{adj} = 0.001$, p = 0.47), but there was found a weak dependence of QD on tree size ($R^2_{adj}= 0.05$,

$p < 10^{-5}$, Fig. 3). The weighting of nodal values of Fusco & Cronk imbalance removed relationship with the size of a tree (p = 0.35). Therefore, values of Fusco & Cronk imbalance, weighted average, and Colless index can be used in following analyses as independent observations.

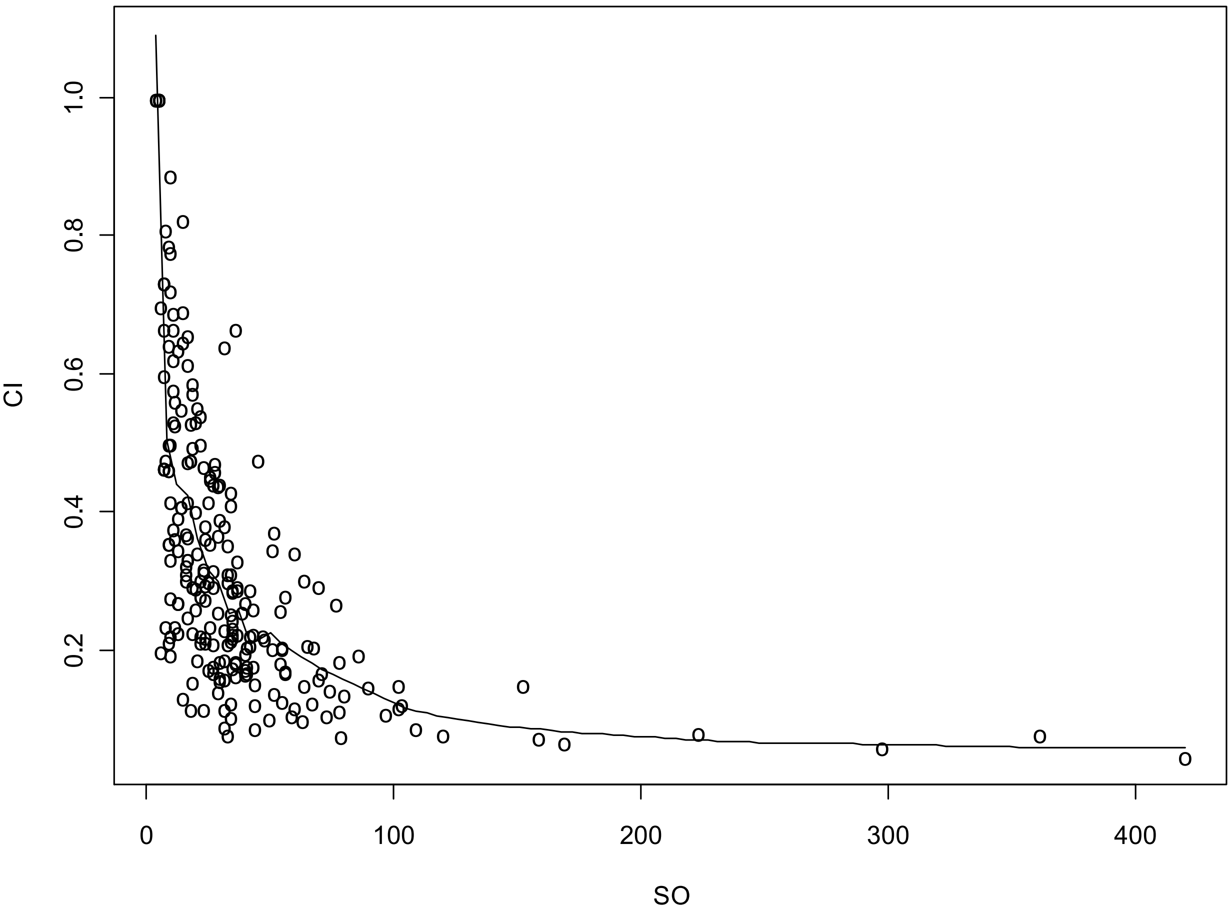


Fig. 1. Plotted relationship between values of Colless index (CI) and the size of a node (SO). LOESS smoother (local regression) was used for fitting the relationship.

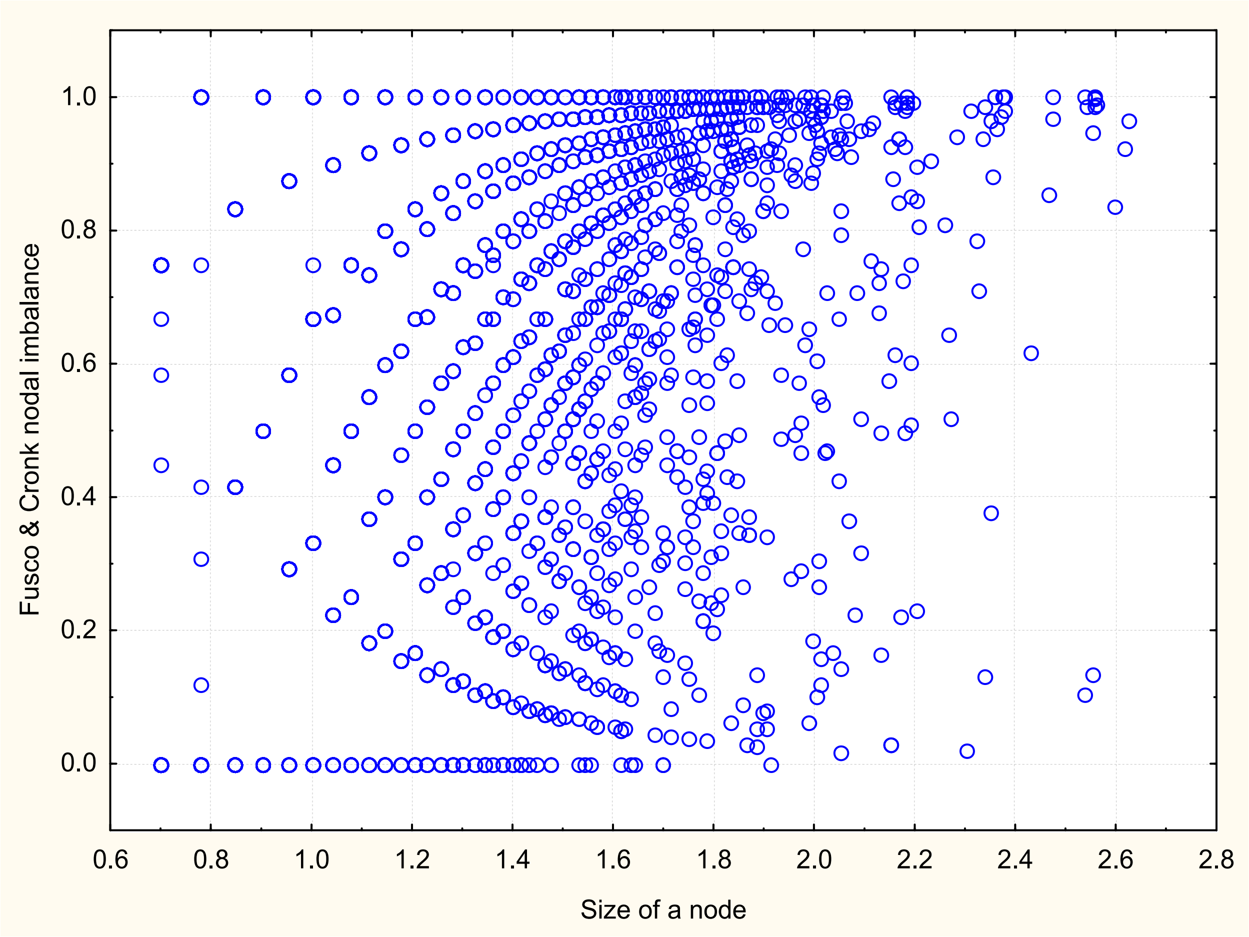


Fig. 2. Relationship between size of a node (log-transformed) and values of Fusco & Cronk nodal imbalance.

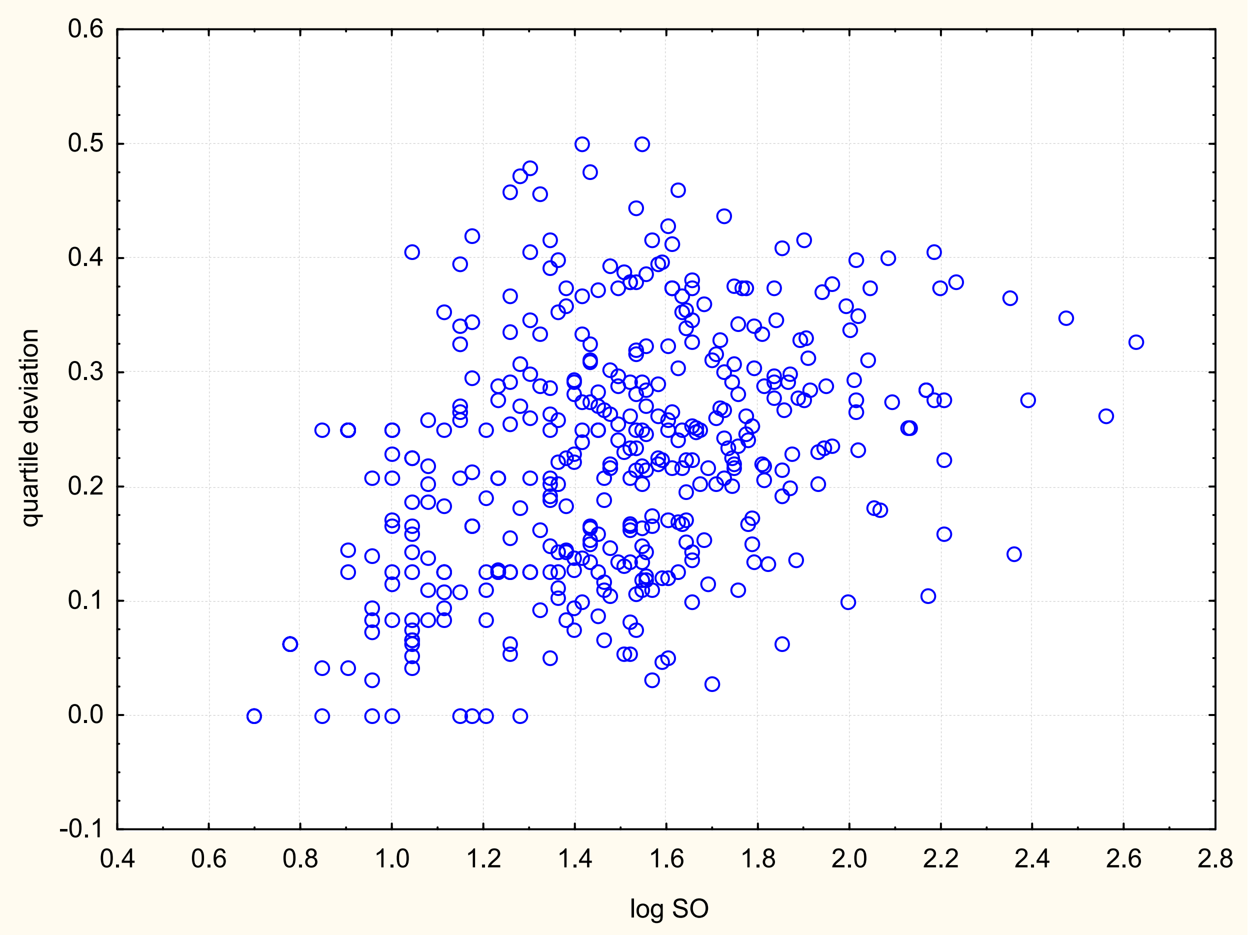


Fig. 3. Dependence of values of quartile deviation on size of a tree (SO, log-transformed).

Is tree balance dependent on phylogenetic quality of characters?

The median, weighted average, and Colless index show no significant relationship between tree balance and variables CI, CIE, RI, and VAR, but significant, albeit rather weak, relationship with C/T, PARSI, and MPT variables. Analysis of MPT (number of most parsimonious trees) with Fusco & Cronk median yielded a significant tendency for trees derived from data matrices producing more MPT to be more imbalanced (p = 0.01). Likewise, the more parsimony informative characters were present in the data matrix, the more symmetrical (measured by Fusco & Cronk median) were the tree topologies. On the contrary, regression analyses revealed strong relationship between quartile deviation and most of the independent variables except to MPT and C/T variables. It was found that when quartile deviation is low (approximately 0.0-0.15), nodal values tend to be rather imbalanced (and, naturally, less dispersed), when QD is high (approximately 0.35–0.5), nodal values are highly dispersed (analyses not shown). Therefore, the appearance of trees with low QD is homogenous (nodes tend to be imbalanced), whereas trees with high QD tend to contain both highly imbalanced and balanced nodes. Results of the analyses performed show tendency for individual trees having disparate values of tree balance (high QD) when they are based on poor data. Additionally, the variable CI explained most variance in values of quartile deviation (Fig. 4), when measured by F-test.

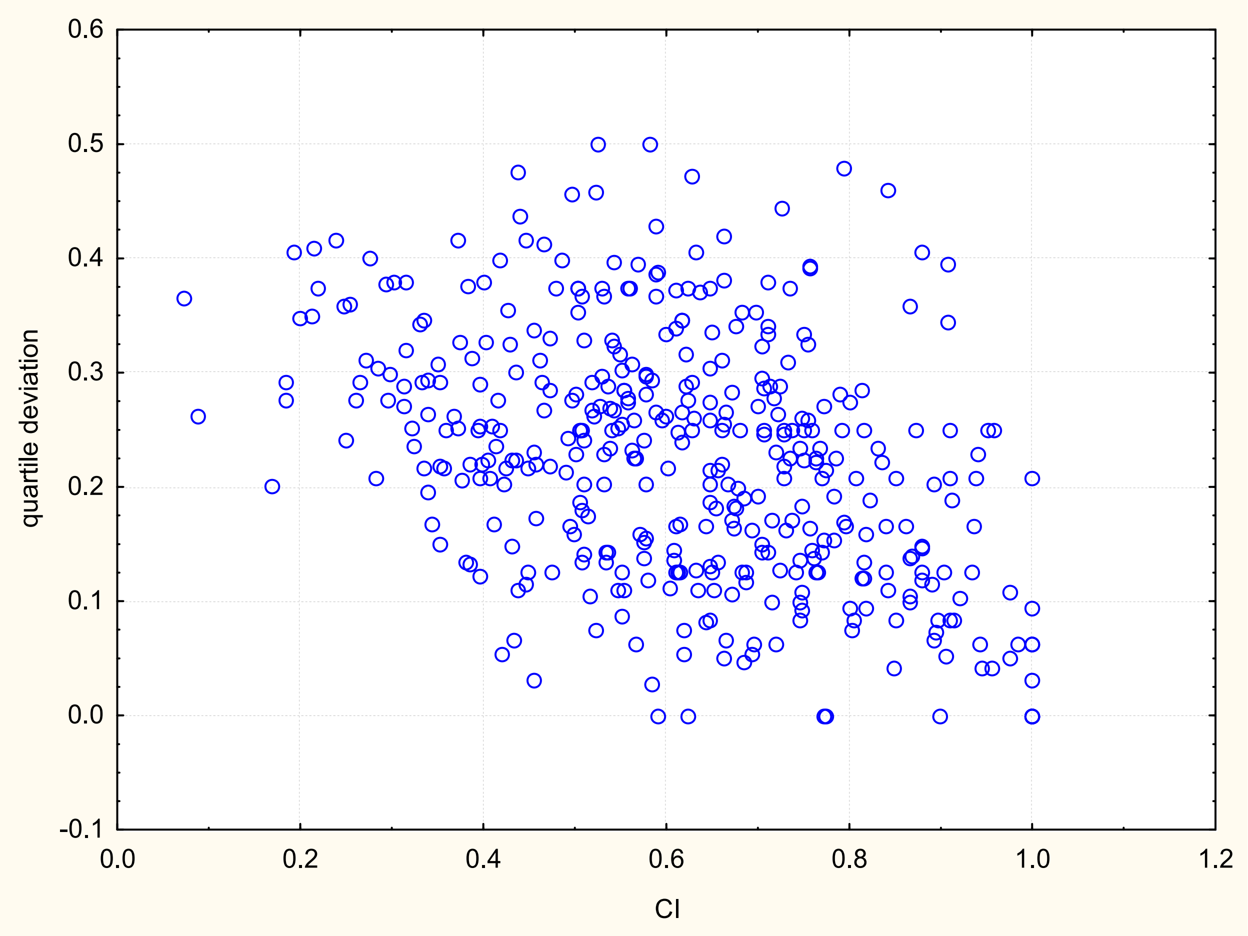


Fig. 4. Relationship between variables of quartile deviation and consistency index (CI).

Table 2: Results of regression and non-parametric analyses. MUW = Mann Whitney U test, KW = Kruskal-Wallis test.

Independent variables	Fusco & Cronk MED	Fusco & Cronk QD	Weighted average
TAX.GROUP	*MWU, $U = 5434$, $n_1 = 176$, $n_2 = 77$, $p = 0.01$*	*MWU, $U = 4959$, $n_1 = 176$, $n_2 = 77$, $p = 0.0006$*	*MWU, $U = 5650$, $n_1 = 176$, $n_2 = 77$, $p = 0.03$*
KNG	$p = 0.08$	*KW, $H = 10,36$, $p = 0.005$*	$p = 0.24$
METH	$p = 0.41$	*MWU, $U = 9596$, $n_1 = 259$, $n_2 = 98$, $p = 0.0003$*	$p = 0.91$
DATA	$p = 0.96$	*MWU, $U = 6771$, $n_1 = 57$, $n_2 = 303$, $p = 0.009$*	$p = 0.51$
RI	$p = 0.08$	*$p < 0.000001, R^2_{adj} = 0.05$, $F_{(1,411)} = 24.13$*	$p = 0.56$
CI	$p = 0.32$	*$p << 0.000001, R^2_{adj} = 0.16$ $F_{(1,411)} = 83.58$*	$p = 0.78$
CIE	$p = 0.51$	*$p << 0.000001, R^2_{adj} = 0.16$, $F_{(1,411)} = 82.68$*	$p = 0.75$
PARSI	$p = 0.02$, $R^2_{adj} = 0.14$, $F_{(1,411)} = 68.81$	*$p << 0.000001, R^2_{adj} = 0.14$, $F_{(1,411)} = 68.81$*	$p = 0.28$
VAR	$p = 0.07$	*$p << 0.000001, R^2_{adj} = 0.11$, $F_{(1,411)} = 53.47$*	$p = 0.27$
MPT	*$p = 0.01$, $R^2_{adj} = 0.03$, $F_{(1,158)} = 6.73$*	$p = 0.74$	$p = 0.50$
C/T	*$p = 0.01$, $R^2_{adj} = 0.01$, $F_{(1,411)} = 6.22$*	$p = 0.08$	*$p = 0.04, R^2_{adj} = 0.007$, $F_{(1,411)} = 3.96$*
SO	$p = 0.47$	*$p < 0.000001$, $R^2_{adj} = 0.05$, $F_{(1,411)} = 25.87$*	$p = 0.35$

Does tree balance differ between taxonomic groups?

All the used indices of tree balance show that there is a significant difference between the imbalance of tree topologies in the Magnoliophyta and Ascomycota. Mann Whitney U test (median: $p = 0.01$, quartile deviation: $p = 0.0006$, weighted average: $p = 0.03$) and Kolmogorov-Smirnov test (Colless index: $p > 0.05$) were used for statistical testing. Moreover, significant difference between Animalia and Fungi was found with the Fusco & Cronk quartile deviation ($p = 0.005$).

Is tree balance dependent of methodology and type of data?

Nonparametric analyses revealed that there is no significant relationship between indices of Colless, Fusco & Cronk median, and weighted average on one side and variables of METH and DATA on the other. However, Mann Whitney U test revealed strong relationship between quartile deviation and METH (p = 0.0003) and DATA (p = 0.009) variables.

Discussion

It is assumed that symmetry of phylogenetic trees could convey information about fluctuation of speciation and extinction rates (Heard 1996; Heath et al. 2008). Nonethess, only a very limited number of studies employed topological information to study macroevolutionary processes (see e.g. Guyer & Slowinski 1993; Katzourakis et al. 2001; Phillimore et al. 2006). These studies analysed variation of speciation and extinction rates and suggested that tested trees were more imbalanced than it would be attributable to Equal-Rate Markov model of random branching (ERM). Katzourakis et al. (2001) and Phillimore et al. (2006) used a large number of ecological, morphological, and life-history correlates, such as body size, diet-breadth, life history or mating system. While Katzourakis et al. (2001) found only a small number of variables, such as sexual selection and diet-breadth, associated with the tree imbalance, Phillimore at al. (2006) explained a large percentage of variance in the tree imbalance by studied ecological variables, e.g. adult dispersal and feeding generalization.

The dataset of the trees was analysed with several indices describing the tree shape. Utilization of distinct measures of imbalance is advantageous, because different indices are not equally sensitive to nonrandom patterns of cladogenesis in different parts of a tree (Kirkpatrick & Slatkin 1993). The Colless index is a frequently used index of the tree balance (Mooers 1995; Mooers et al. 1995; Heard 1996; Salisbury 1996; Heard & Mooers 2002) and was utilized here only as an additional measure, because it does not compute with trees including polytomies. Therefore, only a limited number of phylogenies (approximately half) could be measured with the Colless index. Median and weighted average indices represent global imbalance of the trees, and it is then feasible to compare results of analyses based on these indices with other studies of the tree imbalance (Katzourakis et al. 2001; Agapow & Purvis 2002; Purvis & Agapow 2002). On the other hand, quartile deviation represents a new aspect of the tree shape: distribution of imbalance values across a tree.

So far, the methodological bias was studied either by simulation approach (Sepkoski & Kendrick 1993; Losos & Adler 1995; Huelsenbeck & Kirkpatrick 1996; Agapow & Purvis 2002) or by mining the literature for estimated trees (Guyer & Slowinski 1991; Heard 1992; Harcourt-Brown et al. 2001; Stam 2002). In this thesis, a large quantity of data was analysed to identify possible sources of bias (413 collected trees including up to 420 taxa were accumulated, with 7527 analysed nodes in total). Compared to other studies based on the estimated tree topologies (Savage 1983; Guyer & Slowinski 1991; Mooers 1995), a relatively large number of large topologies was sampled. For example, Savage (1983) collected more than 1,000 incomplete topologies (that is phylogenetic trees that do not contain all the members of a given clade) of sizes 4-7, Guyer & Slowinski (1991) sampled 120 complete trees (that is trees containing all the extant taxa of a given clade) of five tips, and Mooers (1995) accumulated 39 complete trees with 8-14 tips. When the tree size is small (approx. 4–7 tips), all real topologies (including the most extreme) typically tend to be indistinguishable from the ERM model (Rogers 1994). In other words, it is not possible to discern whether complete imbalance of a tree of five taxa suggests that individual lineages had unequal net speciation rates (i.e. speciation minus extinction rates), or the tree shape originated randomly. There is only a limited number of small trees (19 trees smaller than 8) in this data set, and the tree size of a 30 topologies was greater than 100 terminal taxa. Therefore, it is possible to test hypotheses about methodological bias with the accumulated data.

We only consider effect of random / nonrandom omission of taxa from a tree; basically, because extinction process decreases number of extant taxa, all phylogenetic trees are incomplete. The accumulated trees are incomplete, that is they do not contain all the known recent species of a given clade. The impact of tree incompleteness on the tree shape was analysed by several authors (Guyer & Slowinski 1989; Heard 1992; Mooers et al. 1995; Heath et al. 2005). Altogether, these studies illustrate that in estimated trees, incompleteness increases tree imbalance, but, importantly, random pruning of branches in simulated trees does not change the tree shape systematically (Guyer & Slowinski 1989). In addition, also the complete trees were found to be more imbalanced than predicted by the null model based on random speciation and extinction (ERM) (Mooers 1995).

There are other conceivable sources of bias, for example anthropocentric bias (O'Hara 1992; Sandvik 2007). O'Hara (1992) suggested that the estimated trees may be distorted by favouring the position of a *Homo* species, or larger taxa including our species, such as Primates or Vertebrata. Sandvik (2007) found a weak tendency for published cladograms in phylogenetic texbooks to be antropocentrically biased. In the same vein, Zink (2004) suggested that a discrepancy exists between a real number of tips in the trees and the number of tips considered by taxonomists in phylogenies. The study (Zink 2004) suggested that 97% of avian subspecies lacked a genetic structure typical for distinct evolutionary unit. Furthermore, there probably is a bias caused by systematic errors in recognition of very large and very small taxa by systematics (Scotland & Sanderson 2003). Scotland & Sanderson (2003) compared several large phylogenies (e.g. Asteraceae or Aves) with null models of frequency distribution of units containing subunits (so called "hollow curve distribution") and found that real trees contained smaller number of small and large genera than null models.

One of the possible sources of bias is taxonomic bias. All the used indices of tree balance showed that there is no significant difference in the balance of the tree topologies except for the Magnoliophyta versus Ascomycota. Also in all previous studies, the effect of taxonomic affiliation was found insignificant (Guyer & Slowinski 1991; Heard 1992; Mooers et al. 1995; Stam 2002). It is conceivable that the differences in the balance of trees between taxonomic groups have not been previously revealed perhaps due to a low numbers of trees included in the above cited studies. For example, Stam (2002) compared 39 animal and 19 plants phylogenies, and Guyer & Slowinski (1991) sampled 25 cladograms for each of the three monophyletic lineages of Insecta, Tetrapoda, and Magnoliophyta. In this thesis, 177 studies of Magnoliophyta were compared to 77 trees of Ascomycota; other monophyletic lineages (e.g. Deuterostomia, Pinophyta or Bryophyta) were not tested because of small number of trees in these cathegories. Median values of tree sizes were 28.5 and 35 for Magnoliophyta and Ascomycota, respectively, hence the effect of the different tree sizes can be ruled out. Trees of both taxonomic groups were constructed preferably by the maximum parsimony. Therefore, the observed differences might be an outcome of the differences in the net speciation rate between these two groups (although it is possible that the taxonomic differences are artifactual, i.e. caused by different treatment of taxa by fungal and botanical taxonomists).

Besides the taxonomic bias and influence of sample incompleteness on the tree shape, methodological bias is hypothesised to influence tree topologies. Results of performed analyses suggest that the methodological bias (i. e. quality of data, contruction algorithms, type of data) affect tree topologies only to a small extent. A large number of studies analyzing influence of methodological bias on tree shape exists (e.g. Guyer & Slowinski 1990; Kirkpatrick & Slatkin 1993; Mooers 1995; Mooers et al. 1995; Huelsenbeck & Kirkpatrick 1996; Salisbury 1999; Stam 2002). Generally, simulation studies (Huelsenbeck & Kirkpatrick 1996; Salisbury 1999) suggest that relationship between tree shape and the applied methodologies exist, while studies based on estimated trees produce inconsistent results (Mooers 1995; Mooers et al. 1995; Stam 2002). For example, Mooers (1995) and Mooers et al. (1995) suggest that the less corroborated the trees are the more imbalanced is their topology. When there is a preponderance of random positions in a data set, the estimated trees are essentially random, and imbalanced trees are fairly probable when randomly generated (Slowinski 1990). Moreover, it was documented that positive correlation between indices of data quality and tree shape was caused by utilization of index dependent on the tree shape: Stam (2002), who employed index of balance independent of size of a tree, found no relationship between retention and consistency indices and the tree balance. It was demonstrated, both analytically and by a simulation approach (Heard 1992), that tree shape is heavily influenced by the tree size and that tree size explains most variance in values of tree imbalance. From combinatorial reasons, ratio of balanced / imbalanced topologies increases with growing number of taxa. This correlation has an important implication: studies of the tree shape must be controlled for the influence of the tree size. Therefore, it is possible that analyses of quartile deviation were biased by its dependency on the size of a tree. Altogether, it is possible that falsely positive results in some studies of methodological bias were caused by utilization of inappropriate indices of balance.

In sum, our results demonstrate that the tree shape is influenced by the methodological bias only to a small extent. It is therefore possible to employ tree shape in studies of macroevolution. Nonetheless, it is necessary to use index of imbalance independent of size of a tree in order to obtain unbiased results. The differences between number of explained variability in imbalance values between cited macroevolutionary studies (Katzourakis et al. 2001; Phillimore et al. 2006) would be hence explained by the application of index of balance dependent on a tree size (Katzourakis et al. 2001, Purvis et al. 2002), which would bias results of this macroevolutionary study.

References

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. 16: 23-34.

Altaba, C. R. 2009. Universal artifacts affect the branching of phylogenetic trees, not universal scaling laws. PloS ONE 4: e4611.

Agapow, P. M., Purvis, A. 2002. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. Syst. Biol. 51: 866-872.

Binder, M., Hibbett, D. S., Molitoris, H. P. 2001. Phylogenetic relationships of the marine gasteromycete Nia vibrissa. Mycologia 93: 679-688.

Bininda-Emonds, O. R. P., Bryant, H. N., Russell, A. P. 1998. Supraspecific taxa as terminals in cladistic analysis: implicit assumptions of monophyly and a comparison of methods. Biol. J. Linn. Soc. 64: 101-133.

Brady, S. G., Schultz, T. R., Fisher, B. L., Ward, P. S. 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. Proc. Natl. Acad. Sci. USA 103: 18172-18177.

Cavalli-Sforza, L. L., Edwards, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21: 550-570.

Colless, D. H. 1982. Phylogenetics: the theory and practise of phylogenetic systematics II (book review). Syst. Zool. 31: 100-104.

Colless, D. H. 1995. Relative symmetry of cladograms and phenograms: an experimental study. Syst. Biol. 44: 102-108.

Fusco, F., Cronk, Q. C. B. 1995. A New Method for Evaluating the Shape of Large Phylogenies. J. Theor. Biol. 175: 235-243.

Guyer, C., Slowinski, J. B. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null mode. Am. Nat. 134: 907-921.

Guyer, C., Slowinski, J. B. 1991. Comparison of observed phylogenetic topologies with null expectations among three monophyletic lineages. Evolution 45: 340-350.

Guyer, C., Slowinski, J. B. 1993. Adaptive radiation and the topology of large phylogenies. Evolution 47: 253-263.

Harcourt-Brown, K. G., Pearson, P. N., Wilkinson, M. 2001. The imbalance of paleontological trees. Paleontology 27: 188-204.

Harding, E. F. 1971. The probabilitie of rooted tree-shapes generated by random bifurcation. Adv. Appl. Probab. 3: 44-77.

Heard, S. B. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution 46: 1818-1826.

Heard, S. B. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution 50: 2141-2148.

Heard, S. B., Mooers, A. 1996. Imperfect information and the balance of cladograms and phenograms. Syst. Biol. 45: 115-118.

Heath, T. A., Zwickl, D. J., Kim, J., Hillis, D. M. 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst. Biol. 51: 160-166.

Herrada, E. A., Tessone, C. J., Klemm, K., Eguíluz, V. M., Hernández-García, E., Duarte, C. M. 2008. Universal scaling in the branching of the Tree of life. Plos ONE 3: e2757.

Chen, D., Burleigh, J. G., Bansal, B. S., Fernandez-Baca, D. 2008. Phylofinder: an intelligent search engine for phylogenetic tree databases. BMC Evol. Biol. 8: 90.

Katzourakis, A., Purvis, A., Azmeh, S., Rotheray, G., Gilbert, F. 2001. Macroevolution of hoverflies (Diptera: Syrphidae): the effect of using higher-level taxa in studies of biodiversity, and correlates of species richness. J. Evol. Biol. 14: 219-227.

Kirkpatrick, M., Slatkin, M. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47: 1171-1181.

Losos, J. B., Adler, F. R. 1995. Stumped by trees? A generalized null model for patterns of organismal diversity. Am. Nat. 145: 329-342.

Mooers, A. 1995. Tree balance and tree completeness. Evolution 49: 379-384.

Mooers, A., Page, R. D. M., Purvis, A., Harvey, P. H. 1995. Phylogenetic noise leads to unbalanced cladistic tree reconstructions. Syst. Biol. 44: 332-342.

Mooers, A., Heard, S. B. 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72: 31-54.

Nakagawa, S. 2004. Farewell to Bonferoni: The problems of low statistical power and publication bias. Behav. Ecol. 15: 1044-1045.

O'Hara, R. J. 1992. Telling the tree: Narrative representation and the study of evolutionary history. Biol. Philos. 7: 135-160.

Page, R. D. M. 2007. Tbmap: a taxonomic perspective on phylogenetic database Treebase. BMC Bioinform. 8: e158.

Pagel, M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr. 26: 331-348.

Panchen, A. L. 1982. The use of parsimony in testing phylogenetic hypotheses. Zool. J. Linn. Soc. 74: 305-328.

Phillimore, A. B., Freckleton, R. P., Orme, C. D. L., Owens, I. P. F. 2006. Ecology predicts large-scale patterns of phylogenetic diversification in birds. Am. Nat. 168: 220-229.

Pie, M. R., Tscha, M. K. 2009. The macroevolutionary dynamics of ant diversification. Evolution 63: 3023-3030.

Piel, W. H., Donoghue, M. J., Sanderson, M. J. 2001. TreeBASE: a Database of Phylogenetic Information. Information, 2nd International Workshop of Species 2000. Tsukuba, Japan: National Institute for Environmental Studies.

Purvis, A., Agapow, P. M. 2002. Phylogeny imbalance: taxonomic level matters. Syst. Biol. 51: 844-854.

Purvis, A., Katzouarakis, A., Agapow, P. M. 2002. Evaluating phylogenetic tree shape: two modifications to Fusco & Cronk‘s method. J. Theor. Biol. 214: 99-103.

Pybus, O. G., Harvey, P. H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. B 267: 2267-2272.

Rogers, J. S. 1993. Responses of Colless‘ tree imbalance to number of terminal taxa. Syst. Biol. 42: 102-105.

Rogers, J. S. 1994. Central moments and probability distribution of Colless's coeficient of tree imbalance. Evolution 48: 2026 – 2036.

Ronsted, N., Weiblen, G. D., Cook, J. M., Salamin, N., Machado, C. A., Savolainen, V. 2005. 60 million years of co-divergence in the fig-wasp symbiosis. Proc. R. Soc. Lond. B 22: 2593-2599.

Sanderson, M. J., Baldwin, B. G., Bharathan, G., Campbell, C. S., von Dohlen, C., Ferguson, D., Porter, J. M., Wojciechowski, M. F., Donaghue, M. J. 1993. The growth of phylogenetic information and the need for a phylogenetic data base. Syst. Biol. 42: 562-568.

Sandvik, H. 2007. Anthropocentrism in cladograms. Biol. Philos. 24: 425-440.

Savolainen, V., Heard, S. B., Powell, M. P., Davies, T. J., Mooers, A. 2002. Is cladogenesis heritable? Syst. Biol. 51: 835-843.

Scotland, R. W., Sanderon, M. J. 2004. The significance of few versus many in the tree of life. Science 303: 643.

Sepkoski, J. J., Kendrick, D. C. 1993. Numerical experiments with model monophyletic and paraphyletic taxa. Paleobiology 19: 168-184.

Sims, H. J., McConway, K. J. 2003. Nonstochastic variation of species-level diversification rates within angiosperms. Evolution 57: 460–479.

Sackin, M. J. 1972. "Good" and "bad" phenograms. Syst. Zool. 21: 225-226.

Salisbury, B. A. 1999. Misinformative characters and phylogeny shape. Syst. Biol. 48: 153-169.

Slowinski, J. B. 1990. Probabilities of n-trees under two models: a demontration that assymetrical interior nodes are not improbable. Syst. Biol. 39: 89-94.

Stam, E. 2002. Does imbalance in phylogenies reflect only bias? Evolution 56: 1292-1295.

Vrba, E. S., Eldredge, N. 1984. Individuals, hierarchies and processes: Towards a more complete evolutionary theory. Paleobiology 10: 146-171.

Yule, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. F. R. S. Phil. Trans. R. Soc. Lond. A 213: 21-87.

Zink, R. M., 2004. The role of subspecies in obscuring avian biological diversity and misleading conservation policy. Proc. R. Soc. Lond. B 271: 561–564.

Online references

TreeBASE 2. 0.

http://www.treebase.org/treebase-web/home.html;jsessionid=C80863188E70814051135A2376EA0E63

Phylofinder (the website is no longer functioning)

http://pilin.cs.iastate.edu/phylofinder/

TreeBASE mirror

http://www.brianomeara.info/treebase

Supplementary

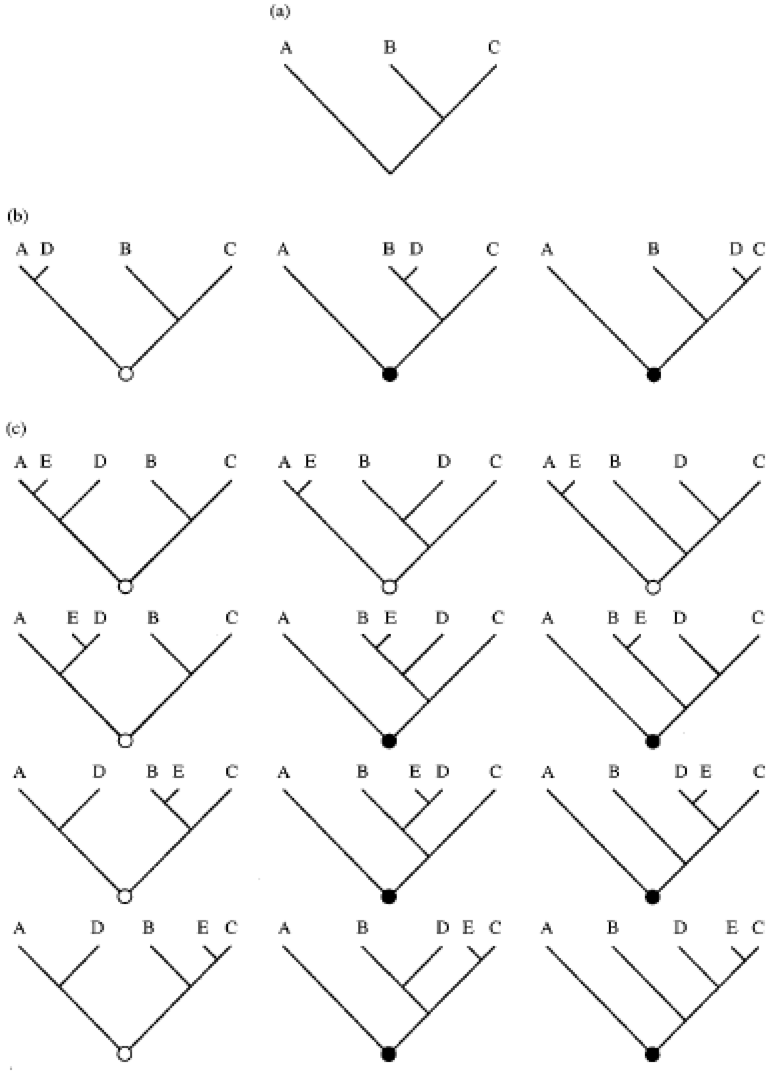


Fig. 1. The possible topologies for tree sizes of three (a), four (b), and five (c). The figure depicts only unlabeled trees, i.e. tree topologies. There are three possible topologies for tree size of four, two of them are comb-like, and one is perfectly symmetrical. Black circles at the root nodes indicate that the entire topology is imbalanced, white circles indicate that the entire topology is balanced. Twelve possible topologies exist for tree size of four; the ratio of imbalanced and balanced phylogenies decreased to 6:6 (Purvis et al. 2002).

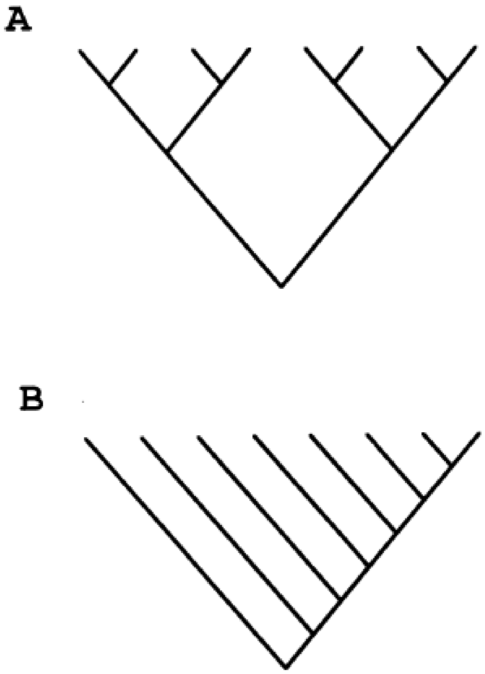


Fig. 2. Completely balanced (A) and imbalanced (B) topologies for tree size 8 (Heard 1992).

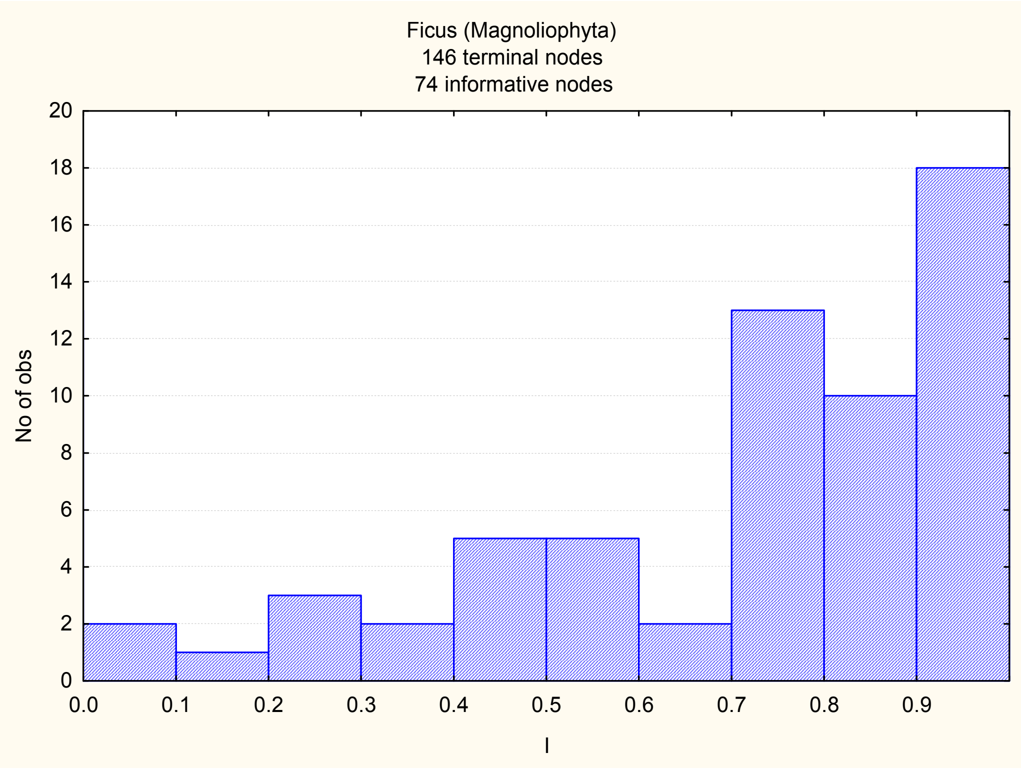


Fig. 3. Frequency distribution of Fusco & Cronk nodal values of imbalance for the tree of *Ficus* (Magnoliophyta). Median for this study is 0.75, quartile deviation is 0.28. (Ronsted et al. 2005).

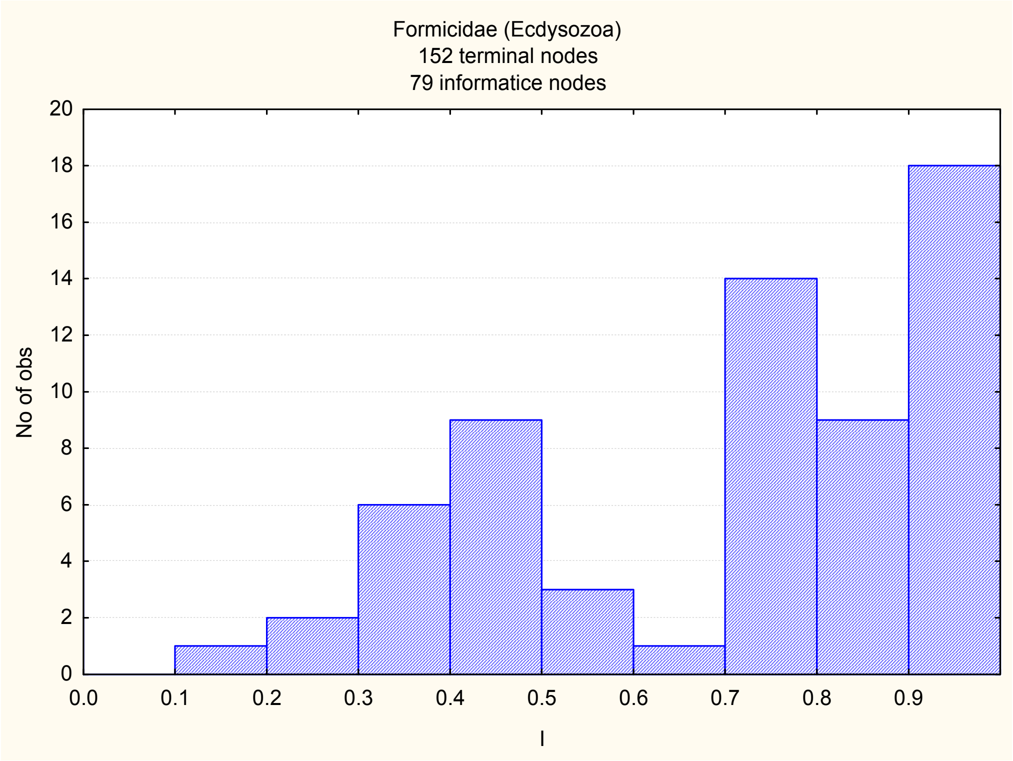


Fig. 4. Frequency distribution of Fusco & Cronk nodal values of imbalance for the tree of Formicidae (Ecdysozoa). Median for this study is 0.62, quartile deviation is 0.31 (Brady et al. 2006).

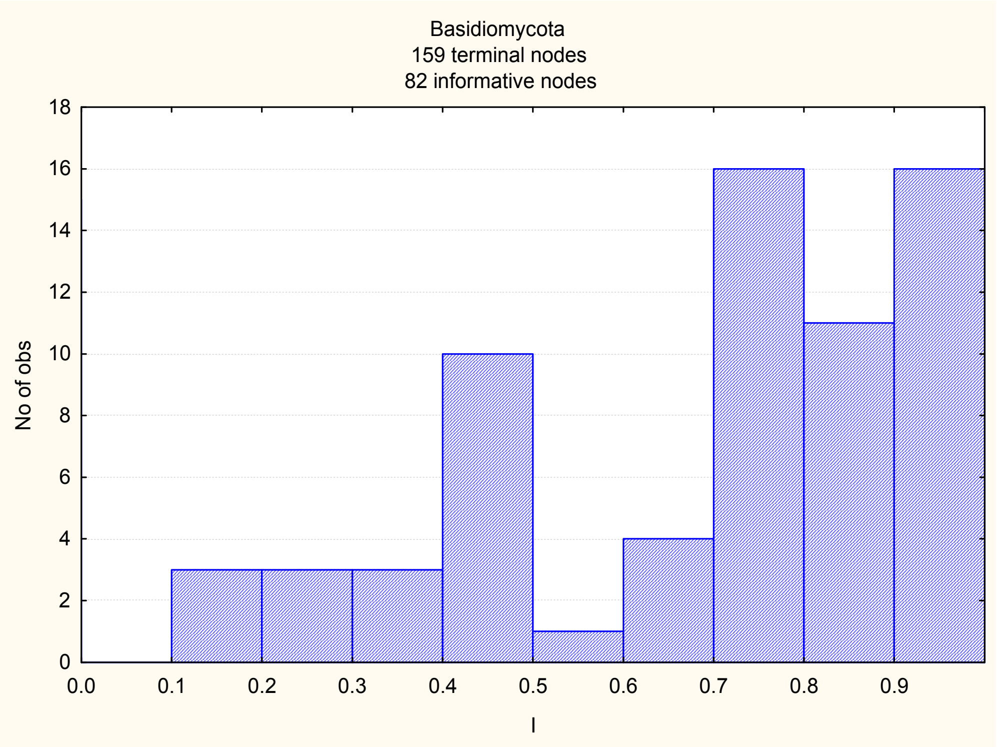


Fig. 5. Frequency distribution of Fusco & Cronk nodal values of imbalance for the tree of Gasteromycetes (Basidiomycota). Median for this study is 0.75, quartile deviation is 0.27 (Binder et al. 2001).

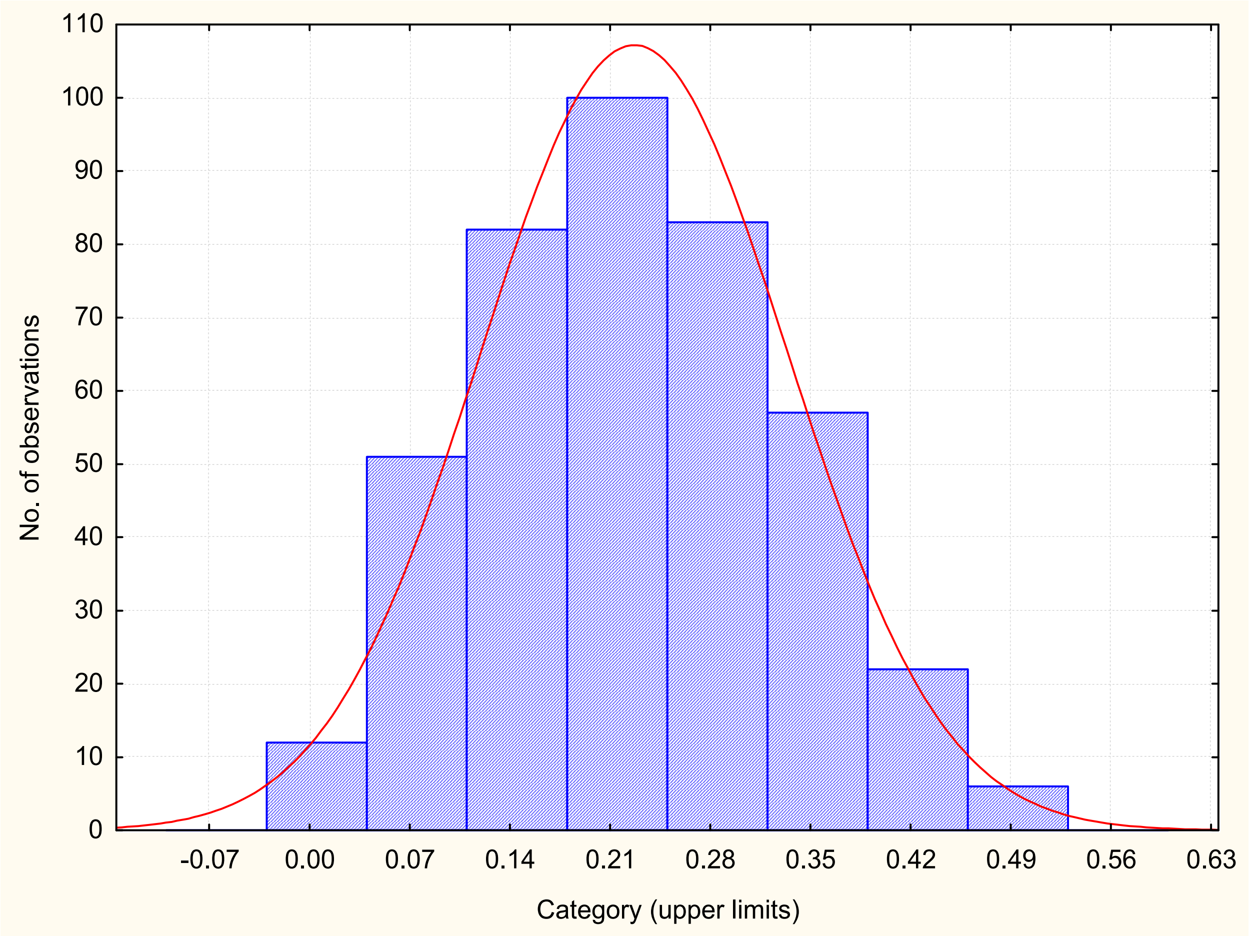


Fig. 6. Frequency distribution of Fusco & Cronk statistics quartile deviation.

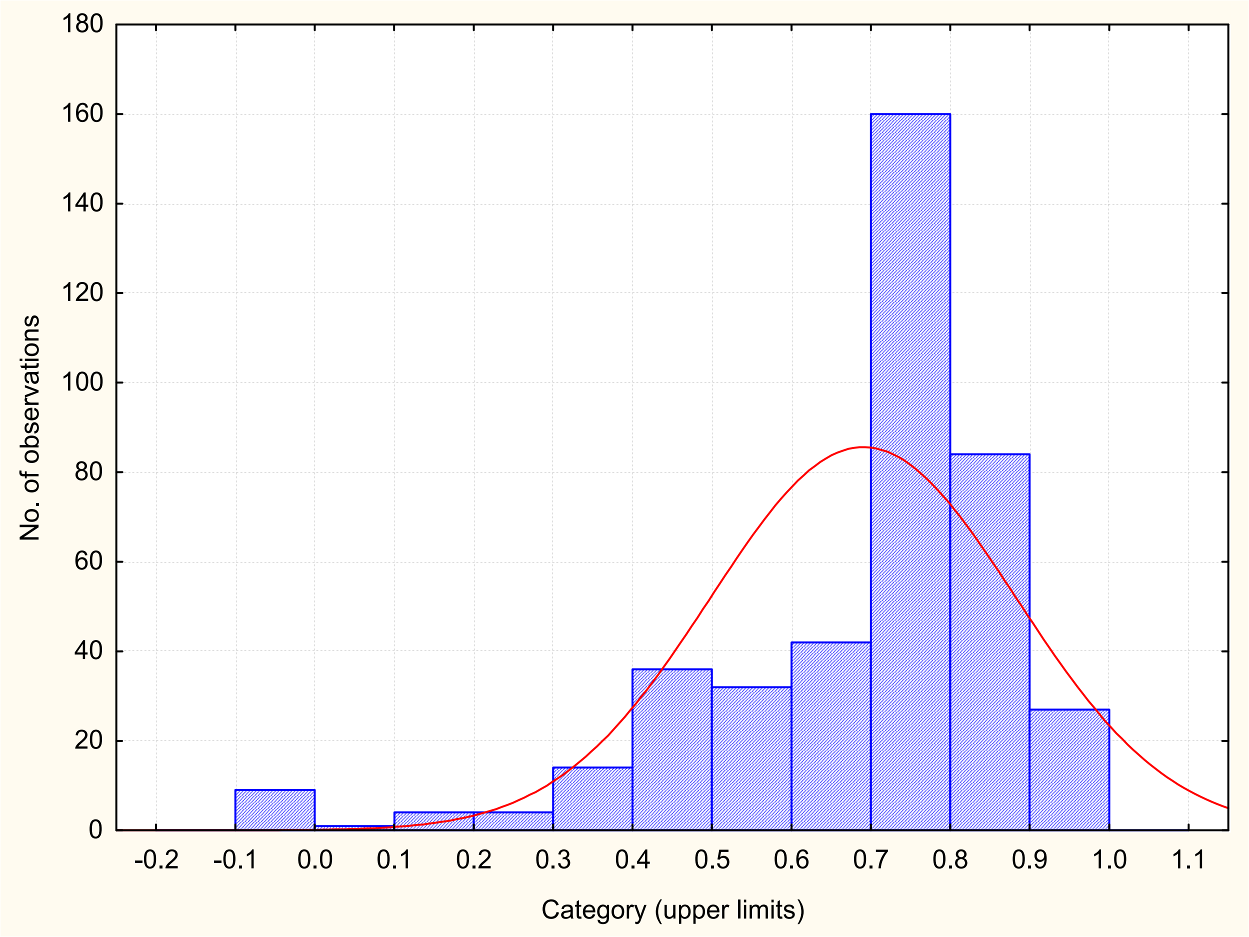


Fig. 7. Frequency distribution of Fusco & Cronk statistics median.

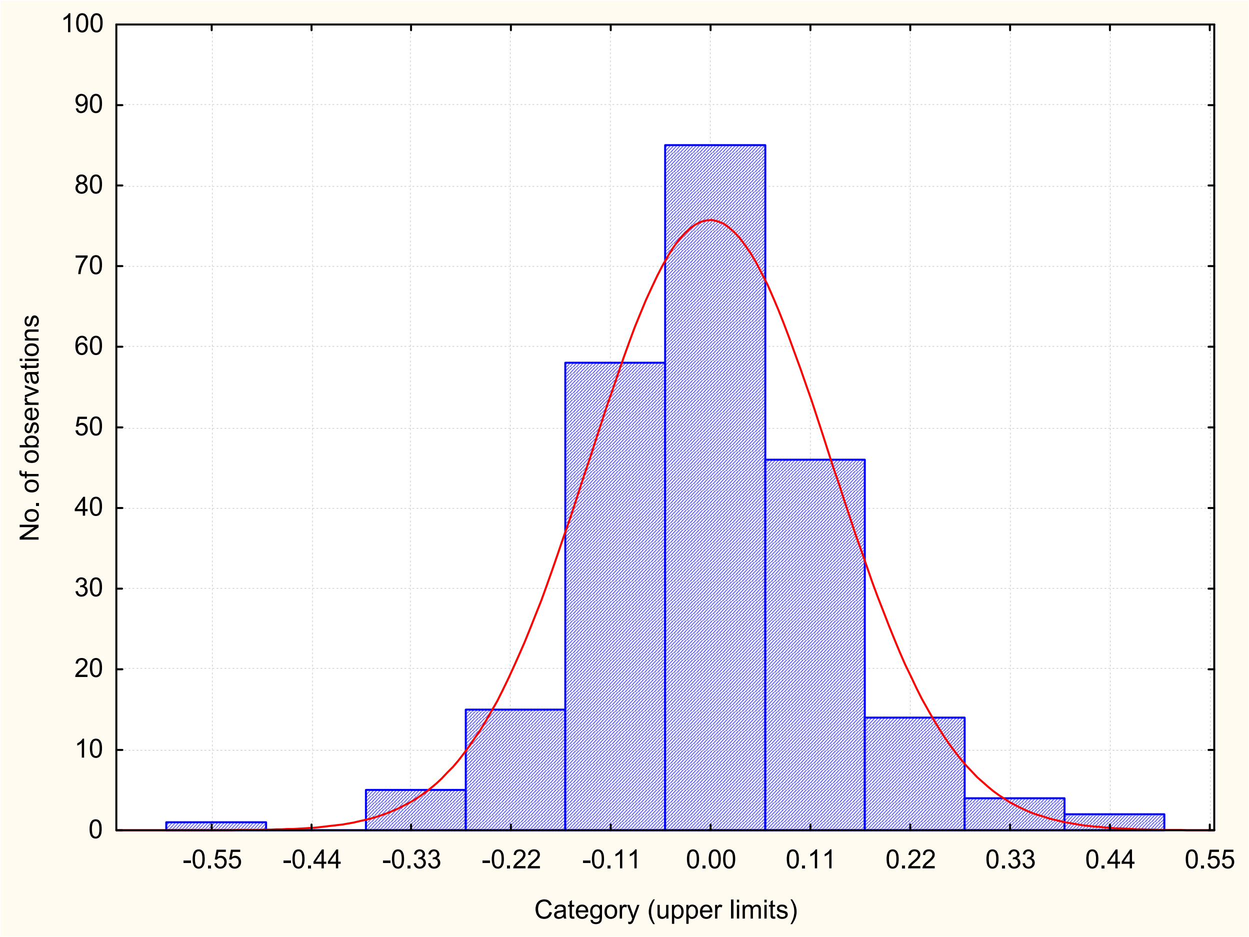


Fig. 8. Frequency distribution of residuals of Colless index.

Table 1.

PROG	**Number of phylogenies**
By hand analysis	8
Clados	1
Clustal X, W	2
DNAdist + Neighbour	2
Garli	9
Hennig86	16
MEGA	1
Mesquite	1
MrBayes	25
Nona	11
Paml	1
PAUP	302
Phylip	4
Phyml	3
Poy	1
RAxML	11
TNT	5
Treefinder	3
Xac	1

Table 2.

DATA	Number of phylogenies
Amino acids	4
Morphological characters	57
Multilocus information (RFLP, RADP, AFLP)	20
Nucleotides	303
Presence/absence data of chemical compounds (metabolites)	2

Table 3.

TAX.GROUP	Number of phylogenies
Ascomycota	77
Basidiomycota	38
Bryophyta	8
Deuterostomia	47
Ecdysozoa	30
Glomeromycota	1
Lophotrochozoa	16
Magnoliophyta	177
Pinophyta	4
Polypodiopsida	10
Zygomycota	6

Table 4.

METH	Number of phylogenies
Bayesian analysis	26
Cluster analysis	1
Maximum likelihood	98
Maximum parsimony with successive weighting	6
Maximum parsimony without successive weighting	258
Neighbour joining	22