Org. Divers. Evol. 1, 257–272 (2001)

http://www.urbanfischer.de/journals/ode

ORGANISMS DIVERSITY & EVOLUTION

Best systematist practice transferred to molecular data

Georg Fuellen[1, 2, *], J. Wolfgang Wägele[1], Robert Giegerich[3]

[1] Ruhr-Universität Bochum, Lehrstuhl Spezielle Zoologie
[2] Integrated Functional Genomics, IZKF, Universitätsklinikum Münster
[3] Universität Bielefeld, Technische Fakultät

Received 2 February 2001 · Accepted 23 April 2001

Abstract

In the first part of the paper, we give a short description of the "minimum conflict" phylogeny estimation algorithm that analyzes molecular data in a way that resembles traditional best practice in morphological systematics. The algorithm calculates the tree from the root to the leaves, focussing on the detection of shared novel (synapomorphic) character states. Following the encaptic order of monophyla, at each step an outgroup serves to distinguish shared novel from shared old character states. The group of species under consideration is split in such a way that no synapomorphies of subgroups are torn apart, indicating that the split divides the group into two monophyla. In the second part of the paper, we describe the validation of our method with both natural and artificial data. Our method is very fast in theory, enabling the analysis of large quantities of data on a genomic scale. A Perl prototype is available on the World-Wide Web, via http://bibiserv.techfak.uni-bielefeld.de/mcope/.

Key words: evolution, phylogeny, molecular systematics, cladistics, divide-and-conquer

Introduction

The systematic analysis of morphological data usually consists of a three-step procedure: **First,** shared character states are noted for distinct groups of species. **Second,** these shared states are analyzed for their phylogenetic information content: How likely are the shared states unique features of putative monophyla? Following the cladistic approach sensu Hennig (1966), these **synapomorphies** are the sole currency of phylogeny estimation. Symplesiomorphies, that are character states shared by a group of species larger than the monophylum considered, do not provide evidence, nor do "shared" character states that in fact developed independently, due to convergent evolution. Usually, outgroup comparison is employed to tell apart synapomorphies and symplesiomorphies. **Third and last,** the phylogenetic tree that is in concordance with the largest number of putative synapomorphies is established. Formal definitions of the terms synapomorphy, symplesiomorphy and convergence are given in the appendix.

The justification for the three-step procedure lies in two principles: Descent with modification, and separation of populations with subsequent reproductive isolation. Since we attempt to formalize the procedure for molecular data, the validity of our "minimum conflict" method is based on the validity of these principles.

Material and methods

The minimum conflict algorithm

A high-level description of our algorithm runs as displayed in Fig. 1; a formal description can be found in Fuellen et al. (2001). The figure should be read columnwise from left to right. The example data used in the figure is kept deliberately simple. The outermost list (items 1 to 5) describe the basic divide-and-conquer scheme: the input alignment (item 1) is subjected to an heuristic search for a split (bipartition) that has a minimum of associated "conflict" due to putative synapomorphies torn apart. In the end, this split is found, and the two putative

***Corresponding author:** Georg Fuellen, IFG/Hautklinik, Von-Esmach-Str. 58, D-48149 Muenster, Germany; e-mail: fuellen@alum.mit.edu

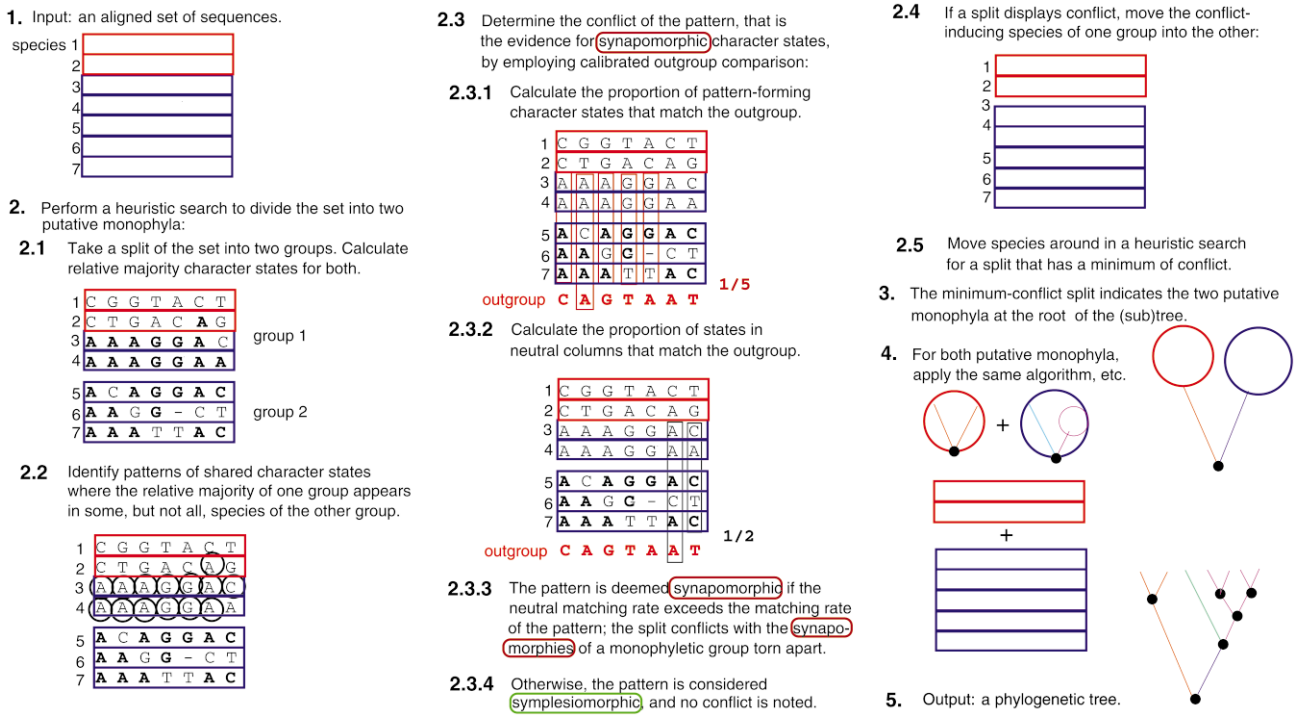

Fig. 1. Schematic overview of the minimum conflict method. (See text.)

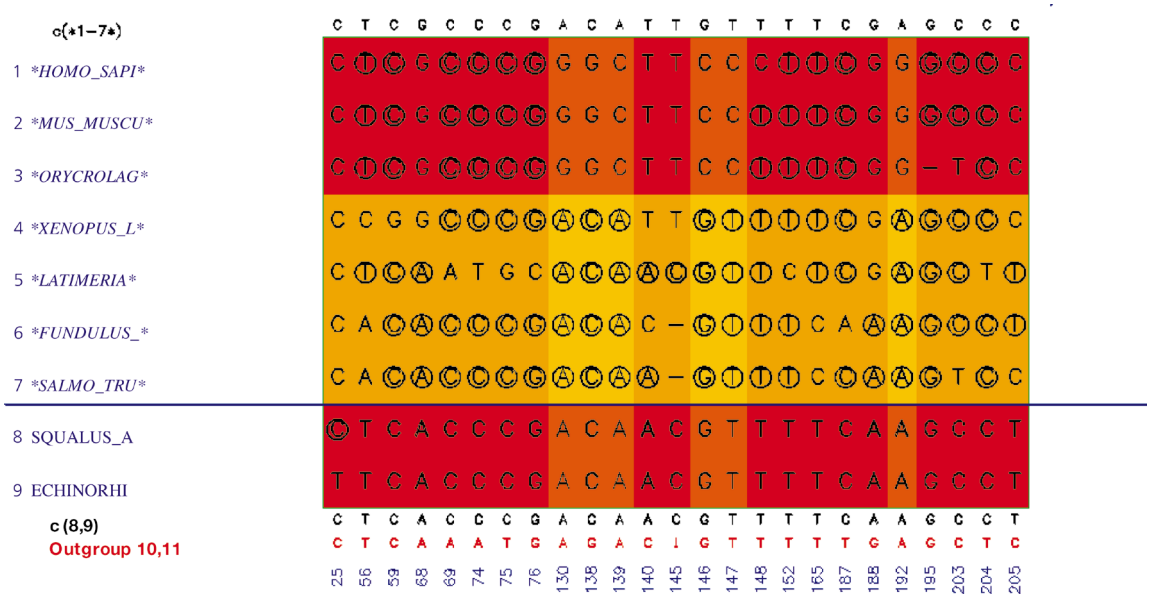

Fig. 8. The first 25 columns of the Chordata alignment, featuring gnathostomatan species (species 1–9) and displaying a pattern for the split of species 1–7 versus 8,9. The outgroup is the relative majority sequence of *Petromyzon* and *Lampetra* (species 10,11). On top of the boxed alignment, the relative majority sequence of species 1–7 is displayed; below the alignment, the relative majority sequence of species 8,9 is shown. The pattern "4–7" is found in columns 130, 138, 139, 146, 147 and 192. It tends to match the outgroup in 5 of 6 columns, and for the full-size alignment, the trend continues and we conclude that the common character states in species 4–9 are symplesiomorphies. Following the procedure outlined in Fig. 1, we calculate a very low conflict value of 0.266, which is also listed in Table 1. Columns 69, 74, 75, 76, 152 and 204 feature the pattern in species 1–4,6,7 (the first entry in Table 1). Despite the few matches shown, for the full-size alignment this pattern tends to match the outgroup even more, and it triggers a conflict of 0.

monophyla render the input alignments for two new invocations of the algorithm (item 4). Items 2.1 to 2.5 describe a split investigation:

- Relative majority sequences are calculated for both groups of species (2.1; majority character states are marked in boldface).
- The patterns of character states shared between species in the alignment are identified and tallied for both groups of species. Patterns are triggered by the occurrence of the relative majority character state of one group in the other. Only the patterns in group 1 are shown in 2.2. The first pattern indicates character states shared between species 3 and 4 in group 1 and the majority of species in group 2.
- Conflict values are calculated for patterns. In 2.3.1 and 2.3.2, we take the pattern "3,4" of the third and fourth taxon in group 1 (species 1–4) as an example.
- In case of conflict, species are transferred between the two parts of the split, resulting in the split of species 1 and 2 versus 3–7 in 2.4.
- Further species transfers may be done in search for minimum conflict.

Following the innermost list (items 2.3.1 to 2.3.4), conflict calculations consist of 3 basic steps:

- We infer the degree of matching between the pattern-forming character states and the outgroup (2.3.1). In our simple example, this matching rate is ⅕ since the pattern "3,4" occurs in the first five columns, but a match with the outgroup can be recorded for column 2 only.
- We infer an analogous degree of matching for character states in "neutral" columns (2.3.2). In our simple example, this matching rate is ½ since a match can be observed in one of the two available "neutral" columns (i.e. in one of the last two columns).
- We compare both matching rates, yielding a "novelty estimate". This comparison is the crucial step of our analysis. The hallmark of synapomorphies is their trend not to match the outgroup (they are "novel" character states). In contrast, symplesiomorphies tend to match (they are "old" character states).

We note the following important aspects:

1. **Calibration.** Outgroup comparison for molecular data must be done in a calibrated way. Just testing the rate *m* with which outgroup character states match the character states tested for synapomorphy status is not enough. After all, the outgroup may have evolved rapidly, gaining many nonconvergent and/or convergent modifications. We do not know the amount of outgroup evolution, nor do we want to estimate it directly. However, we take it into account by calculating the "novelty estimate" in relation to another matching rate, m_0. The variable m_0 is the rate of outgroup matches that are found in those "neutral" columns of the alignment which do not feature any shared states in need of testing. More precisely, these "neutral" columns do *not* display the majority character state of the other group, or they display this state in a subset of the species of *g* just by coincidence, forming a pattern that is only found in very few columns.

2. **Sigmoids.** We use sigmoid functions to model the decision process of a trained taxonomist. In particular, the comparison of matching rates is done using the sigmoid shown in Fig. 2. The more the matching rate in the neutral columns exceeds the matching rate observed for the pattern of shared states, the more confident our verdict is in favor of the hypothesis of synapomorphy (novelty) of the shared character states - we obtain a high novelty estimate. A low novelty estimate lets us favor the hypothesis of symplesiomorphy of the shared states. Then, these are the "leftover" resulting from the "erosion" (substitution) of shared states in the taxa that do not display the pattern.

Moreover, the sigmoid in Fig. 3 shows how the novelty estimate based on outgroup comparison is advised by a far weaker criterion, which is inherently phenetic: shared character states are more likely synapomorphic if they appear in highly evolved sequences, and they are more likely symplesiomorphic if the complementary set of species is highly evolved. As can be seen from the figure, no advice is possible if the outgroup-based novelty estimate is unambiguous.

Finally, another sigmoid function is used to "activate" (i.e. amplify or filter) the advised novelty estimate, and to make the connection to a statistical analysis of reliability. Reliability is given to patterns that are observed in

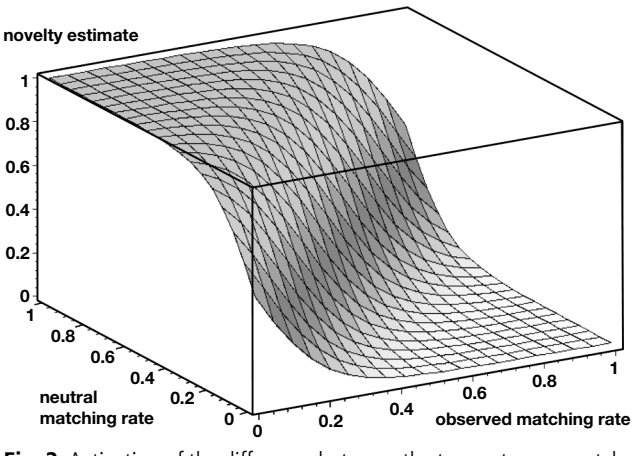

Fig. 2. Activation of the difference between the two outgroup matching rates, the neutral one and the one observed for the pattern. The more the former exceeds the latter, the more likely we are dealing with a pattern due to synapomorphies. The result is the "novelty estimate", our "outgroup-based" criterion for the detection of synapomorphic patterns.

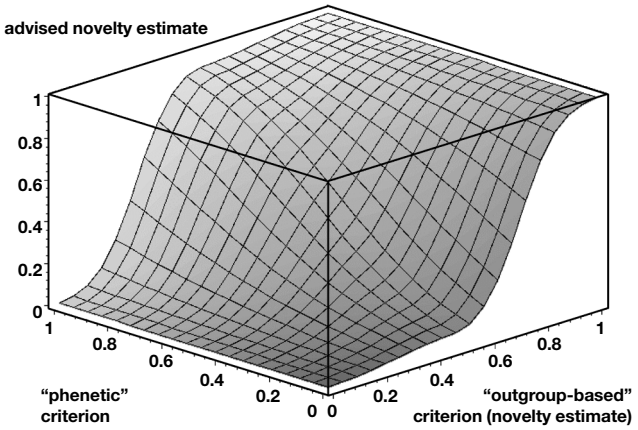

Fig. 3. The "phenetic" criterion advises the "outgroup-based" criterion (the "novelty estimate"). If the "outgroup-based" criterion is unambiguous, no advice is permitted.

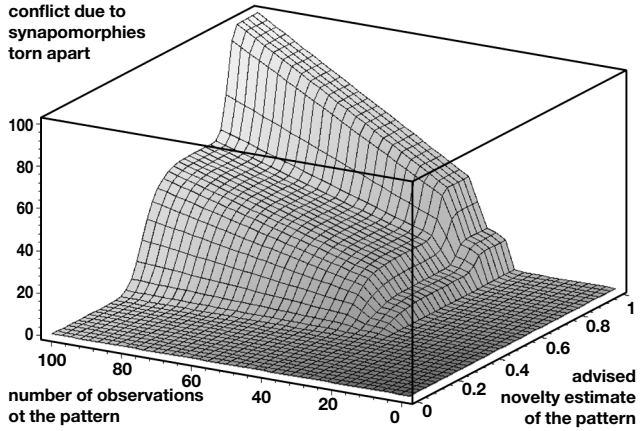

Fig. 4. The conflict triggered by a pattern is calculated from the advised "novelty estimate"; high conflict may result if the pattern is deemed synapomorphic (i.e. due to shared novel character states). However, conflict is not flagged if the number of observations (columns that display the pattern in question) is too small to yield a reliable estimate. Both conditions are evaluated and combined via a sigmoid function.

a "significant" number of alignment columns. Usually, this means that their number exceeds the average number of supporting columns, plus one standard deviation. Such a sigmoid is displayed in Fig. 4, and its effect is that patterns trigger conflict only if they are both reliable and synapomorphic. In Fuellen et al (2001), we employ a slightly modified statistical analysis termed "erosion-corrected reliability estimation". In this case, for each pattern the number of supporting columns is multiplied with the activated advised novelty estimate, and mean and standard deviation of the resulting distribution are calculated. Then, reliability is estimated in the context of this distribution, which cannot have outliers due to patterns that are triggered by a large amount of symplesiomorphic character states, since these have low advised novelty as long as they can be labeled "old" by our criteria. If such "erosive" patterns are predominant in a data set, it helps a lot to estimate reliability based on the erosion-corrected data.

3. **Relative majorities.** Our method detects the consequences of shared states torn apart, forming patterns of the majority character state of one group in the other. This detection is more powerful than a direct search for shared states. In the example of Fig. 1, item 2.1, *relative* majorities are calculated in group 2 (species 5–7; majority character states are marked in boldface), and pattern detection in group 1 (species 1–4) can reveal putative synapomorphies in species 3–7 that are already hidden because of subsequent modifications in group 2. As can be seen for item 2.2, only the first column has a visible putative synapomorphy. But even the fifth column contributes to the pattern, because we resolve ties in relative majority voting by resorting to lexicographic order.

Finally, by considering similarity of patterns, our method may take into account the last two columns as partial support for synapomorphies in 3–7, provided that species 2 is evolving at a higher rate than species 1, and species 4 is more evolved than species 3, where the rate of evolution is estimated in comparison with the outgroup. The details of this inclusion of the "neighbors" of a pattern can be found in Fuellen (2000).

4. **Heuristic search.** We conjecture that the search for minimum conflict is very efficient if there are still synapomorphies visible in the dataset. For all natural data sets analyzed in this paper, the search for minimum-conflict splits follows the **movespecies** heuristic described in Fuellen (2000), starting with the spectrum of all "singleton" splits where one species is separated from all the others. In case of artificial data, the search just starts with the single split that is supported by the largest number of sites.

5. **Choice of outgroup.** To apply the same algorithm for the two putative monophyletic groups found thus far, a new outgroup is needed for each of these. For outgroup selection, we test two candidates: the old outgroup, and the sister group. We favor the candidate promising more homogeneous matching rates for the next iteration. More precisely, we calculate the maximum and the minimum number of matches of individual species with the candidate, and we elect the one with a minimum spread between maximum and minimum. In the beginning, the researcher has to select an appropriate first outgroup based on his/her expert knowledge. This outgroup should be informative about the status of the character states in the ingroup under consideration; it should not be too close, nor should it be too far away. Outgroup selection is important for our method, cf. the discussion.

6. **Reasons for patterns.** A split of a set of species may trigger patterns for many reasons. As discussed, synapomorphies testifying the exclusive common her-

itage of subsets may be torn apart by a split. Then, at least one group in the split cannot be monophyletic, and we expect to observe a pattern. Or, symplesiomorphies may be torn apart. In this case, we say that the pattern is "due to erosion" of the symplesiomorphic character states in some taxa, creating a pattern in the taxa that are unaffected. Outgroup comparison is designed to tell apart these two cases. Or, convergences may induce patterns - as discussed in Fuellen et al. (2001), our approach cannot yet handle this case in a satisfactory way.

7. **Computational speed.** For balanced trees, the divide-and-conquer scheme enables the fast analysis of large sets of data. In the best case, a set of 100 taxa yields two subsets of 50 taxa on the next stage, etc. In particular, we do not need any tree topology searches like branch swapping, nearest neighbor interchange, etc. (cf. Swofford et al. 1996). In the worst case of a caterpillar tree, we still obtain a running time that is less than quadratic in the size of the alignment, cf. Fuellen (2000).

Related work

We find some related work in the papers of Woas (1990), Richardson & Stern (1997), and Wilkinson (1998). Woas' "sequence of splitting steps" is suited for morphological data only, the "apomorphic" characters underlying a split are determined by expert knowledge, and their number is simply tallied without involving systematic outgroup comparison and maintenance. The divide-and-conquer idea, however, is clearly visible, making explicit a method of successive analysis that has been practiced by some morphologists all along, following Hennigian principles. Richardson & Stern employ a questionable notion of "synapomorphy", a character state found in one group of species but not in the complementary group, and that is different from a "plesiomorphic" character state present in both groups. Then, they compute and evaluate splits in a completely different fashion, following the notion of an ingroup/outgroup relationship between the two sets of taxa that comprise the split to be found. In other words, the outgroup is not separate from the set of species analyzed. Furthermore, the method requires human intervention in particular for successive steps of analysis. Wilkinson's split analysis is based on permutations of the data, it does not involve any outgroup comparison, no statistical analysis of observations, and no sigmoid functions, nor does he suggest a hierarchical analysis of the data following the encaptic order of putative monophyla, and updating the outgroup along the way.

Established phylogeny estimation methods

For comparison, the *Phylip* package (Felsenstein 1993) was used to estimate phylogenetic trees by maximum parsimony and neighbor joining, and *fastDNAml* (Olsen et al. 1994) was used for maximum likelihood. In all cases, default parameters were used, supplementing the defaults of *Phylip/fastDNAml* by the ones used for the Pasteur Institute Web Interface (Letondal 2001). *Phylip* defaults imply a Kimura-2-parameter model for the distance matrix estimation, with a ratio of transition to transversion of 2.0. *fastDNAml* defaults imply equal empirical base frequencies of 0.25, a ratio of transition to transversion type substitutions of 2.0, input order jumbling (up to 10 times) until the same tree is found 2 times, and *quickadd* rearrangement. In all cases, 1000 bootstraps were performed with random seed 1. It is future work to compare our method with variants of the established approaches, as well as with other novel approaches.

Data – general considerations

Intensive validation on both natural and artificial data is desirable for any new approach to phylogeny estimation. We have selected natural data from two sources. We assemble datasets from an alignment database, and we reinvestigate published studies. A simulation study using artificial data provides supplementary evidence. In this case, the phylogeny is known, and we can identify the cases where our algorithm prefers an incorrect split. Some intensive research gives the impression that there are currently no benchmark datasets for phylogeny estimation, neither published in the literature, or on the World-Wide Web. The tradeoff between unrealistic assumptions employed for the generation of artificial data, and the impossibility to know the true phylogeny for natural data, poses major problems for benchmarking phylogeny estimation methods.

We have constructed our own datasets from the mitochondrial 12S-rDNA alignment of the RDP (Ribosomal Database Project, Maidak et al. (2000), using data as of June 2000), which is guided by structural information. An important point for this kind of study is the choice of species – how can it be done in a rigorous way, and still yield a set of species for which an "undisputed" tree can be constructed? The RDP database offers a sequence query facility (the "Phylogenetic Tree Browser") that has a crude phylogenetic organization which we can finetune, and then we can run a rigorous procedure to obtain a selection of species that is both objective and classifiable to a very high degree. This procedure amounts to the mechanic rule "Always take *the first two* taxa", at each level of the finetuned phylogeny. The rule also helps us to select species such that the tree is "almost" undisputed; adding a third taxon would imply that a debate is possible on the correct classification of the three taxa. (Our rule does not select the most "representative" taxa; "representative" is a subjective criterion

that is sacrificed in favor of a strict rule that just uses the rather arbitrary order in the listings given to us. We note that, usually, reconstructing phylogenies becomes easier if "representative" species are used for the various groups.)

Tetrapoda data

Tetrapoda are included in the listing of vertebrate mitochondrial 12S-rDNA from the RDP "Phylogenetic Tree Browser", which consists of 6 sublistings: "Jawless vertebrate", "Amphibia", "Fish", "Birds", "Reptiles", and "Mammals". We ignore "Reptiles", since it consists of taxa for which the correct placement with respect to "Birds" and "Mammals" is not certain. Then, tetrapod taxa may be selected from "Amphibia", "Birds" and "Mammals", while the outgroup comprises the first two taxa from the remaining two groups, "Jawless vertebrate" and "Fish". "Jawless vertebrate" includes one taxon only (*Petromyzon*), and the first two fish with a complete 12S-rDNA sequence are *Crossostoma* and *Cyprinus*. The relative majority sequence of these 3 taxa comprises the outgroup.

Regarding the Tetrapoda, we assume that an undisputed tree can feature the following monophyla: Amniota (versus Amphibia), Mammalia (versus Aves (Birds)), Eutheria (versus Marsupialia) and Primates (versus Rodentia). The only amphibians with non-fragmentaric 12S-rDNA can be found in "Group I"/"Subgroup II" of the RDP listing, and we take the first two *Rana catesbeiana* sequences from there. The first two birds with complete 12S-rDNA found in the listings are in "Group I"/"Chicken and Relatives"; these are *Gallus* and *Coturnix*. The first two such marsupials are *Phalanger orientalis* and *Phascolarctos cinereus*. For Amphibia, Aves and Marsupialia, there is no second group of taxa for which both an undisputed placement is possible, and at the same time the 12S-rDNA is not fragmentary. For Primates and Rodentia, we can select the first two such groups with two representatives each, always selecting the first two taxa, skipping those with fragmentaric DNA.

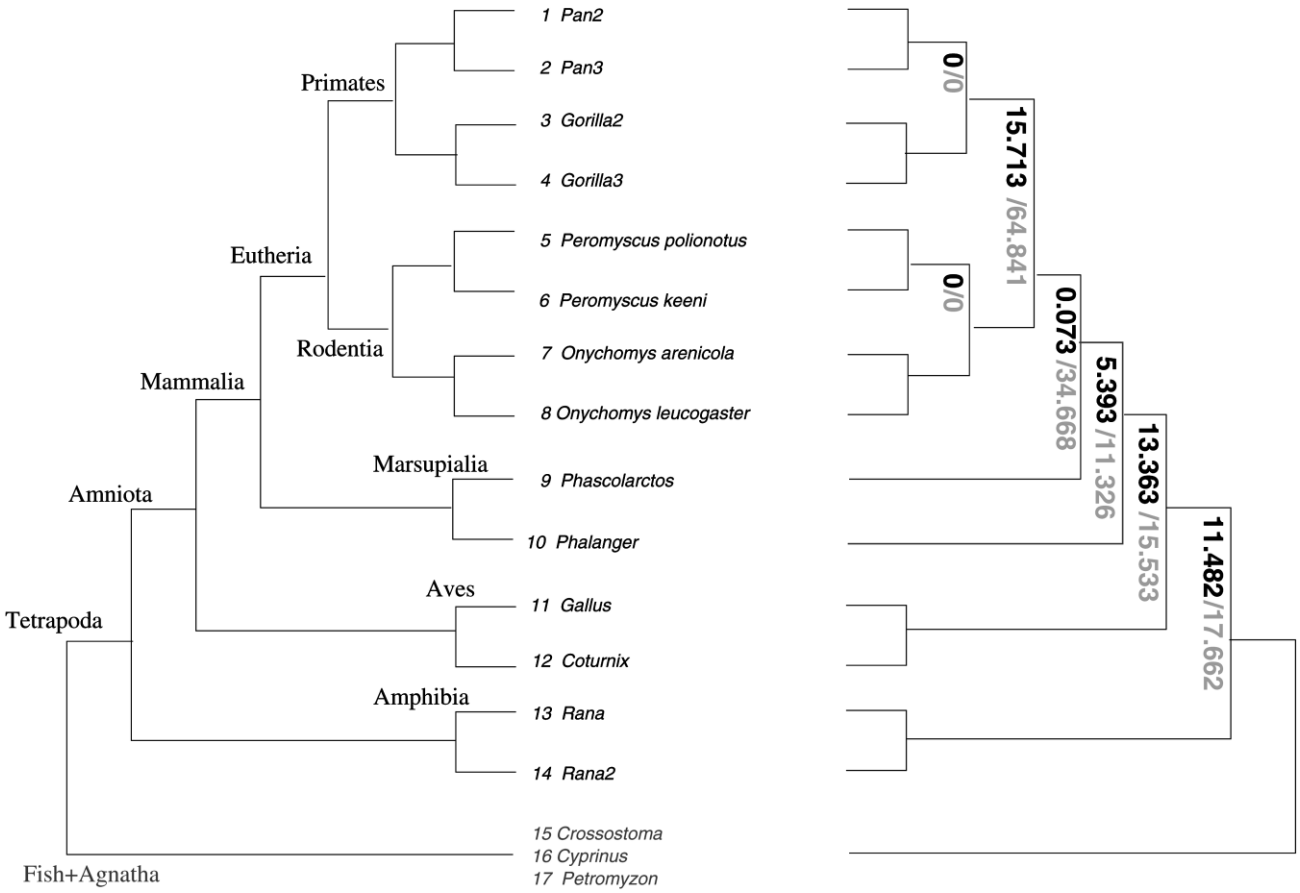

Fig. 5. Undisputed tree (left) and minimum conflict tree (right) for the Tetrapoda dataset (mitochondrial 12S-rDNA). The undisputed tree is recovered by all standard methods with bootstrap values exceeding 91% (1000 bootstraps). Nodes in the minimum conflict tree are labeled with the associated minimum conflict value of the split represented by the node, and the conflict value of the runner-up split that came next in the heuristic search for the best split. The runner-up split is not shown, but its conflict is indicated in grey.

Fig. 5, left, displays the "undisputed" tree for the species that we have just selected, assuming monophyletic Amniota, Mammalia, Eutheria, and Primates. The aligned sequences feature 867 sites that do not contain a symbol for a missing residue, 521 of which are variable.

Mammalia data

The Mammalia dataset is also taken from the mitochondrial 12S-rDNA alignment of the RDP database, following the same procedures as for the Tetrapoda dataset, except that we are interested in the debated split between the Eutheria (placental mammals), Marsupialia (opossums, kangaroos, etc.), and Monotremata (platypus, echidnas).

Analyzing mammalian taxa and following the "take *the first two*" rule, we retrieved the species included in the tree of Fig. 6, left. The alignment features 827 sites without a missing residue, 413 of which are variable.

Chordata data

The Chordata dataset is a subset of the Metazoa dataset analyzed by Rödding & Wägele (1998). The underlying alignment is provided by the authors, who used a computerized procedure (ClustalW, Thompson et al. 1994) and then maximized the number of invariable sites by eye. It contains 1974 sites, 662 of which are variable. The most plausible tree is given in Fig. 7, left.

Artificial data

For the generation of artificial datasets, the tool ROSE (Random Generation of Nucleotide Sequences, Stoye et al. 1998), Version 1.0.1, was used. ROSE allows a wide array of parameters; we restrict our analysis to nucleotide sequences generated under the following setup:

- A random tree topology with 32 leaves is constructed by ROSE. The m u t a b i l i t y, that is the percentage of nucleotides modified along one branch of the tree

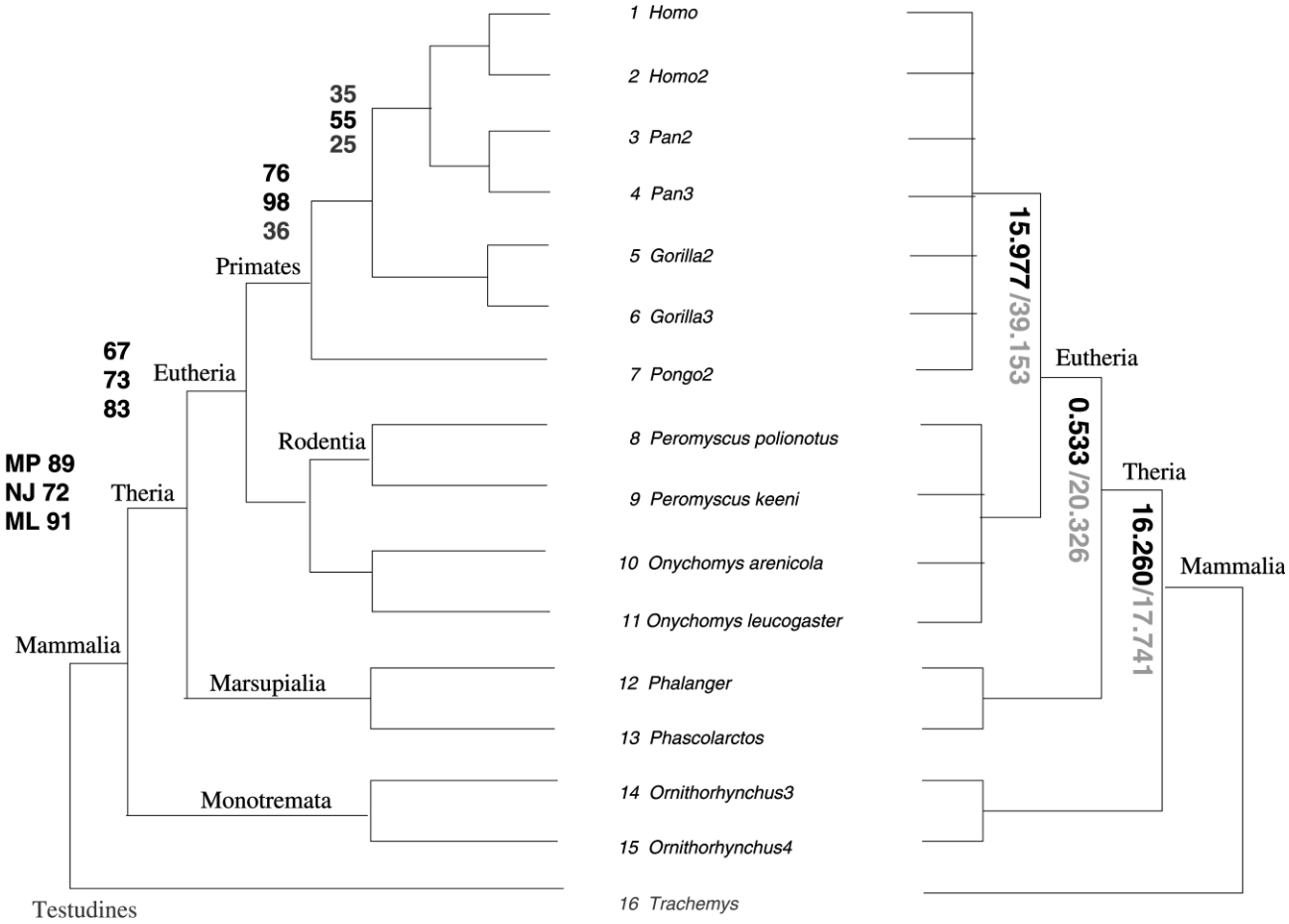

Fig. 6. Most plausible tree (left) and minimum conflict tree (right) for the Mammalia dataset (mitochondrial 12S-rDNA). Nodes in the most plausible tree are labeled with their bootstrap support (1000 bootstraps) obtained via maximum parsimony (MP), neighbor joining (NJ) and maximum likelihood (ML). In case of bootstrap values below 50 the corresponding node does not appear in the consensus tree estimated by the method in question. If bootstrap values exceed 96% for all methods, they are not listed. The other labels are explained in Fig. 5.

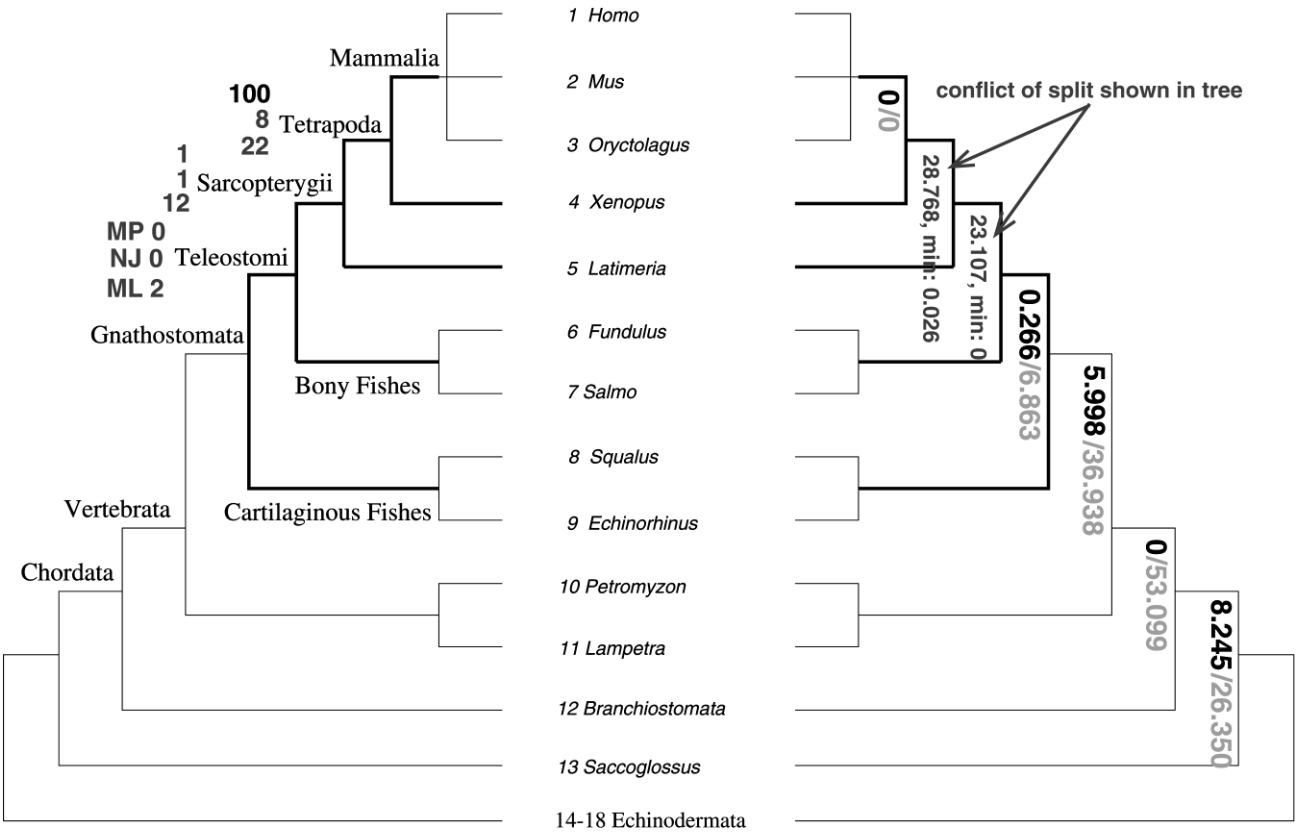

Fig. 7. Most plausible tree (left) and conflict tree (right) for the Chordata dataset (18S-rDNA). On the left, parsimony, neighbor joining and likelihood bootstrap values are attached; on the right, conflict labels are added, cf. Fig. 5 and 6. In two cases, the tree on the right does not reflect the minimum conflict tree. These cases are marked by an arrow: the first label reveals the conflict value for the split represented by the node. This conflict is not minimum, and the conflict of the minimum-conflict split for the set of taxa at the node is indicated as well. This minimum-conflict split is incorrect if we follow morphological data; it separates species 1–3 from 4–7 in case of species 1–7 (conflict 0), and species 1,3 from 2,4,5 in case of species 1–5 (conflict 0.026). Within the Gnathostomata, standard methods favor the following nodes in contrast to the "Teleostomi versus *Squalus* and *Echinorhinus*" split: Maximum parsimony: "*Latimeria* versus the other species" (bootstrap 100%); Neighbor joining: "Mammalia versus the other species" (bootstrap 70%); and Maximum likelihood: "*Fundulus* and *Salmo* versus the other species" (bootstrap 56%). Minimum conflict recovers the correct split.

(also known as the "percent accepted mutations", or PAM) is set by editing the tree lengths, and feeding the tree back to ROSE, for sequence generation only. This manipulation is necessary since, per default, ROSE 1.0.1 generates trees with a very unequal branch length distribution, if the PAM-value (expressed in terms of "average distance") is supplied directly (see Fuellen 2000).

- 32 sequences are generated as described, with average sequence lengths of 500, 1000 and 1500 nucleotides. We then split the tree at the root, and the first subtree contributes the sequences to be analyzed, whereas the second subtree contributes the outgroup by taking the relative majority character state at each sequence position of the corresponding subalignment. This mechanism results in trees of different size and topology; the average size is 16 sequences.
- The following simple substitution probability matrix is used. Nucleotides are substituted with 1% probability per unit of branchlength (1 PAM), and all kinds of substitutions are given the same probability of 0.003333.

$$M = \begin{pmatrix} m_{A\to A} & m_{A\to C} & m_{A\to G} & m_{A\to T} \\ m_{C\to A} & m_{C\to C} & m_{C\to G} & m_{C\to T} \\ m_{G\to A} & m_{G\to C} & m_{G\to G} & m_{G\to T} \\ m_{T\to A} & m_{T\to C} & m_{T\to G} & m_{T\to T} \end{pmatrix}$$

$$= \begin{pmatrix} .99 & .003333 & .003333 & .003333 \\ .003333 & .99 & .003333 & .003333 \\ .003333 & .003333 & .99 & .003333 \\ .003333 & .003333 & .003333 & .99 \end{pmatrix}$$

- In ROSE, insertions and deletions depend on the "average distance" value (called "Relatedness" in the

ROSE 1.0.1 interface), user-supplied "thresholds", and user-supplied indel length functions. As described in Stoye et al. (1998), the PAM-value is multiplied with the depth of the tree, yielding the "average distance". The insertion and deletion thresholds are both set to 0.1, and the length of both insertions and deletions follow the following distribution:

$$frequency\ (length) = 1/(2 \cdot 1.5^{length}),$$

not considering lengths larger than 10. The following table lists the length distribution explicitly; note that indels larger than 10 may nevertheless appear in the sequences, due to the multiple indels.

length	1	2	3	4	5
frequency	0.333	0.222	0.148	0.099	0.066
length	6	7	8	9	10
frequency	0.044	0.029	0.020	0.013	0.009

32 independently created phylogenies were analyzed using minimum conflict. In other words, 32 runs with the same average number of sequences (i.e. approx. 16), the same average sequence length (i.e. 500, 1000 or 1500 nucleotides), and the same mutability were performed for each data point. In a few cases, execution of the software was terminated prematurely due to external factors; in any case, at least 30 phylogenies were created.

For each run, the **error count** was calculated as the number of incorrectly established splits across the whole tree. Whenever an incorrect split was favored, the error count was incremented and the calculation was resumed with the correct split as if nothing happened.

The **error rate** of a single run is defined as the relative frequency of error, i.e. the error count is divided by the number of splits to be estimated for the tree under consideration. The error rate of a set of runs is the average error rate taken over all runs performed.

Results

First, we compare the accuracy of our minimum-conflict method with standard methods, using natural data. We encounter cases where our method is inferior as well as cases where standard methods perform worse.

Tetrapoda data

The undisputed tree in Fig. 5, left, is supported by bootstrap values exceeding 91% using the standard methods maximum parsimony, neighbor joining, and maximum likelihood. Fig. 5, right, displays the minimum-conflict tree, where each node is labeled by the minimum conflict value that made us prefer the split in question, followed by the conflict value of the runner-up split encountered by the heuristic search. In other words, the first node establishes the Amniota versus Amphibia split, which has nevertheless a spurious conflict of 11.482. However, the second-best split (separating one bird and the Amphibia from the other species) has a much higher conflict of 17.662, so we can feel confident that the correct split is supported by our analysis of the data. This confidence cannot be gained for the next node, where the correct split (Mammalia versus Aves) has minimum conflict, but there are close runners-up. (The two runners-up are the splits where one marsupial species branches off.) The next node, where *Phalanger* branches off and the other marsupial forms a subtree with the other mammals, is in fact erroneous even though there is no close runner-up; non-monophyletic Marsupialia are very implausible, and we may have run into an artifact, possibly due to a poor choice of the outgroup, cf. the discussion. The following splits are all correct, but resolution is missing in the end due to an insufficient number of variable columns. While 521 variable columns are analyzed in the beginning for the Amniota versus Amphibia split, this number drops to 462 for the Mammalia versus Aves split, and it is 369 for the incorrect split involving non-monophyletic Marsupialia. Analysis continues with 349 variable columns, and the last good split of Primates versus Rodentia is based on 275 columns. Finally we deal with 47 and 61 variable columns, respectively, and resolution is lost.

Employing erosion-corrected reliability estimation (cf. Material and methods, The minimum conflict algorithm, note 2), minimum conflict infers the undisputed tree (data not shown).

Mammalia data

The tree shown in Fig. 6 may be debated as far as the position of the Monotremata is concerned. On the left, bootstrap values obtained for the standard methods are displayed; some plausible groups within the Primates receive insufficient support in the parsimony and likelihood analyses. On the right, the minimum conflict tree is displayed; the split of species 1–13 versus 14,15 (Theria versus Monotremata) has a minimum conflict of 16.260, and it is closely followed by 12 versus 1–11,13–15 (conflict 17.741), 1–12 versus 13–15 (conflict 20.417), 12,13 versus 1–11,14,15 (conflict 22.280), 8–11 versus 1–7,12–15 (conflict 25.960) and 1–11 versus 12–15 (conflict 28.732). The last split listed refers to the "Marsupionta" hypothesis, proposing a monophyletic group of Marsupialia and Monotremata (cf. Janke et al. 1996, 1997, Penny & Hasegawa 1997). The other splits (Eutheria versus Marsupialia, and Primates versus Rodentia) are well supported. No further resolution is possible for the primates; there are seven splits with zero conflict.

While 413 variable sites are given in the alignment of species 1–15, 378 are given for species 1–13, and 294 are still available in the alignment of the Eutherian species 1–11, no resolution is possible for the 107 variable columns in the projected alignment of the primate species 1–7.

In contrast to the Tetrapoda data, employing erosion-corrected reliability estimation yields an "inferior" tree where the split Eutheria versus "Marsupionta" has a minimum conflict of 14.570, followed by the Theria versus Monotremata split with conflict 16.260. However, the difference between the minimum-conflict split and the followup-split is small, as it was when we obtained the more plausible tree above. In general, caution should be exercised with respect to the confidence in tree nodes based on a narrow decision.

Chordata data

As can be seen in Fig. 7, left, standard methods are not able to recover the Teleostomi group (species 1–7); at most two percent of the trees estimated include this putative monophylum. We get the tree right up to this point (Fig. 7, right). Thereafter, minimum conflict is as incorrect as the standard methods, favoring a zero-conflict split 1–3 versus 4–7 over the correct split 1–5 versus 6,7, which obtains a conflict of 23.107. For the Chordata dataset, we recognize 662 variable sites in the beginning, and this number drops to 341 for species 1–9, and to 323 for the subtree of species 1–7. In the end, 50 variable columns are left in the alignment of the sequences of species 1–3. Erosion-corrected reliability estimation yields the same tree.

Other natural data

We reinvestigated another study, Friedrich & Tautz (1995), and again the minimum conflict tree is more plausible, even though not all putative correct monophyla are recovered by our method either, see Fuellen (2000). We have also constructed a set of 14 Bilaterian 18S-rDNA sequences from the RDP database, following the "Always take *the first two* taxa" rule, and minimum conflict recovers the undisputed tree topology with no errors (see Fuellen 2000, Fuellen et al. 2001). In contrast, neighbor joining and maximum likelihood cannot correctly recover the first split of the set of species, and their bootstrap support for the correct split is only 5% and 28%, respectively. For a set of Gnathostomata taxa, we encountered some more problems, as described in Fuellen (2000). These problems consist of incorrect minimum-conflict splits with very close followups, and standard methods perform no better (data not shown).

Artificial data

Results obtained by inspecting random trees with sequences of approximately 1500, 1000 and 500 nucleotides are shown in the following set of figures, Figs. 9–11. In all 3 figures, the vertical axis is labelled with the average error rate established over at least 30 runs, and the horizontal axis is labelled with the specifics of the run. This PAM-value, or mutability, is the number of applications of the substitution probability matrix along *one* branch of the artificial tree, from the ancestral to the descendant node. Naturally, the percentage of substitutions introduced along a path of branches is much larger.

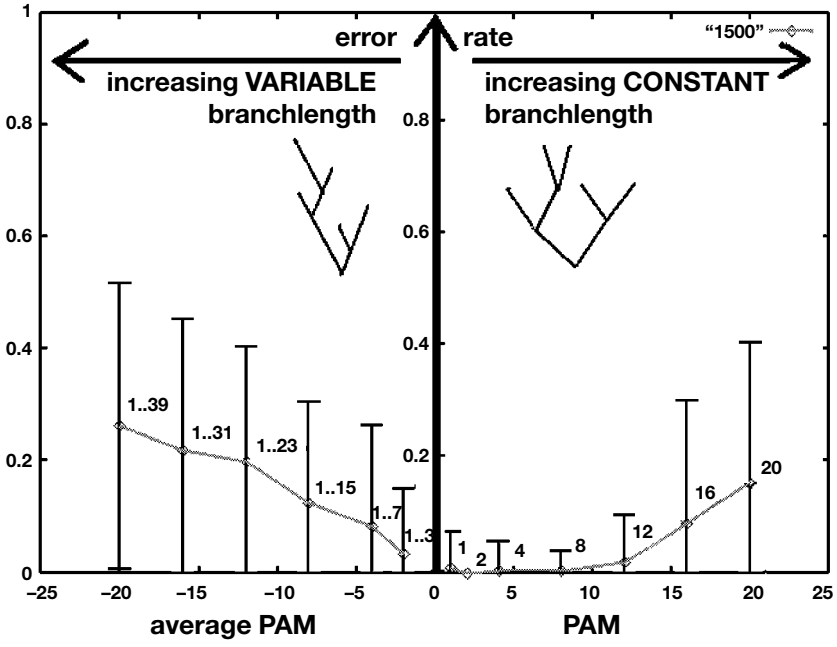

Fig. 9. Error rates for artificial data, 1500 nucleotides. The vertical axis is the frequency of incorrectly estimated splits, averaged over approx. 32 runs with trees of random topology consisting of approx. 16 species. The horizontal axis is interpreted as follows: on the right side, the percent accepted mutations along one branch of a random tree with equal branchlengths is given; on the left side, the **average** percent accepted mutations is given, where trees are constructed with variable branchlengths. These branchlengths are in the interval indicated in the plot, ranging from 1 to 39 PAM for an average of 20 PAM, and from 1 to 3 PAM for an average of 2 PAM. Error bars indicate an approximate 95% confidence interval. Sample patterns for split 1–7 versus 8, 9 are displayed in fig. 8 on page 258.

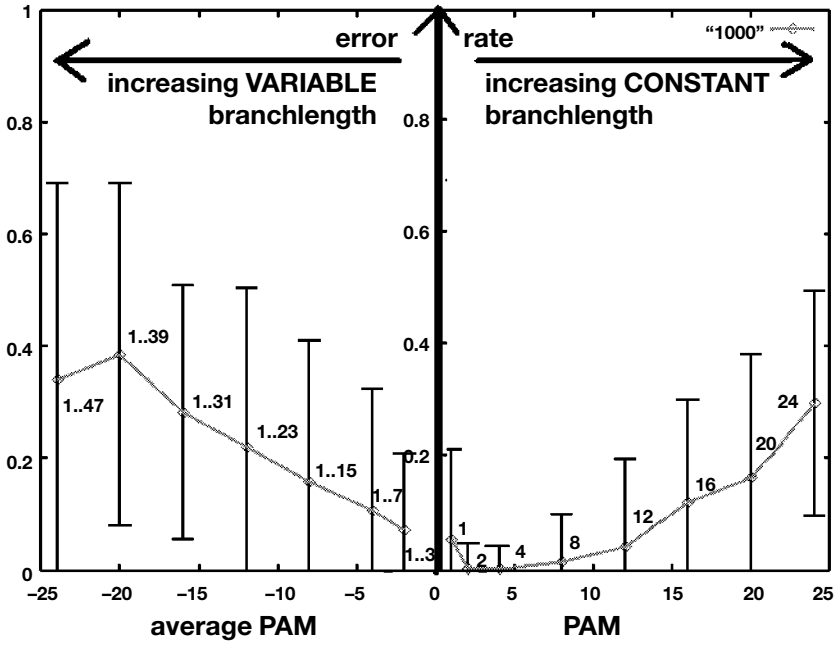

Fig. 10. Error rates for artificial data, 1000 nucleotides. See Fig. 9 for further explanations.

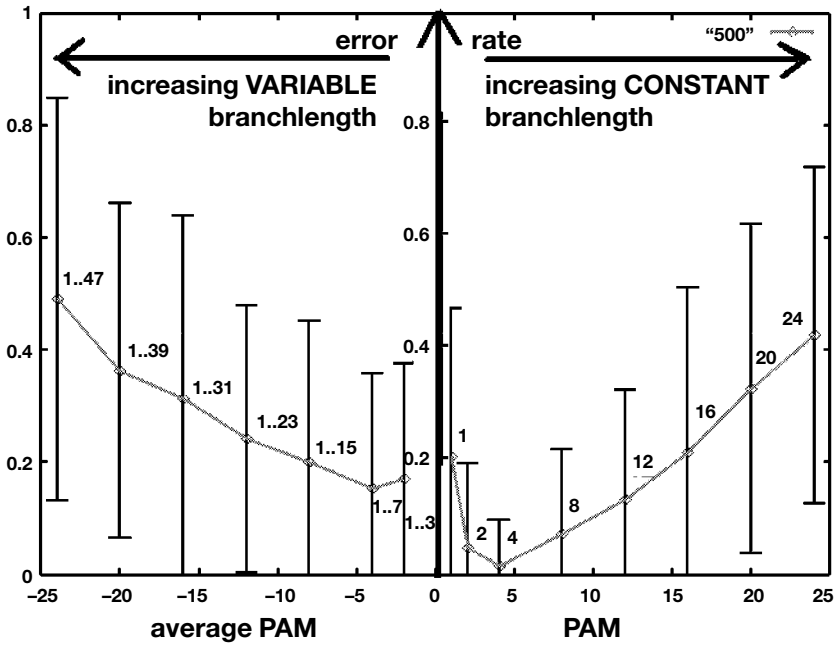

Fig. 11. Error rates for artificial data, 500 nucleotides. See Fig. 9 for further explanations.

For example, the percentage of substitutions between two sister group sequences is almost twice the mutability, unless multiple hits cause saturation effects in case of very large mutabilities. For the right-hand side of the figures, the mutability is the same for each branch, whereas on the left-hand side it varies as indicated, and the axis refers to the *average* mutability, marked as such by the minus ("–") sign.

The figures include so-called "error" bars for purposes of decoration only, based on an unsound attempt to calculate a standard "error" of the mean as a basis for 95% confidence intervals. Our calculation of "error" bars is unsound because error rates do not feature a symmetric distribution around the mean value - there are no negative rates. We also note that 30 runs are just enough to calculate "error" bars, see Swinscow (1997).

Using approx. 16 sequences with 1500 nucleotides, 32 independently created phylogenies established an error rate of zero or almost zero given a mutability of 1–12 PAM (Fig. 9, right). Larger mutability triggers

more and more error, like 8% for 16 PAM, and 17% for 20 PAM. If the mutability varies randomly between 1 and 3 PAM on different branches of the tree, we observe about 3% error. Again, longer branchlengths result in more erroneous tree estimates (Fig. 9, left). To some degree, the large error rates to the left can be expected; after all, mutability variation between 1 and 39 PAM implies almost complete randomizations taking place across some branches of the tree, and randomization effects are even much stronger along paths of the tree.

Fig. 10, right, displays similar results for an average sequence length of 1000 nucleotides. We observe an average error rate of zero or almost zero given a mutability of 2–8 PAM. A smaller mutability of 1 PAM renders the alignment less informative, so we can indeed expect a higher error rate. Again, larger mutability triggers more and more error (Fig. 10, right). If the mutability varies randomly between 1 and 3 PAM on different branches of the tree, about 7% error results, and for variation between 1 and 7 PAM we observe about 10% error. Again, larger branchlengths result in more erronenous tree estimates (Fig. 10, left).

For sequences with only 500 nucleotides, we note much higher error rates (Fig. 11, right), unless mutability is around 4 PAM. In particular, the rise in error for small mutability is more pronounced. Moreover, in these cases variation of branchlength within the trees studied increases the error rate quite a lot (Fig. 11, left). For example, we record about 16% incorrectly estimated bipartitions in the case of variation between 1 and 7 PAM. However, for half of these errors the correct bipartition has a conflict value very close to the erroneous one (data not shown), and polytomies should have been flagged.

Discussion

Theoretical considerations

Let us compare our method to other approaches. For a detailed explanation of these, and a comprehensive overview of phylogenetic systematics in general, the reader is asked to consult Swofford et al. (1996). We are aware that some of the issues raised are hotly debated, and that our rather simple-minded comparison is useful mostly to put our method in perspective.

Distance methods for phylogeny estimation calculate distances between **pairs** of sequences, based on the number of character states that do not match. The pairwise distances are used to build up a phylogenetic tree. Our method, however, is character-based, analyzing multiple sequences simultaneously in a position-by-position fashion. This makes more use of the information provided by the individual sequence positions. Furthermore, modifications may occur in the development of the present-day species only, after the last separation event. These amplify distances between close relatives, and may misguide the analysis. In an attempt to deal with this specific type of long-branch artifact, corrections of distance estimates are often employed, and they are obtained by using specific models of character state evolution.

Such detailed models of evolution, including specific "substitution rates" for different classes of modifications, cannot be universally valid, and their estimation from data analyzed before constitutes circular reasoning: what if the trees used to estimate model parameters are incorrect? In this case, a systematic error is introduced into the model, and it may be reinforced by the further analyses of similar data.

Like our approach, **maximum likelihood** analyses are character-based. These evaluate a so-called likelihood function for each sequence position, and combine the results. Likelihood gives high scores to trees for which the modifications conform to an estimate about which classes of modifications are likely, and which ones are not. Therefore, the likelihood function relies on a detailed model of character state evolution, which we would like to avoid for the same reasons as in the case of distance methods. Furthermore, symplesiomorphies as well as noise caused by random modifications and the variation of substitution rates along branches of the tree may nevertheless influence the reliability of the results of maximum likelihood analyses.

Parsimony analysis is also character-based, evaluating another (but similar, see Tuffley & Steel 1997) scoring function for each sequence position and combining the results. Parsimony gives high scores to trees which explain the data with a minimum of modifications. No detailed model is needed for the parsimony function, but a lot of the noise caused by random modifications may influence the result of parsimony. If positions are weighted differently (see Swofford et al. 1996: 502–503, for a review), and/or the correction suggested by Salisbury (1999) is applied, the problem may at best disappear at the expense of additional complexity that allows for other systematic errors. Most importantly, we now discuss how parsimony may be misled by symplesiomorphies.

Parsimony can "fall into the erosion trap" (i.e. it may place taxa displaying symplesiomorphies into one subtree, cf. Fullen et al. 2001) because trees for which the less evolved sequences form a subtree require less modifications, if the more evolved sequences match the character states estimated for the node at which they are all attached. This matching may be viewed as a low level of long-branch attraction (Felsenstein 1978).

In Fig. 12, panels 1 and 2 feature the same character states at the leaves of the tree. The difference is in the tree itself: species 2, representing "the long branches" in this case, is attached to different edges, and we assume

that the left-side tree is the true tree. In panels 3 and 4, the character state of species 6 is T, but the tree in panel 3 is the same as in panel 1, and the tree in panel 4 coincides with the one in panel 2. Inferred modifications are marked by thick lines. If species 6 has a character state that is the same as for species 1,3–5, but different from species 2, we need two modifications to explain its evolution, no matter which tree we take (panel 1 or panel 2). However, if species 6 has a character state matching species 2, the true tree on the left requires three inferred modifications (panel 3), and parsimony will favor the false tree where species 2 and species 6 are "together", and there is no need for a modification "between them" (panel 4). In the most parsimonious tree, the short-branch species 1,3–5 then form one subtree. For panel 3, we remark that three modifications are also needed if we assume that the last common ancestor of 1–6 has character state T. Moreover, there is more than one tree that is both false as well as most parsimonious; species 2 and 6 in the tree to the right (panels 2 and 4) may as well be placed in a common subtree, and the number of modifications is two for such a tree as well.

In contrast to all three standard methods, our method builds the tree bottom-up, from the root to the leaves. At each level we do a simultaneous analysis of the relevant sequences only. As we have seen, this strong focus has a very pleasant side-effect: at least for balanced trees, calculations can be very fast.

Of the existing approaches just discussed, parsimony and maximum likelihood in particular are reaching their limits for large datasets, especially because these require a (heuristic) search of the space of all possible trees, evaluating the scoring function very many times. Researchers are reworking these methods in order to make them faster. For example, "PUZZLE" (Strimmer & von Haesler 1996) is an heuristic method that does maximum likelihood calculations for sets of 4 species, and then assembles the subproblem solutions recursively. For parsimony, "Iterative Fixation" (Salisbury 1999) is one algorithm that combines species into archetypes to speed up calculations. However, these two approaches are limited by both the drawbacks of the underlying method, and the heuristic nature of the speedup.

Then again, the divide-and-conquer paradigm already plays a key role in tackling large phylogenies. A distance-based generic algorithm called "Disk-Covering" is described in Huson et al. (1998) and Huson et al. (1999). "Disk-Covering" is similar to our approach because it also divides the set of species into subsets, and then combines the subtree solutions. Its reliance on distance calculations makes it susceptible to the problems already discussed for distance trees. Nevertheless, "Disk-Covering" holds a lot of promise, in particular because it has shown desirable characteristics (so-called "fast convergence") for sufficiently large sequences.

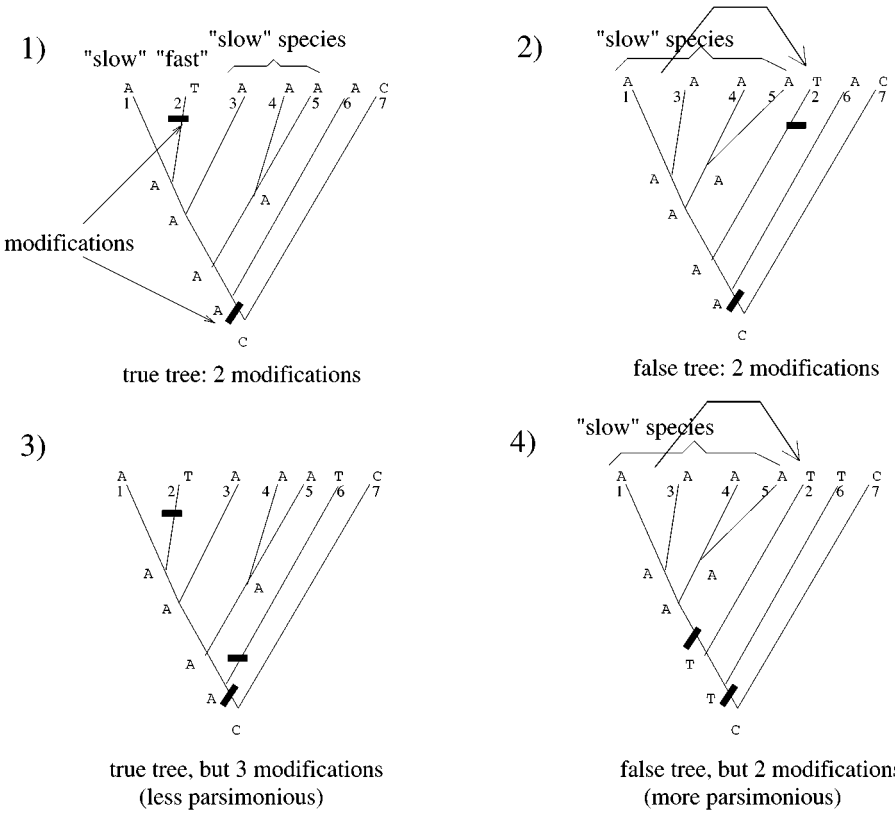

Fig. 12. A conceptual example where parsimony is misled by symplesiomorphies. (See text.)

Natural data

To summarize our results, we note that if we compare the performance of our method with standard methods, we find cases where all methods agree on the plausible monophyla, cases where all methods fail, and cases where only some methods fail. Failure of our own method for the Tetrapoda dataset may be attributed to an unsatisfactory selection of the outgroup, since the problem disappears in the case of **para**phyletic outgroup maintenance: if we select the new outgroup of a set of species as the **union** of a) the first species of the sister group and b) the old outgroup, Marsupialia are recognized as a monophyletic group (minimum-conflict 0, runnerup-split with conflict 13.83). We believe that this is a general improvement over the current usage of **mono**phyletic outgroups; paraphyletic outgroups are able to pinpoint a maximum of symplesiomorphies that are not lost in the majority of its constituents. We only use the first species of the sister group to avoid sampling problems; otherwise a large number of species in the sister group would dominate the outgroup. However, it would be even better to use not the first, but the best-suited species from the sister group, and we are currently investigating this issue.

Failure of parsimony and likelihood within the Primate group of the Mammalia dataset is probably due to noisy data, since the incorrect groupings do not have sufficient bootstrap values either – they range between 43 and 45% supporting the groups *Gorilla* and *Pan* in case of parsimony, and *Homo*, *Gorilla* and *Orang* in case of likelihood.

Failure of all standard methods in recovering the Teleostomi group within the Chordata dataset can be attributed to symplesiomorphies, as follows. If we investigate the correct split within the Gnathostomata (i.e. Teleostomi versus *Squalus* and *Echinorhinus*, denoted 1–7 versus 8,9 following the numbering scheme in Fig. 7) using minimum conflict, we observe 5 patterns within the Teleostomi with more than 10 supporting columns. They are listed in Table 1. None of these patterns triggers conflict because all have low advised novelty estimates – we can assume that each pattern is caused by symplesiomorphies, i.e. character states still present in the species forming the pattern, but eroded in the species not part of the pattern. More precisely, these symplesiomorphies encompass the pattern species within the Teleostomi and the (majority of the) two species of the sister group of the Teleostomi (*Squalus* and *Echinorhinus*). We assume that the three different kinds of internal nodes suggested by the standard methods are based on such symplesiomorphies, because each of them corresponds to one pattern in the table: the node "*Latimeria* versus the other species" (split 5 v 1–4,6–9) is found by parsimony with bootstrap support of 100%, and it corresponds to the erosive pattern "1–4,6,7". In other words, it is based on symplesiomorphies that encompass species 1–4,6–9. In the presence of a method artifact, bootstrap support of 100% is indeed no evidence for correctness of the phylogenetic tree calculated. The split "Mammalia versus the other species" (split 1–3 v 4–9) is established by neighbor joining (bootstrap 70%), and it corresponds to the pattern "4–7". It is based on symplesiomorphies in species 4–9. Their pattern is marked in yellow in Fig. 8, on page 258. Finally, the split "*Fundulus* and *Salmo* versus the other species" (split 6,7 v 1–5,8,9) is favored by maximum likelihood (bootstrap 56%), corresponding to the pattern "1–5", and based on symplesiomorphies in 1–5,8,9.

Like the standard methods, minimum conflict cannot recover the Sarcopterygii within the Chordata dataset. On the one hand, the correct split between (a) Sarcopterygii and (b) *Fundulus* and *Salmo* (1–5 v 6,7) triggers a significant pattern in *Xenopus* and *Latimeria*. Its observed matching rate of 59% with the outgroup (i.e. *Petromyzon* and *Lampetra*) is far below the neutral matching rate of 67%, yielding a high novelty estimate. One possible reason for such a pattern is the accumulation of convergences in *Xenopus* and *Latimeria*. On the other hand, the incorrect split separating Mammalia from the other species (1–3 v 4–7) triggers no conflict within the other species (4–7); the pattern in *Xenopus* and *Latimeria* is significant, but it features a very low novelty estimate due an excess of the observed matching rate (68%) in relation to the neutral matching rate of

Table 1. Symplesiomorphic Patterns and Corresponding Splits involving the Gnathostomata group of the Chordata dataset. Species numbers as in Fig. 7.

Pattern for the correct split 1–7 versus 8,9	Advised novelty estimate	Conflict	Incorrect split (corresponding to the pattern)	Method favoring the incorrect split	Bootstrap support
1–4,6,7	0.255	0	5 versus 1–4,6–9	parsimony	100
4–7	0.335	0.266	1–3 versus 4–9	neighbor joining	70
1–5	0.054	0	6,7 versus 1–5,8,9	likelihood	56
1–5,7	0.056	0	6 versus 1–5,7–9	–	–
1–6	0.039	0	7 versus 1–6,8–9	–	–

59%. Again, convergences in *Xenopus* and *Latimeria* may be the reason for this artifact, diminishing the neutral matching rate.

Artificial data

Current knowledge about the relationship between the process used for the generation of artificial data and the evolutionary processes that are behind natural data is very limited. Calculating mutabilities / nucleotide substitution rates from natural data is not straightforward, as exemplified by the complex procedures used by Van de Peer et al., (1996, 1997). Therefore, some intensive research is needed to find good estimates for the mutabilities that we can expect in natural data, not to mention their variation across the different branches, and we should not derive any detailed conclusions yet based on the performance of our method on artificial data, not even from the extensive comparative tests performed in Fuellen et al. (2001).

Acknowledgements

We are grateful to the Integrated Functional Genomics unit of the Interdisciplinary Center for Clinical Research (IZKF) of the University of Muenster for financial support. The first author would like to thank Dr. C. Held for valuable comments, and the Perl community for the open software used. This work was funded in part by a scholar ship from the „Deutsche Forschungsgemeinschaft – Graduiertenkolleg Strukturbildungsprozesse“.

Appendix

It is instructive to give a formal definition of the Hennigian terms (cf. Hennig 1966,Wägele 1996) used in the introduction, and throughout the paper.

We are given a tree τ, and a monophyletic group g of species. Then, a **synapomorphy, or shared novelty** n in g, is a character state n that is shared by at least two species in g, which inherit it from the last common ancestor of g. In Fig. 13, synapomorphies are displayed in panels 1 and 2. A synapomorphy is also called a **shared derived character state.**

Given a synapomorphy n, a **convergence** to n is a character state that first appeared in the last common ancestor of a group of species g' disjunct from g, was then inherited by at least one species in g', and is matching with the character state of the synapomorphy. (If g' is a singleton, the last common ancestor coincides with g'.) In Fig. 13, a convergence is shown in panel 3. A convergence is also called a **chance similarity,** or an **analogy.**

Finally, a **symplesiomorphy** is the synapomorphy of a larger group of species. Given such a monophyletic supergroup $g^+ \supsetneq g$, a **symplesiomorphy** for g in g^+ is a synapomorphy in g^+ that was inherited by at least one species in g. A **symplesiomorphy** is also known as a **leftover**, or a **shared old character state**. It is an "old" state for g and defined with respect to g^+, in which it is a new character state, a shared novelty that gives evidence of the last common ancestor of g^+. A symplesiomorphy can be found in Fig. 13, panel 4.

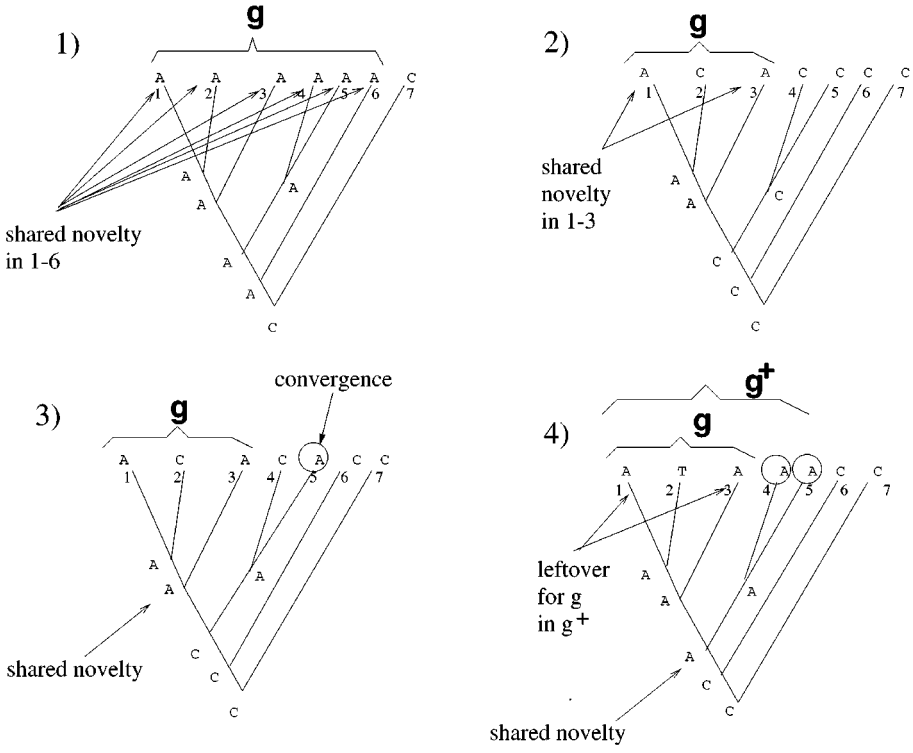

Fig. 13. Shared novelty (synapomorphy, panels 1 and 2), convergence (panel 3) and leftover (symplesiomorphy, panel 4).

References

Felsenstein, J. (1978): Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27: 401–410.

Felsenstein, J. (1993): Phylip (phylogeny inference package) version 3.5c. URL: http://evolution.genetics.washington.edu/phylip.html

Friedrich, M. & Tautz, D. (1995): Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376: 165–167.

Fuellen, G. (2000): Computing Phylogenies by Comparing Biosequences Following Principles of Traditional Systematics. Doct. dissertation, Univ. Bielefeld, Germany. URL: http://archiv.ub.uni-bielefeld.de/disshabi/2000/0026/diss.pdf

Fuellen, G., Wägele, J. & Giegerich, R. (2001): Minimum conflict – a divide-and-conquer approach to phylogeny estimation. Bioinformatics 12: in press.

Hennig, W. (1966): Phylogenetic Systematics. University of Illinois Press.

Huson, D., Nettles, S., Parida, L., Warnow, T. & Yooseph, S. (1998): The disk-covering method for tree reconstruction. Pp. 62–75 in: Battiti, R. & Bertossi, A. (eds) Proceedings of „Algorithms and Experiments“ (ALEX), Trento, Italy. URL: http://rtm.science.unitn.it/alex98/proceedings.html

Huson, D., Nettles, S. & Warnow, T. (1999): Disk-covering, a fast-converging method for phylogenetic tree reconstruction. J. Computat. Biol. 6: 369–86.

Janke, A., Gemmell, N., Feldmaier-Fuchs, G., von Haeseler, A. & Pääbo, S. (1996): The mitochondrial genome of a monotreme – the platypus (*Ornithorhynchus anatinus*). J. Mol. Ecol. 42: 153–159.

Janke, A., Xu, X. & Arnason, U. (1997): The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. Proc. Natl Acad. Sci. USA 94: 1276–1281.

Letondal, C. (2001): A web interface generator for molecular biology programs in unix. Bioinformatics 17: 73–82.

Maidak, B., Cole, J., Lilburn, T. et al. (2000): The RDP (Ribosomal Database Project) continues. Nucleic Acids Res. 28: 173–174.

Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeek, R. (1994): fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. 10: 41–48.

Penny, D. & Hasegawa, M. (1997): The platypus put in its place. Nature 387: 549–550.

Richardson, B. & Stern, W. (1997): Testing phylogenetic hypotheses using a Hennigian approach and DNA sequence data. J. Zool. Syst. Evol. Res. 35: 131–136.

Rödding, F. & Wägele, J. (1998): Origin and phylogeny of the metazoans as reconstructed with rDNA sequences. Progr. Mol. Subcell. Biol. 21: 45–70.

Salisbury, B. (1999): Strongest evidence: maximum apparent phylogenetic signal as a new cladistic optimality criterion. Cladistics 15: 137–149.

Stoye, J., Evers, D. & Meyer, F. (1998): Rose: generating sequence families. Bioinformatics 14: 157–163.

Strimmer, K. & von Haeseler, A. (1996): Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13: 964–969.

Swinscow, T. (1997): Statistics at Square One. BMJ Publishing Group; revised by M. J. Campbell. URL: http://www.bmj.com/collections/statsbk/index.shtml

Swofford, D., Olsen, G., Waddell, P. & Hillis, D. (1996): Phylogenetic infererence. Pp. 407–514 in: Hillis, D., Moritz, C. & Mable, B. (eds) Molecular Systematics. Sinauer Associates Inc., Sunderland, MA, USA.

Thompson, J., Higgins, D. & Gibson, T. (1994): CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

Tuffley, C. & Steel, M. (1997): Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59: 581–607.

Van de Peer, Y., Chapelle, S. & De Wachter, R. (1996): A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Res. 24: 3381–91.

Van de Peer, Y., Jansen, J., De Rijk, P. & De Wachter, R. (1997): Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res. 25: 111–116.

Wägele, J.-W. (1996): First principles of phylogenetic systematics, a basis for numerical methods used for morphological and molecular characters. Vie Milieu 46: 125–138.

Wilkinson, M. (1998): Split support and split conflict randomization tests in phylogenetic inference. Syst. Biol. 47: 673–695.

Woas, S. (1990): Die phylogenetischen Entwicklungslinien der höheren Oribatiden (Acari) I. Zur Monophylie der Poronota. Andrias 7: 91–168.