BIOINFORMATICS

Vol. 17 no. 12 2001
Pages 1168–1178

Minimum conflict: a divide-and-conquer approach to phylogeny estimation

Georg Fuellen[1, 3,*], Johann-Wolfgang Wägele[1] and Robert Giegerich[2]

[1]Ruhr-Universität Bochum, Lehrstuhl für spez. Zoologie, D-44780 Bochum, Germany, [2]Universität Bielefeld, Technische Fakultät, D-33594 Bielefeld, Germany and [3]Universität Münster, Integrated Functional Genomics, IZKF, D-48149 Münster, Germany

Received on December 1, 2000; revised on May 8, 2001; accepted on May 25, 2001

ABSTRACT

Motivation: Fast and reliable phylogeny estimation is rapidly gaining importance as more and more genomic sequence information is becoming available, and the study of the evolution of genes and genomes accelerates our understanding in biology and medicine alike. Branch attraction phenomena due to unequal amounts of evolutionary change in different parts of the phylogeny are one major problem for current methods, placing the species that evolved fast in one part of the phylogenetic tree, and the species that evolved slowly in the other.

Results: We describe a way to avoid the artifactual attraction of species that evolved slowly, by detecting shared old character states using a calibrated comparison with an outgroup. The corresponding focus on shared novel character states yields a fast and transparent phylogeny estimation algorithm, by application of the divide-and-conquer principle, and heuristic search: shared novelties give evidence of the exclusive common heritage (monophyly) of a subset of the species. They indicate conflict in a split of all species considered, if the split tears them apart. Only the split at the root of the phylogenetic tree cannot have such conflict. Therefore, we can work top-down, from the root to the leaves, by heuristically searching for a minimum-conflict split, and tackling the resulting two subsets in the same way.

The algorithm, called 'minimum conflict phylogeny estimation' (MCOPE), has been validated successfully using both natural and artificial data. In particular, we reanalyze published trees, yielding more plausible phylogenies, and we analyze small 'undisputed' trees on the basis of alignments considering structural homology.

Availability: MCOPE is available via http://bibiserv.techfak.uni-bielefeld.de/mcope/.

Contact: fuellen@alum.mit.edu

1 INTRODUCTION

Phylogeny estimation, that is the inference of the evolutionary history of the various life forms (species) on Earth, is a widely studied problem that is not yet solved to satisfaction (Swofford *et al.*, 1996; Wägele, 1996b). Nevertheless, due to improvements in nucleotide sequencing technology, larger and larger data sets are in need of phylogenetic analysis, featuring hundreds of species and thousands of nucleotides. In fact, whole genomes are becoming available, making an all-encompassing phylogenetic analysis possible for the first time. Whole genomes comprise huge data sets in the order of billions of nucleotides, and it would be worthwhile to align the data as far as possible, and to estimate phylogenetic trees from the data that comprise all the inheritable information of the different species. Such an analysis not only reveals insights into history, but it is crucial for our understanding of molecules, organisms and ecosystems as they are today, see e.g. Harvey *et al.* (1996)–'Nothing in biology makes sense except in the light of evolution' (Dobzhansky, 1973).

Previous work

Computer algorithms have been used extensively to approach phylogeny estimation. There are three major classes of algorithms—distance methods, parsimony, and maximum likelihood (Swofford *et al.*, 1996). All methods have found dedicated followers, and the (sometimes furious) debates on the methods' (dis)advantages clearly indicate that none of these solves the problem to satisfaction. This paper suggests a new type of method, clearly distinct from the ones mentioned. It is based on very simple principles, which have a long history in systematics; we cast them into an algorithmic format suitable for molecular data.

*To whom correspondence should be addressed.

Shared novelties

Investigating the phylogenetic relationships between m species, let an alignment of length r

$$A = \begin{array}{lclllllll} s_1 & = & s_{1,1} & s_{1,2} & \cdots & s_{1,j} & \cdots & s_{1,r} \\ s_2 & = & s_{2,1} & s_{2,2} & \cdots & s_{2,j} & \cdots & s_{2,r} \\ \cdots & & & & \cdots & & & \\ s_i & = & s_{i,1} & s_{i,2} & \cdots & s_{i,j} & \cdots & s_{i,r} \\ \cdots & & & & \cdots & & & \\ s_m & = & s_{m,1} & s_{m,2} & \cdots & s_{m,j} & \cdots & s_{m,r} \end{array}$$

be given, as a rectangular arrangement of the corresponding m biosequences. Gaps are inserted into the sequences yielding the 'padded' sequences $s_1, \ldots, s_m$, highlighting homologous sites in spite of insertions and deletions. The character states in an alignment column are supposed to have evolved from a common ancestral character state, and every column corresponds to one individual 'character'. The most natural way of analyzing such data resulting from the 'descent with modification' process of evolution is to look for the modifications. In the case of biomolecular sequences, these are modifications of character states (nucleotides) that appeared anew in an ancestral species. They give evidence of the exclusive common heritage (or, monophyly) of all the species to which that ancestral species gave rise. Consequently, our method for recovering the phylogenetic tree tries to detect novel character states shared between species because these species are *the sole descendants* of an ancestral species. In formal terms, we are given a tree $\mathcal{T}$, and a monophyletic group g of species. Then, a `shared novelty` n in g is a character state n that was inherited by at least two species in g, and first appeared as a substitution in the last common ancestor of g. In cladistic terms, a shared novelty is also called a `synapomorphy`, or a `shared derived character state` (cf. Hennig, 1966; Wägele, 1996a). If we can identify the shared novelties correctly, we can estimate the correct phylogeny instantanously.

Erosion

A major challenge for phylogeny estimation using molecular data is that shared novelties do not usually appear as insertions into the sequence, i.e. they do not usually stand out, aligning with gap characters. Instead, they are substitutions in alignment columns that already display a shared novelty that appeared earlier (i.e. a `symplesiomorphy`, also known as a `shared old, or primitive, character state`, cf. Hennig, 1966; Wägele, 1996a). Such substitutions do not always introduce random noise (random homoplasies) into the data. Instead, if many substitutions affect the same set of species, they can lead to artifacts that mislead standard phylogeny estimation methods in a systematic way. One such phenomenon, which we call 'erosion', will be described next.

Consider Figure 1, derived from the Crustacea data presented in Section 4. On the right, part of the alignment of the 18S-rDNA sequences of twelve Crustacean species is displayed. Species 12 is the `outgroup`, i.e. a species that is not among the descendants of the last common ancestor of the species 1–11 to be investigated. On the left, the putative phylogeny based on morphological features is given, and we can derive the following hypotheses. In the columns marked by arrows in green, shared novelties give evidence of monophyly of species 1–8, but hide shared novelties showing monophyly of 1–11. The consequence is an illusion of shared novelties in 9–11, triggered by shared old character states. If there are many of these, standard methods of phylogeny inference are led astray: for the full-size data set, neighbor joining, parsimony and likelihood pull 9–11 into one group. In particular, only a few character states in 12 that coincide with the shared novelties in 1–8 can render the tree with subtree 9–11 the most parsimonious one (cf. Fuellen, 2000). Such a low level of 'long-branch attraction' (cf. Felsenstein, 1978) can be avoided by investigating matching rates with the outgroup; they indicate that character states shared by 9–11 are old (symplesiomorphic), and we can then completely disregard *all* alignment columns that support an apparent monophylum 9–11. In other words, shared character states of 1–11 `eroded` away in the fast-evolving sequences of 1–8, and based on the evidence given by the matching rates with the outgroup, we ignore the shared old character states left over in 9–11 that give misleading information about the phylogeny.

2 METHOD AND ALGORITHM

Let us formalize our phylogeny estimation algorithm based on telling apart shared novelties and shared old character states by outgroup comparison. First we introduce a specific notion of consensus sequences.

Majority sequences

Given an alignment A and a group g of species for which we want to calculate the majority sequence, we employ `relative majority voting` in a column-by-column fashion. The majority character state of column j is denoted by $c_j(g)$, and it is obtained by first ordering the symbols found in column j by frequency. Then, symbols of same frequency are ordered lexicographically, and the first symbol is taken, unless it is the gap symbol. In the latter case, the majority character state is set to '!'.

Inconsistency patterns

Given a split $G = g$ versus $\overline{g} = g \vee \overline{g}$, where $g \cup \overline{g}$ covers all species currently investigated, a character state in g (or $\overline{g}$) that is part of a variable subcolumn is called an `inconsistency`, if it is matching with the majority character state of the complementary group ($\overline{g}$ or g). Given

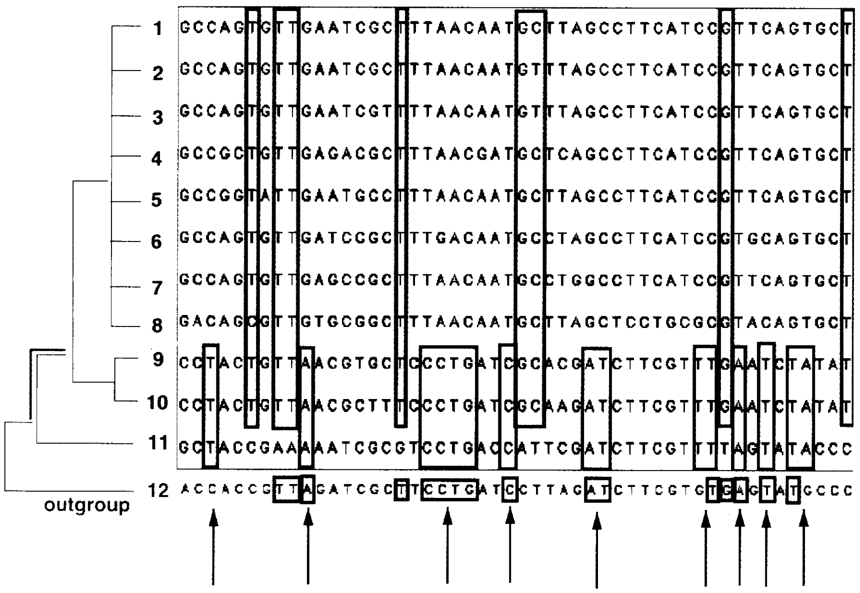
1 GCCAGTGTTGAATCGCTTTAACAATGCTTAGCCTTCATCCGTTCAGTGCT
2 GCCAGTGTTGAATCGCTTTAACAATGTTTAGCCTTCATCCGTTCAGTGCT
3 GCCAGTGTTGAATCGTTTTAACAATGTTTAGCCTTCATCCGTTCAGTGCT
4 GCCGCTGTTGAGACGCTTTAACGATGCTCAGCCTTCATCCGTTCAGTGCT
5 GCCGGTATTGAATGCCTTTAACAATGCTTAGCCTTCATCCGTTCAGTGCT
6 GCCAGTGTTGATCCGCTTTGACAATGCCTAGCCTTCATCCGTGCAGTGCT
7 GCCAGTGTTGAGCCGCTTTAACAATGCCTGGCCTTCATCCGTTCAGTGCT
8 GACAGCGTTGTGCGGCTTTAACAATGCTTAGCTCCTGCGCGTACAGTGCT
9 CCTACTGTTAACGTGCTCCCTGATCGCACGATCTTCGTTTGAATCTATAT
10 CCTACTGTTAACGCTTTCCCTGATCGCAAGATCTTCGTTTGAATCTATAT
11 GCTACCGAAAAATCGCGTCCTGACCATTCGATCTTCGTTTTAGTATACCC
12 ACCACCGTTAGATCGCGTCCTGATCCTTACATCTTCGTGTGAGTATGCCC
outgroup

Fig. 1. Outgroup comparison reveals shared old character states. On the left, a tree based on morphological data is shown (see Section 4, Crustacea data). On the right, the beginning of the alignment of the corresponding 18S rDNA is displayed. Two types of shared character states are put into boxes: type A (marked in green) is shared by 9–11 and it tends to match the outgroup 12, while type B (marked in brown) is shared by 1–10, and it does not tend to match. The conclusion is that type A is composed of shared old states, while type B contains mainly shared novelties of 1–10.

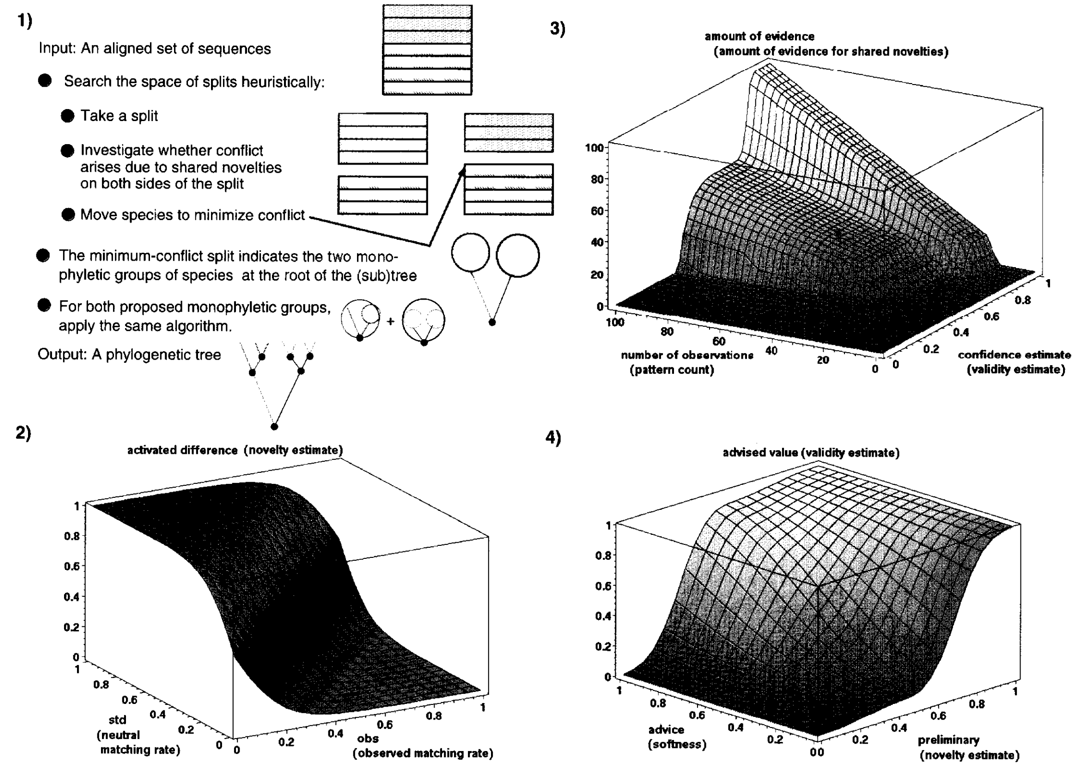

Fig. 2. Panel 1: schematic overview of minimum conflict (MCOPE, see text). Panel 2: activation of the difference between two values, std and obs. The more the former exceeds the latter, the larger the activated difference. In the case of inconsistency patterns, the variables are called as indicated in brackets. Panel 3: weighted activation of the number of observations. Only a sufficiently large number of observations that come with a sufficient confidence trigger a significant amount of evidence. In the case of inconsistency patterns, the variables are called as indicated in brackets. Panel 4: the 'advice' value influences the 'preliminary' value.

a column j, a subset $t = t(j)$ of g (or $\overline{g}$) is a `pattern` in j, if inconsistencies are found in all species of t, but in no species of its complement ($g - t$ or $\overline{g} - t$). By inspecting the entire alignment, we can prepare two lists of patterns. For each subgroup g and $\overline{g}$, $T(g)$ and $T(\overline{g})$ list the subsets of species from the subgroup that are patterns. The `pattern count` $s(t)$ of pattern t is the number of its supporting columns. These are the columns in which inconsistencies are observed exactly in the species making up t.

A split of a set of species may trigger patterns for many reasons. For example, shared novelties giving evidence of the monophyly 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 well-supported pattern in this group. Or, shared old character states left over by erosion may be torn apart. Or, convergences may induce patterns, cf. the end of Section 4.

Phylogeny estimation based on inconsistency patterns

Given a pattern, we will quantify its *conflict*, i.e. its evidence for shared novelties torn apart. Given a split, its conflict will be the maximum pattern conflict observed. Then, we can combine heuristic search and divide-and-conquer into a phylogeny estimation algorithm. Starting with any split of the set of species analyzed, conflict arises if shared novelties can be found for a subset of the species: if this subset is torn apart by the split, the shared novelties are then found on both sides, they are torn apart themselves. If a split with no (or minimum) conflict can be found by *moving high-conflict species* between the two sides, we assume that we have found the most ancient separation which does not tear apart any shared novelties. (See Fuellen, 2000, for details of the heuristic search.) The question is how the analysis can be continued. The most natural answer is to use *divide-and-conquer*.

We proceed top-down, from the root to the leaves, as follows. Given a set of species,

- find the split with minimum conflict,
- divide the set of species into two putative monophyletic groups, according to the split with minimum conflict,
- call this procedure for each group recursively as long as its size is larger than 2.

An overview of our method, thus termed 'minimum conflict phylogeny estimation' (MCOPE) is found in Figure 2, panel 1.

Sigmoid activation

To evaluate the conflict induced by a split, we will evaluate inconsistency patterns in a cascade of calculations designed to filter out those patterns due to erosion, and to keep those patterns that give evidence of the monophyly (shared novelties) of a subset of the species considered. To trigger clear decisions whenever possible, filters will be used based on sigmoid neural network activation functions.

As a first example, Figure 2, panel 2, displays the sigmoid activation of the difference of two values (std and obs), with the property that the return value is *the higher, the larger* the excess of std with respect to obs is. The formula is

$$\text{excess}_\theta(\text{std}, \text{obs}) := \frac{1}{1 + \mathrm{e}^{-(\text{std}-\text{obs})/\theta}},$$

based on the usual sigmoid neural network activation function (see e.g. Michie *et al.*, 1994). If std and obs are in the range $[0, 1]$, their difference is in $[-1, 1]$, and the smoothness of the slope of the activation is set to $\theta = 0.1$, which is our standard smoothness for this input range.

Amount of evidence

Analyzing an inconsistency pattern, we face the problem of evaluating evidence from many small observations, to which a single overall confidence value is assigned. In our case, each column that bears the inconsistency pattern is an observation. Based on all columns, we will estimate an overall confidence that the split under consideration triggers the pattern because it tears apart shared novelties. In other words, we are looking for evidence for the hypothesis that the split does *not* separate two monophyletic groups; we assume that in this case shared novelties giving evidence of monophyly are torn apart, creating recognizable patterns.

Evidence evaluation is based on the following general rules:

- evidence requires a sufficiently high overall confidence;
- evidence can only be derived from sufficiently many observations;
- if there are sufficiently many observations for which we are in doubt, we are in a plateau of partial evidence.

Comparing evidence for different hypotheses of non-monophyly derived for different splits, a single cutoff value is not appropriate. Instead, we will provide a continuous assessment of evidence based on a sigmoid formula, given the number of observations $s(t)$ of pattern t and an estimate $v(t)$ that expresses a confidence that the observations have some distinct property (i.e. that they are due to shared novelties that are torn apart). We define the `amount of evidence` as follows, given thresholds r_0 and v_0:

$$\text{evidence}^{\bowtie}_{\eta}(s(t), v(t)) := \text{excess}_{\theta_r,\eta}(r(t), r_0) \cdot r(t),$$

where

$$r(t) := \mathrm{excess}^{\bowtie}_{\theta_v,\eta}(v(t), v_0) \cdot s(t)$$

is the number of observations weighted by confidence, that is the `confidence-corrected pattern count`. As discussed below, the symbols $\bowtie$ and η indicate slight modifications of the excess formula just introduced.

For a specific instantiation of the thresholds, the function is shown in Figure 2, panel 3, and we already note that it matches our general rules. The basic idea of the formula is to suppress a small number of observations as well as any observations with low confidence, *multiplying* the number of observations $s(t)$ by the activated confidence estimate, $\mathrm{excess}^{\bowtie}_{\theta_v,\eta}(v(t), v_0)$, and then activating the resulting confidence-corrected number of observations: the more $r(t)$ exceeds the threshold count r_0, the more $r(t)$ can retain its value. If $r(t)$ is much smaller than r_0, it is squashed to zero. Such an activation reflects that a small number of observations cannot be used to reach reliable conclusions. Furthermore, taking $\mathrm{excess}^{\bowtie}_{\theta_v,\eta}(v(t), v_0)$ instead of $v(t)$ implies that strong confidence is amplified, and weak confidence is squashed. (The formula for the amount of evidence is an improvement over the formula presented in Fuellen (2000); at that time, the pattern count $s(t)$ was compared to the threshold s_0, which was based on the mean and standard deviation of the distribution of all pattern counts. Now, we compare the confidence-corrected pattern count $r(t)$ to a threshold r_0 that will be based on the mean and standard deviation of the distribution of all confidence-corrected pattern counts. In cases of massive erosion, exemplified in some of the artificial data sets presented later on, this distinction matters: using the old formula, valid patterns due to shared novelties may be mislabeled as insufficient if they are found together with an abundance of low-confidence patterns triggered by erosion. These low-confidence patterns inflate s_0 but not r_0.)

Both excess formulas are modified. On the one hand,

$$\mathrm{excess}_{\theta_r,\eta}(r(t), r_0) := \frac{1}{1 + \mathrm{e}^{-((r(t)-r_0)/\theta_r)^\eta}},$$

uses an odd integer η like 5 to squash any residual evidence rigorously for cases where there are not enough observations. On the other hand,

$$\mathrm{excess}^{\bowtie}_{\theta_v,\eta}(v(t), v_0)$$
$$:= \begin{cases} \frac{1}{1+\mathrm{e}^{-((v(t)-v_0)/\theta_v)^\eta}} & \text{if } v(t) \geqslant v_0, \\ \frac{1}{1+\mathrm{e}^{-((v(t)-v_0)/(\frac{v_0}{1-v_0}\cdot\theta_v))^\eta}} & \text{otherwise} \end{cases}$$

aligns the smoothness of the activation with the size of the interval; this is important for $v_0 \neq 0.5$. As can be seen from Figure 2, panel 3, using $v_0 = 0.75$ as the threshold, the interval [0, 0.75] is larger than [0.75, 1], and the slope of the function in [0, 0.75] needs to be smaller, if we want to cover the intervals in a symmetric way. In effect, this 'symmetric scaling' creates a plateau of partial evidence for $1/3 \leqslant v(t) < 3/4$. Again, exponent $\eta = 5$ suppresses very doubtful cases rigorously. Both $v_0 = 0.75$ and $\theta_v = 0.1$ are values found empirically. However, we will estimate r_0 from the data set, and θ_r is then given by the rule $\theta_r / r_0 = \theta_v / v_0$.

Investigating patterns

Given a split $G = g \vee \overline{g}$, an outgroup gO, and an inconsistency pattern t for this split, we investigate whether the pattern can be explained by erosion, or not. Our investigation relies on two types of matching rates with respect to the outgroup, as described in the next subsections. If a pattern cannot be explained by erosion, the validity estimate that we introduce to measure our confidence will be maximum, and we will assume that shared novelties are torn apart by the split.

Matching rates

A `matching rate` is the frequency of matching character states, defined for two disjoint groups of species I_1 and I_2 to be compared, and a set of columns J in which the comparison takes place. In formal terms,

$$m(I_1, I_2, J) := \frac{|\{j \in J : c_j(I_1) = c_j(I_2)\}|}{|J|},$$

where $c_j(I)$ is the majority character state of the species making up I, at column j.

Let an alignment of ℓ species, $A = (a_i)_{i \in \{i_1,\ldots,i_\ell\}} = a_{i_1}, \ldots, a_i, \ldots, a_{i_\ell}$, where $\{i_1, \ldots, i_\ell\} \subseteq \{1, \ldots, m\}$, of length q be given. We ignore constant columns. One example of a matching rate is the relative number of 'preserved' character states displayed. Given an outgroup gO, this `preservation rate` of species i is defined as $p(i) = m(i, gO, \{j_1, \ldots, j_q\})$.

Species softness

The first but weak criterion to detect whether a pattern is due to erosion is based on preservation rates. The `species softness` $q(t)$ of the species involved in a pattern t found in group g is given by

$$q(t) := \mathrm{excess}_\theta(\min_{i \in g-t} p(i), \max_{i \in t} p(i)).$$

The species in t are soft if the minimum preservation rate 'outside' t is larger than the maximum rate 'inside' t. If the species involved in pattern t are soft, we have a *weak* hint that *no* erosion took place, simply because erosion usually happens to the less preserved species, leaving shared old character states in the more preserved ones. However, species softness is neither a necessary nor a sufficient erosion criterion since random substitutions in individual species may overshadow any difference due to the erosive process itself.

Pattern novelty

As indicated in Figure 1, we can base a stronger erosion criterion on outgroup comparison. The basic idea is that, for a split that tears apart shared character states, inconsistency patterns matching the outgroup are likely due to erosion affecting the shared novelties of a larger group of species, and they are *not* due to shared novelties of a subset of the species investigated. However, matching with the outgroup needs to be calibrated because it also reflects the evolutionary distance between the outgroup and the group currently under consideration.

- The outgroup gO may be close, or farther away, because it may be subject to substitutions.
- Shared novelties of the group $g \cup \overline{g}$ under consideration affect the distance as well.

Therefore, we compare two matching rates. The 'neutral' matching rate used for standardization is based on the columns without the pattern considered. The comparison of this matching rate with the matching rate of the pattern-supporting columns can reveal whether the pattern is due to erosion, as long as the 'neutral' columns that do not exhibit the pattern are subject to approximately the same amount of substitution with respect to the outgroup.

For example, if the outgroup is subject to nonconvergent substitutions, both matching rates will be modified: they tend to go down by the same amount. In the case of convergent substitutions, both matching rates will tend to go up. If the current group is subject to substitutions (i.e. if shared novelties appear for $g \cup \overline{g}$), both matching rates tend to go down as well if the pattern is due to erosion. (The neutral matching rate goes down because these substitutions do not tend to match the outgroup. The observed matching rate for a pattern t goes down because these substitutions are subject to the same erosive process contributing to t.)

Given a split $G = g \vee \overline{g}$ of species $\{i_1, \ldots, i_\ell\} \subseteq \{1, \ldots, m\}$, let us assume that we have obtained the list of patterns $T(g)$ observed in group g in alignment A. We fix a `minimum column count` δ, which is the minimum number of supporting columns that a pattern needs in order to be investigated. Patterns with less support are `ignored`. Empirically, δ is taken as the base-2 logarithm of the number of variable alignment columns. This value is sufficiently low (it is 8.0 for 256 variable columns) that no relevant information should be lost. The `observed matching rate` is $m(t, gO, \mathcal{C}(t))$, where $\mathcal{C}(t)$ are the columns supporting t. The `neutral matching rate` $m(t, gO, \mathcal{C}'(t))$ is based on the `neutral columns` of t

$$\mathcal{C}'(t) := \{j \in \{j_1, \ldots, j_q\} : t(j) = \emptyset \text{ or } t(j) \text{ is ignored}\}.$$

The neutral matching rate checks outgroup matches considering the same species t as the observed matching rate, but for a different set of columns. This set $\mathcal{C}'(t)$ consists of columns featuring no pattern, and columns featuring an ignored pattern for the species in g. The former may feature only deviations in g that are not inconsistent because they do not match with the majority of the other group, and columns that are constant in g—in both cases, no pattern is observed. As the overall number of random deviations in a data set increases, so does the number of inconsistent ones, and the first subset of $\mathcal{C}'(t)$ shrinks, whereas the second subset grows.

The `novelty estimate` $n(t)$ is the excess of the neutral matching rate in comparison to the observed matching rate:

$$n(t) := \text{excess}_\theta(m(t, gO, \mathcal{C}'(t)), m(t, gO, \mathcal{C}(t))).$$

The larger the excess, the more likely *no* erosion took place. (See Figure 2, panel 2, for a plot where the activated difference is the novelty estimate.)

Considering split 1–8 v 9–11 in Figure 1, the character states of the majority sequence of 1–8 form a pattern '9, 10' in the columns marked in brown. We calculate an 'observed' matching rate of $4/8 = 0.5$ for this pattern, and a 'neutral' matching rate of $27/33 = 0.818$. An excess of the 'neutral' matching rate indicates shared novelties in 1–10. In contrast, the observed matching rate for the same pattern in the split 1–10 v 11 is $12/15 = 0.8$ (observed in the columns marked in green), and the neutral matching rate is $15/22 = 0.682$, and we conclude a case of erosion with shared old character states left over in 9–11.

Pattern validity

We combine the novelty estimate $n(t)$ of a pattern t and its species softness $q(t)$ into one `validity estimate` $v(t)$, using an *advised* function based on the excess formula, and weighting the result by one half:

$$v(t) := \frac{n(t) + \mathit{advised}_\theta(n(t), q(t))}{2}.$$

As can be seen from Figure 2, panel 4, *advised* lets the species softness $q(t)$ influence the novelty estimate $n(t)$ depending on the ambiguity of the latter. The closer $n(t)$ is to 0.5, the more advice is taken. The underlying formulas are

$$\begin{aligned}
&\omega = \text{abs}(n(t) - 0.5) + 0.5,\\
&\mathit{advised}_\theta(n(t), q(t))\\
&\qquad := \text{excess}_\theta(\omega \cdot n(t) + (1 - \omega) \cdot q(t), 0.5),
\end{aligned}$$

where ω is the weight that is given to $n(t)$ depending on its ambiguity.

Pattern reliability

We call an unignored pattern `unreliable`, if it still has too few supporting columns such that the observed matching rate may be distorted easily by very few substitutions, resulting in an incorrect validity estimate. To formalize this notion, let μ and σ be mean and standard deviation of the distribution of *all* pattern counts of inconsistency patterns in group g; for 'unignored' patterns, the validity is known, and the confidence-corrected pattern count is used. However, if a confidence-corrected pattern count is less than the minimum column count δ, we use δ, which then acts as a lower bound. If pattern counts were distributed according to a Poisson distribution with mean μ, the standard deviation of their distribution would be $\sqrt{\mu}$. Let the excess of this value with respect to the observed standard deviation, $\rho = \text{excess}_{\theta_\sigma}(\sqrt{\mu}, \sigma)$, be the `regularity` of the distribution. Regularity is close to zero if σ is very large because there are outliers, but it is close to one if the distribution is well-behaved. The slope-smoothness θ_σ of the activation is set such that $\theta_\sigma/\sqrt{\mu} = \theta_r/r_0 = \theta_v/v_0$.

If there are no outliers, we set the threshold for the pattern count r_0 to $\mu + \nu_s \cdot \sigma$, where $\nu_s = 5$ is found empirically, based on the idea that reliability should be assigned to 'significant' counts, and 'significant' can be expressed in statistical terms as the standard error of the mean, $\mu + \nu_s \cdot \sigma$. However, if there are outliers, σ can become very large, and inflate r_0. Then, we set r_0 to $\mu + \nu_S \cdot \sigma$, where ν_S is set to 1. To accommodate both cases in a smooth way, we set

$$r_0 := \rho(\mu + \nu_s \cdot \sigma) + (1 - \rho)(\mu + \nu_S \cdot \sigma).$$

Pattern conflict and split conflict

The `pattern conflict` of pattern t, also called the `amount of evidence for shared novelties` that is 'behind the conflict in the inconsistency pattern', is now given by

$$\widehat{s_v}(t) := \text{evidence}_\eta^{\bowtie}(s(t), v(t)).$$

Given a split $G = g \text{ v } \overline{g}$, the `split conflict` is the maximum pattern conflict taken over all patterns:

$$\text{conflict}(g \text{ v } \overline{g}) := \max_{t \in T(g) \cup T(\overline{g})} \widehat{s_v}(t).$$

As we have seen, a heuristic search for minimum split conflict returns the best candidate for the most ancient separation, and we can then continue our analysis viewing the two separated sets of species as new problems that can be tackled in the same way.

Outgroup maintenance

Once we have found a minimum-conflict split $G = g \text{ v } \overline{g}$ and we tackle g, we have a choice of two outgroup candidates for g: the old outgroup gO, and the sister group $\overline{g}$. It is obvious that a far-away outgroup may not be able to give sufficiently accurate matching rate information. If the outgroup is very distant, matching rates around 0.25 are observed (in the case of nucleotide sequences), and the comparison of matching rates tends to be useless. However, if the outgroup is too close, neutral matching rates close to 1 are the result, and it becomes impossible for the observed matching rate to exceed the neutral one by a sufficient amount. Furthermore, for neutral matching rates of, say, 0.9, sampling error can easily diminish the observed matching rate such that no excess is possible even though erosion took place. In other words, outgroup comparison is not very informative if the outgroup is too close, nor if it is too far away. Therefore, the `homogeneity` of the amount of deviations introduced into a new group g by comparing its species to the outgroup candidate g' is a good criterion for outgroup selection. If the alignment under consideration has variable columns J, the homogeneity is

$$p_\Delta(g, g') := 1 - (\max_{i \in g} m(i, g', J) - \min_{i \in g} m(i, g', J)).$$

Then, the outgroup of g is selected by the formula

$$O(g) = \begin{cases} \overline{g} & \text{if } p_\Delta(g, \overline{g}) \geqslant p_\Delta(g, gO), \\ gO & \text{otherwise.} \end{cases}$$

The outgroup of $\overline{g}$ is selected in an analogous way.

3 IMPLEMENTATION

MCOPE software is written using the Perl programming language (Wall *et al.*, 1996). It consists of object-oriented modules for alignment manipulation (see Chervitz *et al.*, 1999), phylogeny manipulation, phylogeny exploration and corresponding alignment visualization. The PGPLOT plotting library (Pearson, 1997) and its Perl interface (Glazebrook, 1997) are used for graphics, and a bitvector implementation (Beyer, 1998) is used for handling splits.

4 RESULTS AND DISCUSSION

Intensive validation on both natural and artificial data has been performed with good results. We have selected natural data from two sources. We reinvestigate published studies, and we assemble data sets from an alignment database. We take great care that the latter are assembled in an objective manner.

In the following, all columns with unknown nucleotides/missing data (usually coded '?' or 'N' in the alignment) are removed. If columns with gaps are not removed upfront, the suppression of 'runs' of inconsistencies found in a consecutive sequence of alignment columns is necessary, since these usually indicate no valid pattern, but deletions and sequencing gaps. Therefore, we

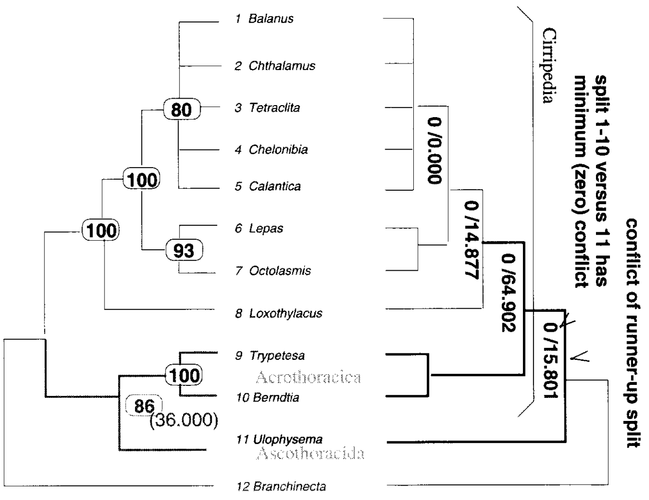

Fig. 3. Left: tree published by Spears *et al.* (1994, redrawn; numbers indicate bootstrap support). Right: minimum conflict tree for the Crustacea data set. The tree on the left is implausible as far as the monophylum 9–11 is concerned, despite high bootstrap support. The tree on the right indicates the plausible monophyla postulated from morphological data, and it displays the tree that follows from the minimum-conflict splits. Their actual conflict value is indicated, followed by the conflict of the second-best split.

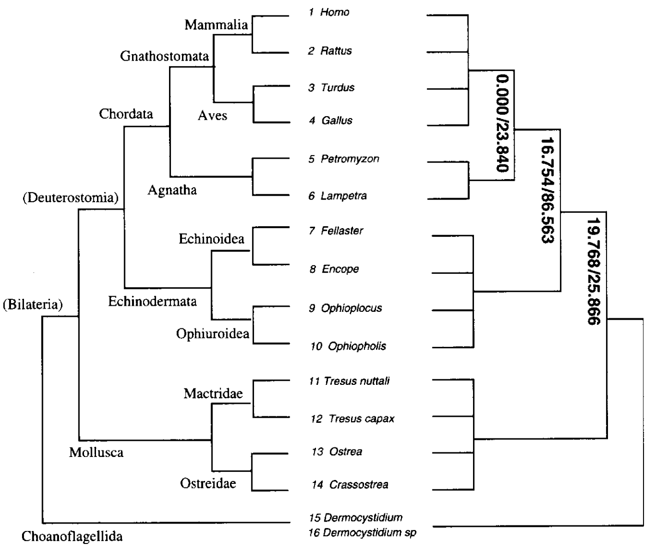

Fig. 4. 'Undisputed' tree (left) and minimum conflict tree (right) for the Bilateria data set. Labels for the minimum-conflict tree follow conventions as in Figure 3.

identify the indices involved in runs, per default defining 'consecutive' such that a run is not interrupted by a single exception. For each run, we retain only the first few columns (i.e. the first 4), and ignore the others.

Finally, a similarity measure on the set of patterns is employed to amplify the signal of weak patterns, by considering 'neighboring' ones, as described in Fuellen (2000). Omitting this extra step, tree topologies and conflict values presented in the following are essentially the same, except for the first split in Figure 4 (Bilateria data). This deep branching cannot be recovered without extra measures; alternatively to the consideration of 'neighboring' patterns, the inclusion of columns with unknown nucleotides enables us to 'see' so far.

Crustacea data

The Crustacea data set published by Spears *et al.* (1994) has been used as our running example. The data comprise 18S-rDNA from twelve species, one species (*Branchinecta*, 12, used as the outgroup) from the Branchiopoda group, and 11 species to be analyzed, from the Thecostraca group. Following morphological data, the Thecostraca split into Cirripedia and Ascothoracica, and the Cirripedia split into Acrothoracida and Thoracica. Thoracica in turn are comprised of Rhizocephala (represented by *Loxothylacus*) and Thoracica *sensu stricto*. The tree *topology* from Figure 3, right, is assumed to be correct (see Spears *et al.*, 1994; Newman, 1987; Newman *et al.*, 1969).

Figure 3, left, features the tree obtained by Spears *et al.* (1994). The authors comment their tree as follows: 'Parsimony, invariants and neighbor-joining analyses all showed the Ascothoracida and Acrothoracica to be sister taxa [···]. Although we certainly do not reject the considerable molecular data supporting a close relationship between the Acrothoracica and Ascothoracica, we suggest that the Acrothoracica diverged very early from the cirripedian lineage [···]'.

Consider again Figure 1. The character states shared between 9–11 are presumably due to erosion, since they tend to match the outgroup, as outlined towards the end of the section on pattern novelty. (In the partial alignment, we calculate their 'observed matching rate' of $12/15 = 0.8$, while the neutral rate based on 'neutral' columns evaluates to $15/22 = 0.686$. For the full-size alignment, the observed rate is $94/116 = 0.810$, and the neutral rate is $78/119 = 0.655$.) Conversely, character states shared between 1–10 do not tend to match the outgroup, indicating that these are, at least in part, shared novelties. (For the full-size alignment, the observed rate is $19/36 = 0.528$, and the neutral is $194/236 = 0.822$.) Indeed, the plausible tree featuring the Cirripedia (species 1–10) as a monophylum is clearly found by minimum conflict (Figure 3, right.) The label attached to an internal node of the minimum conflict tree lists the minimum conflict value established for this node, followed by the conflict of the second-best split, as determined by the heuristic search. In other words, the split 1–10 v 11 triggers zero conflict, followed by a split with a conflict of 15.801; this is the split 1–8, 11 v 9, 10. Furthermore, on the left of the figure, the conflict value obtained for the split featured by the doubtful tree (36.000) is noted next to its parsimony bootstrap value.

Support for the incorrect split 1–8 v 9–11 is triggered by an erosion artifact, and the split has a high bootstrap value of 86% in parsimony analysis. Let us note that high bootstrap values (see e.g. Swofford *et al.*, 1996) do not exclude systematic error—e.g. these indicate maximum support (100%) for any data if the method just builds up a caterpillar tree of the sequences in input order; resampling will yield such an artifact tree every time. Since the number of variable columns is 298 for the alignment of species 1–11, the minimum column count for the first set of conflict calculations is $\lceil \log_2 298 \rceil = 8$. The projected alignments featuring species 1–10, 1–8 and 1–7 have 267, 137 and 75 variable columns, respectively, and the minimum column counts are 8, 7 and 6. The calculation of the Crustacea minimum conflict tree then continues with an insufficient number of just 45 variable columns. The outgroup selected for 1–10 as well as for 1–8 is species 12, but for 1–7, species 8 triggers a smaller spread of matching rates, and is therefore elected as outgroup.

We obtained the minimum conflict tree just discussed ignoring those parts of the alignment featuring gaps. Using alignment columns with gaps as well, we estimate a tree which again strongly supports 1–10 v 11. Thereafter the method fails, detecting several splits with zero conflict (i.e. 1–8 v 9, 10, 1–9 v 10 and 1–8, 10 v 9)—such a polytomy indicates that gaps may be misleading for this data set.

For natural data, a comparison of our method with standard methods is given in Table 1. The *Phylip* package (Felsenstein, 1993) was used to estimate trees by maximum parsimony, neighbor joining and UPGMA, and *fastDNAml* (Olsen *et al.*, 1994) was used for maximum likelihood. *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. The Robinson–Foulds score (Robinson and Foulds, 1981, also known as the 'partition metric' dT) is used to compare trees; it is the size of the symmetric difference between the edges of the estimated tree, and of the presumably correct tree. Correct inference of the 1–10 v 11 split for the Crustacea data set is reflected by a zero dT distance in case of minimum conflict. (The phylogeny of species 1–5 is ignored because it is not known.) We also reanalyzed the Chordata part of the tree estimated for Rödding and Wägele (1998), see Fuellen (2000). Again, the minimum conflict tree is more plausible, even though not all putative correct monophyla are recovered by minimum conflict either, yielding a dT difference of 3 with the presumably correct tree.

Table 1. Performance of minimum conflict (MC), neighbor joining (NJ), UPGMA (UP), parsimony (MP) and likelihood (ML) for various kinds of natural data

Method	Published data sets		Data sets from RDP			
	Crustacea	Chordata	Bilateria	Mammalia	Gnathostomata	Tetrapoda
MC	0	3	0	4	8	0
NJ	2	5	2	0	8	0
UP	2	5	0	2	12	2
MP	2	5	0	0	6	0
ML	2	5	2	0	9	0

Performance is measured via the partition metric of Robinson and Foulds (1981). Bilateria and Crustacea data sets are described in this text. Chordata, Mammalia, Gnathostomata and Tetrapoda data sets are described in Fuellen *et al.* (2001).

Bilateria data

We have developed a procedure for the systematic construction of natural datasets where the 'true' tree topology is undisputed, selecting sequences from the alignment of the Ribosomal Database Project (RDP); Maidak *et al.*, 2000 database in a mechanical manner. The RDP alignment is guided by structural information, and the database offers a sequence query facility (the 'Phylogenetic Tree Browser') that has a crude phylogenetic organization which we can finetune. Our procedure amounts to the mechanic rule 'Always take *the first two* taxa', at each level of the finetuned phylogeny. The rule 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' (i.e. least derived) 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. The species selection process is described in detail by Fuellen, 2000.)

If we apply the strict species selection process to Bilateria taxa in the RDP database as of July 2000, we sample the species in the tree in Figure 4 on the left. Moreover, this tree is hard to dispute. The minimum conflict tree on the right recovers the phylogeny correctly, even though resolution deteriorates for the most internal nodes. The reason is a lack of variable columns—after all, we ignore all the columns that include unknown nuclotides/missing data. Minimum conflict starts off with 522 variable characters in the alignment of species 1–14,

Table 2. Performance of minimum conflict (MC), neighbor joining (NJ), UPGMA (UP), parsimony (MP) and likelihood (ML) for various kinds of artificial data

	1500 nucleotides				1000 + 500 nucleotides				1000 + 500 nucleotides			
	Jukes–Cantor, tree with constant branchlength		Jukes–Cantor, tree with variable branchlength		Jukes–Cantor with 2 PAM but				Jukes–Cantor with 2 PAM but			
	4	12	1 to 7	1 to 23	1	2	3	4	1	2	3	4
					very long branches anywhere				very long terminal branches			
MC	0.19	0.69	2.00	6.25	2.62	5.62	8.81	10.94	3.12	4.69	7.25	8.50
NJ	0	0	0	2.94	0.31	5.94	9.44	10.06	0.25	8.38	12.81	14.31
UP	1.12	0.94	4.62	5.44	4.94	7.56	10.25	10.81	5.94	10.75	13.88	15.88
MP	0	0	0	0.44	0.06	5.12	9.50	9.94	0	7.44	12.19	14.25
ML	0	0	0	0.56	0	3.12	4.69	7.00	0	2.19	4.12	6.38

Performance is measured via the partition metric of Robinson and Foulds (1981), averaged over 32 runs for each data point. Likelihood results are separated since the method has a natural advantage in case of artificial data generated using a distinct model of sequence evolution. Branch length is measured in PAM, percent accepted mutations per branch.

continues with 408 variable sites in the alignment of species 1–10, and can still resolve the correct split of 1–6, given 206 variable sites. Lists of many zero-conflict splits result for the remaining subtrees of species 1–4, 4–8 and 9–12, featuring 135, 177 and 191 variable characters, respectively.

For the Bilateria, neither neighbor joining nor likelihood reconstruct the correct tree, placing Mollusca as a sister taxon to Chordata; parsimony yields the correct tree. Three more RDP-based data sets are described in Fuellen *et al.* (2001), applying the species selection process to gnathostomatan 18S-rDNA and mammalian as well as tetrapodian 12S-rDNA. Comparative Robinson–Foulds scores for these data are given in Table 1. (For the mammalian data, MCOPE is penalized because it favors the 'Marsupionta' hypothesis (conflict 14.570), but the presumably correct split is the closest runner-up (conflict 16.260). Without erosion-corrected reliability estimation, the 'Marsupionta' split was a close runner-up to the presumably correct split, cf. Fuellen, 2000.)

Artificial data

For the generation of artificial data, we use the tool 'Rose' (Stoye *et al.*, 1998) as described in Fuellen (2000), generating trees and corresponding alignments with approx 16 species and 1500 sites. Rose allows the generation of sequences based on the Jukes–Cantor model of sequence evolution (Jukes and Cantor, 1969) including the creation of indels. In all cases, maximum likelihood is the method that performs best in recovering the tree topology from the alignment, and this is no surprise since the core of the method is the estimation of a model of sequence evolution, and this is an easy task for this kind of input. We can show that under certain conditions, minimum conflict performs superior to the other standard methods, i.e. neighbor joining, UPGMA and parsimony. In such scenarios, the set of sites is divided into two parts: a (larger) set of sites that evolve very fast in certain branches (causing the artifacts that mislead standard methods), and a set of sites that follow artificial evolution with equal branchlengths (causing shared novelties to appear and be sustained.) In fact, such a division of sites renders the artificial evolution more resemblant to the scenario known from morphological systematics: among many misleading characters, some well-supported synapomorphies can then be found.

Comparative results are presented in Table 2. Under conditions of equal branch length as well as variable branch length with no very long branches, minimum conflict performs inferior, cf. Table 2, columns 1–4. However, minimum conflict catches up if 1000 sites are evolved using a tree with between 2 and 4 very long branches featuring 128% accepted mutations (PAM) instead of 2 PAM in the other branches, and another 500 sites evolve at a constant rate of 2 PAM per branch (cf. Table 2, columns 6–8. Note that the long branches are located anywhere in the tree.) Minimum conflict is superior if there are two or more very long branches that are all terminal branches, cf. columns 10–12. If there is just one long branch, terminal or not, minimum conflict performs inferior (columns 5 and 9).

Conclusions

Reviewing weaknesses and strengths of the minimum conflict algorithm, we find two major situations where the method may be misled. (1) If the ancestor of a group of species gave rise to one fast-evolving and one slowly-evolving branch, and the split under investigation tears the slowly evolving monophylum apart, the resulting pattern will be caused by the shared novelties of the slowly evolving monophylum, as well as shared old character states due to erosion, and it is possible that the latter will dominate the former. Then, erosion is flagged and no conflict is noted even though a monophyletic group is torn apart. (2) Patterns may *not* be due to shared novelties, but instead due to convergences. Then, no erosion is flagged and conflict is noted even though the split may separate two monophyla. In both cases, the implication is that even if we can always detect erosion correctly, the estimated tree may still be incorrect. Another focus of future research is to improve the handling of gaps and of putative polytomies.

Nevertheless, minimum conflict phylogeny estimation performs well in practice. It is fast due to its divide-and-conquer approach, and many species can be handled simultaneously; no search through the space of tree topologies is necessary. Moreover, minimum conflict is transparent in the following way: for each decision taken, it clearly identifies the sites that feature putative shared novelties in conflict with the assumption of monophyly of a particular set of species, and it identifies sites with putative shared old character states. This allows the researcher to evaluate these decisions in terms of his or her own expertise.

ACKNOWLEDGEMENTS

We are grateful to the Integrated Functional Genomics unit of the Interdisciplinary Center for Clinical Research (IZKF) of the University of Münster for financial support. The first author would like to thank the Perl community for the open software used. This work was funded in part by a stipend from the 'Deutsche Forschungsgemeinschaft—Graduiertenkolleg Strukturbildungsprozesse'.

REFERENCES

Beyer,S. (1998) Bit::Vector—efficient base class implementing bit vectors. URL:www.engelschall.com/u/sb/download/Bit-Vector/.

Chervitz,S.A., Fuellen,G., Dagdigian,C., Brenner,S.E., Birney,E. and Korf,I. (1999) Bioperl: standard perl modules for bioinformatics. *BITS J.*, **1**. Article URL: www.bitsjournal.com/bioperl.html, Journal URL: www.bitsjournal.com/.

Dobzhansky,T. (1973) Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teacher*, **35**, 125–129.

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.

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

Fuellen,G., Wägele,J.W. and Giegerich,R. (2001) Best systematist practice tranferred to molecular data. *Organisms, Diversity and Evolution*, to appear. URL www.TechFak.Uni-Bielefeld.DE/~fuellen/mcpaper2.pdf.

Glazebrook,K. (1997) PGPLOT—allow subroutines in the PGPLOT graphics library to be called from Perl. URL: www.aao.gov.au/local/www/kgb/pgperl/.

Harvey,P.H., Leigh Brown,A.J., Maynard Smith,J. and Nee,S. (eds) (1996) *New Uses for New Phylogenies*. Oxford University Press, New York.

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

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.

Maidak,B.L., Cole,J.R., Lilburn,T.G., Parker,C.T., Saxman,P.R., Stredwick,J.M., Garrity,G.M., Li,B., Olsen,G.J., Pramanik,S., Schmidt,T. and Tiedje,J. (2000) The Ribosomal Database Project (RDP) continues. *Nucleic Acids Res.*, **28**, 173–174.

Michie,D., Spiegelhalter,D.J. and Taylor,C.C. (1994) *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, URL: www.amsta.leeds.ac.uk/~charles/statlog/.

Newman,W.A. (1987) Evolution of cirripedes and their major groups. In Schram,F.R. (ed.), *Crustacean Issues—Barnacle Biology*, vol 5, Balkema, Rotterdam.

Newman,W.A., Zullo,V.A. and Withers,T.H. (1969) Cirripedia. In Moore,R.C. (ed.), *Treatise on Invertebrate Paleontology Part R, Arthropoda*, vol 4, Geological Society of America, Boulder, Colorado, and the University of Kansas Press, pp. R206–R295.

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

Pearson,T. (1997) PGPLOT graphics subroutine library. URL: astro.caltech.edu/~tjp/pgplot/.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

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

Spears,T., Abele,L.G. and Applegate,M.A. (1994) Phylogenetic studies of Cirripedes and selected relatives (Thecostraca) based on 18S rDNA sequence analysis. *J. Crust. Biol.*, **14**, 641–656.

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

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) Phylogenetic infererence. In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics*. Sinauer, Sunderland, MA, USA, pp. 407–514.

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

Wägele,J.-W. (1996b) Identification of apomorphies and the role of groundpatterns in molecular systematics. *J. Zoo. Syst. Evol. Res.*, **34**, 31–39.

Wall,L., Christiansen,T. and Schwartz,L. (1996) *Programming Perl*. O'Reilly.