J. Zoo. Syst. Evol. Research **34** (1996) 31–39
©1996 Blackwell Wissenschafts-Verlag, Berlin
ISSN 0947-5745

Received on 1 December 1995

Identification of apomorphies and the role of groundpatterns in molecular systematics

J. -W. WÄGELE

Abstract

Putative apomorphic character states are the only relevant phylogenetic signal contained in sets of sequence data. Using the sequence position as a character, a way to identify putative apomorphies prior to phylogenetic analysis is proposed. It is shown that distance-matrix methods use trivial characters. The concept of the asymmetrical split is presented for determination of character polarity. It is furthermore argued that groundpatterns (node sequences) should be reconstructed prior to the study of relationships between taxa of high phylogenetic age. The 'evolutionary noise' contained in groundpatterns can be illustrated with a network of distances using a split-decomposition analysis.

Key words: molecular systematics – phylogenetic systematics – character analysis – phylogenetic signal – apomorphy – split decomposition – maximum parsimony – neighbour joining

1 Introduction

Hennigian logics (see Hennig 1950, 1966; Farris et al. 1970; Ax 1988) can be derived directly from the process of evolution: characters that appear for the first time in an ancestor species (Hennigian term for these features: autapomorphies) are inherited by descendants. Therefore, such similarities found in descendants (synapomorphies) can be traced back to an (usually unknown) ancestor, whose character pattern (= groundpattern of a monophylum) can be reconstructed. Not all similarities are autapomorphies of the ancestor species; some are older (Hennigian term: plesiomorphies) and can be traced back to a phylogenetically older ancestor. Some species may share a similarity not inherited from a common ancestor (analogy). The two classes of inherited similarities are homologies.

These simple laws are valid for any genetically fixed character. Problems arise when similarites are the product of chance or convergency (analogies). Analogies are false synapomorphies and cause a false reconstruction of groundpatterns and of phylogenetic relationships. Characters of poor information content such as numerical characters or nucleotide positions do not allow *a priori* the distinction between homology and analogy (Wägele 1995). Only the addition of a larger number of such characters to a larger unit (e.g. a gene) increases the probability that similarites are not the result of chance or convergencies but of inheritance from a common ancestor.

Sequences must also contain apomorphic and plesiomorphic states of positions. When it is stated that apomorphies do not exist in sequence analysis, something else is meant: homology of a character state cannot be identifed *a priori* without additional information. Nevertheless Hennigian logics must also be valid on the sequence level: only synapomorphic nucleotides (character states) are evidence of monophyly in a group. The quantitative dominance of other states of positions (plesiomorphies, autapomorphies, analogies) are responsible for erroneous estimation of evolutionary distances and for false tree patterns.

On the sequence level, groundpatterns are identical with node sequences (e.g. Lundberg 1972). These are of course hypothetical; their reconstructed composition depends on the terminal taxa used for the reconstruction of tree topology and the algorithms used. Node sequences can be calculated with PAUP. Node sequences differ considerably from consensus sequences: the latter contain only the commonest states of positions of terminal taxa (e.g. Day and Morris 1993).

2 Methods

For this study the following computer programs were used: PAUP (Swofford 1991) (see also Fitch 1977; Swofford and Maddison 1987; Swofford and Olsen 1990), MEGA (Kumar et al. 1993), SPLITSTREE (Huson and Wetzel 1994), and CLUSTALV (Higgins and Sharp 1988). 12SrDNA invertebrate sequences used herein are those analysed by Ballard et al. (1992). Sequences were aligned with CLUSTALV, multiple alignment was corrected by hand with the help of ESEE (Cabot and Beckenbach 1989) to increase homology (see also Wägele and Stanjek 1995).

The mathematical basis for SPLITSTREE was published and explained by Bandelt and Dress (1992). The advantage is that the diagrams do not represent fully resolved trees but show analogies, information usually lost in customary procedures. Diagrams are constructed without any assumptions concerning the mechanisms of sequence evolution. A tree is only reconstructed when all splits are pairwise compatible. Further details are explained below.

3 Identification of apomorphic character states

Assuming that homology of sequence positions is obtained by some procedure of sequence alignment, each aligned position is a putative (hypothetical) homologous character and the type of nucleotide at this position is a (putative) character state.

A single substitution or deletion is a historical event occurring in an evolving biopopulation. The event can be detected by a new character state present in descendant species. This state is an apomorphy if it is conserved in the descendants. A homologous position will then have two states: the plesiomorphic (older) and the apomorphic (new) state. A further substitution at this position (secondary substitution) will mask the apomorphy originally characteristic for members of a taxon. The plesiomorphic state of a taxon can be substituted in a similar way. This phenomenon is usually circumscribed as the 'reduction of evolutionary distance by multiple hits'; the resulting genetic distance differs from the evolutionary distance.

Assuming that it is highly probable that sequences have been aligned correctly, those positions that contain character states apomorphic for a monophylum will have two nucleotides, one for the monophylum (ingroup) and one for the remaining

species (outgroup) whenever: 1. No secondary substitutions occurred at these positions within all ingroup species; and 2. The plesiomorphic state remains conserved in all outgroup species.

These *binary* positions will support a split between ingroup and outgroup. This type of split will be named symmetrical in the following text.

Since, in natural populations, mutations can occur at any time, a subsequent substitution or deletion within a population can cause one of the following effects:

- Transformation of an apomorphic character of the ingroup into a new character that occurs only in one species or a subset of species (autapomorphic character of a terminal taxon); this state is trivial (not informative);
- Transformation of an apomorphic character of the ingroup into a seeming plesiomorphic state in one species or a subset of species ('back mutation');
- Transformation of a plesiomorphy of an outgroup species or a subset of outgroup species into a convergency (analogous to the apomorphy of the ingroup);
- Transformation of a plesiomorphy of an outgroup species or a subset of outgroup species into a (new) autapomorphy (trivial character state).

The following cases resulting from such 'noise' are important for a phylogenetic analysis: 1. Some positions are asymmetrical split-supporting positions, with one (apomorphic) state in all ingroup species and different two or three states in outgroup species; 2. Some outgroup species will have character states analogous to the ingroup species only at a few positions; 3. Some ingroup species will have positions, where a secondary substitution produced a convergency to an outgroup taxon; and 4. Some ingroup species will have autapomorphies (caused by secondary substitutions) at only few positions instead of the apomorphy of the ingroup. For these species a lower number of synapomorphies will be counted than in other species of the same ingroup.

These effects cannot be discerned with distance-matrix methods of phylogeny inference. In ideal cases, maximum parsimony methods allow the *a posteriori* identification of secondary substitions whenever reliable node sequences are reconstructed and compared with the terminal taxa of a tree topology.

It is well known that, contrary to analyses of complex morphological characters, the absence of further information content of the sequence position does not allow *a priori* discrimination between homology and analogy. However, an *a priori* analysis (without further assumptions) of the information content of sequence data is possible with the split-decomposition method developed by Bandelt and Dress (1992). Diagrams calculated with this method are based on 'clean', symmetrical split-supporting positions. Split-supporting positions become 'dirty' when one or more of the above-mentioned cases (2–4) occurs. 'Clean' symmetrical splits will show none of the cases (1–4).

Split-decomposition allows construction of networks of distances that show splits caused by both apomorphic and analogous substitutions, without discriminating between these cases. If the number of analogies resulting from 'evolutionary noise' is small, the network will approach the form of an unrooted tree. If apomorphies and analogies are similarly frequent, it will be impossible to find out which of two incompatible splits in a group of 4 taxa is, with some probability, the historically correct one. Such data sets are not informative enough for phylogenetic analyses and so the data must be discarded. If the number of synapomorphic characters is small in comparison with autapomorphies of terminal taxa and convergencies, the tree may take the shape of a bush.

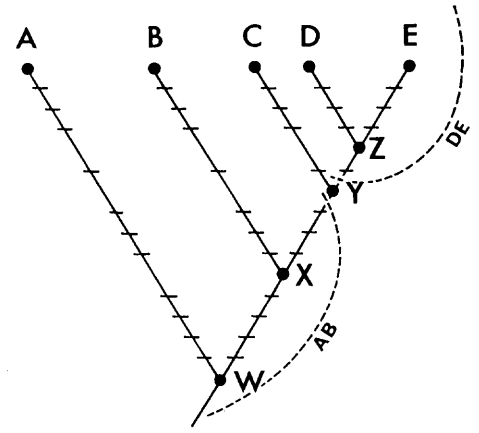
An analysis based on 'clean', symmetrical split-supporting positions will not allow polarization of the characters. In general, it is impossible to identify an apomorphic state if only character states at single sites are investigated.

To obtain more information, patterns of sites must be studied. It might be thought that such patterns are represented, for example, by pairwise distances. However, since distance data are composed of single numbers and the complexity behind these numbers is usually not analysed, it is impossible to discern *a priori* between character states. Branch lengths in rooted tree topologies calculated with character-based methods such as maximum parsimony (Kluge and Farris 1969; Farris et al. 1970; Swofford and Maddison 1987) can be traced back to substitutions that occurred, e.g. between an internal node and a terminal taxon, but only *a posteriori*, i.e. after construction of the tree. The result of character analysis depends on the tree topology. The *a priori* character analysis typical for a Hennigian (morphological) study is not possible.

In the following section, a method is proposed that allows *a priori* analysis of the information contained in sequences in a way that resembles the Hennigian procedure.

As explained above, symmetrical split-decomposition does not allow determination of the root area or of character polarity.

a



b

functional ingroup	functional outgroup	substitutions in ingroup	substitutions in outgroup
* DE /	ABC	7	29
ABC /	DE	30	6
* CDE /	AB	14	22
AB /	CDE	25	11
* BCDE /	A	25	11
ABCD /	E	33	3

Fig. 1. a. Model tree showing historical events of sequence evolution. A–E = extant species; W–Z = ancestor species. Broken lines indicate the composition of a natural ingroup (DE) and an artificial ingroup (AB); b. Table showing the number of substitutions counted within groups; * = natural ingroups. Note that substitutions occurring in a stem-line of a natural ingroup are characters of the ingroup (e.g. substitution between Y and Z is apomorphic for group DE)

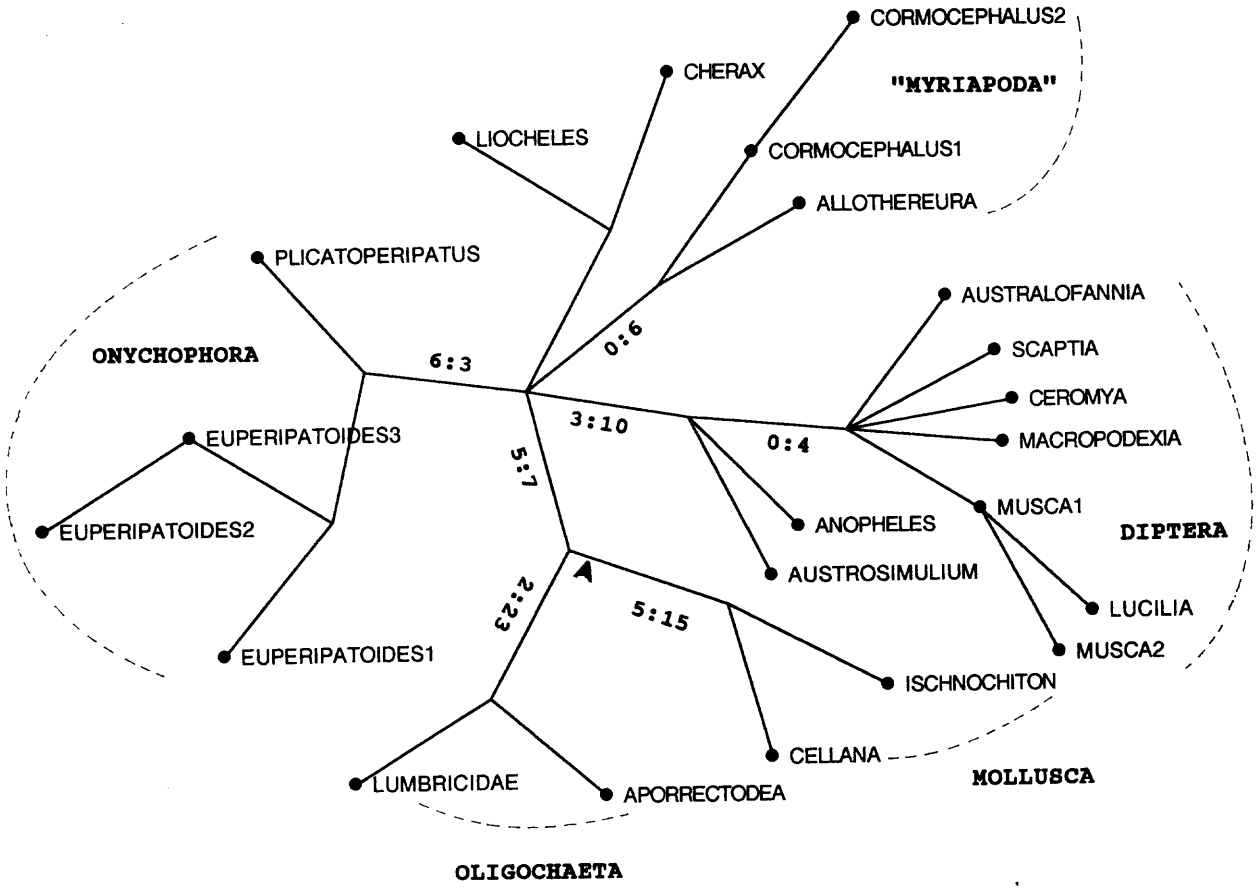


Fig. 2. Example using real data. SPLITSTREE graph (Hamming-distances) drawn with equal edges (not to scale) of 22 sequences originally used by Ballard et al. (1992). CLUSTALV alignment corrected by hand as in Wägele and Stanjek (1995). The alignment has 341 sites; of these, 82 are not parsimony-informative. Numbers on edges indicate a minimum number of identified putative apomorphies for each group (left and right side group of the edge). Only 'clean' supporting positions were counted (details in text)

But, polarity is obtained, when asymmetrical supporting positions are also considered.

If a functional ingroup X is supported by distinctly more putative apomorphic characters against an outgroup Y than vice versa (Y against X), then group X is more likely to be monophyletic than group Y, or, in other words, X : Y probably has a natural ingroup : outgroup relationship.

Why do natural ingroups (monophyla) usually have more asymmetrical supporting positions than outgroups? Consider a tree with a certain history and a number of substitutions on each branch (Fig. 1 shows a tree that evolved with a molecular clock; not too large (normal) deviations from the molecular clock model will have no influence on the results). Count the total number of substitutions that occur within a functional ingroup and a functional outgroup (Fig. 1b): closely related groups (natural ingroups) have fewer substitutions (DE and CDE in Fig. 1). When a large ingroup is compared with a single outgroup species, the natural ingroup (BCDE) still has fewer substitutions than an artificial ingroup of similar number of sequences (e.g. ABCD in Fig. 1). Now remember that substitutions supporting the historically correct splits are those occurring on the ancestor-line W–X–Y–Z of the example (Fig. 1a). These substitutions produce the only 'phylogenetic signal' useful for the identification of monophyletic groups (natural ingroups). All remaining substitutions that occur on other branches can accidentally mask this signal (by secondary substitution). The probability that secondary substitution occurs is much higher in outgroups than in ingroups due to the higher absolute number of substitutions. Therefore, positions splitting symmetrically at time t (i.e. being binary) can become asymmetrical at time t + 1. The asymmetric split-supporting positions support monophyly of the functional ingroup, where only one type of nucleotide occurs. The result of this phenomenon is that natural ingroups are supported by more asymmetric positions than outgroups. In a second step of character analysis, the symmetric split-supporting positions will be polarized in the same sense as the asymmetric positions that support the same split. The pattern that contains information about the probability that a group is monophyletic is composed of all supporting positions.

Positions supporting ingroups are putative apomorphies. The total number of putative apomorphies in relation to the number of variable positions contained in a data set is a measure of the information content of the data.

An example based on real data is seen in Figure 2. Re-aligned 12SrDNA sequences used by Ballard et al. (1992) to discuss arthropod phylogeny (see also Wägele and Wetzel 1994; Wägele and Stanjek 1995) were submitted to a symmetrical split-decomposition analysis (Bandelt and Dress 1992) with the SPLITSTREE program. The result is an unrooted

tree with a central point from which sequences of insects (here only Diptera), crustaceans, a scorpion, 'myriapods' (only Chilopoda) and lower invertebrates branch off. In a second step, positions supporting terminal groups found in this tree were polarized according to the procedure explained above. The number of putative autapomorphies is always higher for one of the two groups separated by each split. For example, Diptera against outgroup are supported by 10 : 3 positions (10 putative apomorphies for the Diptera, 3 for the remaining taxa), Oligochaeta : outgroup by 23 : 2, Euperipatoides : outgroup by 14 : 0. The number of putative autapomorphies is not high; it reflects the insecurity for hypotheses of monophyly based on these data. The area where a polarization is not obvious is the region near the root. In the example presented here, the root must be sought near the split between (Annelida + Mollusca)/ Arthropoda, but the exact position of the root can not be determined. The central point of the bush (Fig. 2) is not the root. As soon as one leaves the root area, polarization of the branches is obvious. The result is congruent with phylogeny of arthropods as reconstructed from anatomical, physiological and morphological characters: the root inserts on the line that separates molluscs from the remaining taxa (= Articulata) (arrowhead in Fig. 2). It is interesting to see that 12SrDNA indicates a rapid radiation of arthropod taxa from which the Onychophora, Chelicerata, Crustacea, Chilopoda ('Myriapoda') and Insecta originate; symmetrical split decomposition of these data does not allow resolution of the dichotomic topology of this phase of radiation because the data are not informative enough.

It must be stressed that, in this example, only 'clean' (binary) positions were counted. Taking into consideration the fact that convergences may occur within outgroups and secondary substitutions within ingroups, the number of supporting positions can be increased with further 'dirty' positions. A mathematical procedure for the *a priori* identification of a maximum number of putative apomorphies will be studied in the near future.

Counting the number of putative apomorphies, one obtains a number resembling the genetic distance used for distance-matrix methods of phylogeny inference. This number is identified prior to the phylogenetic analysis. Contrary to the genetic distance calculated for pairs of sequences the number of apomorphies represents an estimate of the genetic distance between groups of taxa; it is an estimate of the number of historical substitutions that occurred between two internal nodes. When searching apomorphies, the phylogenetic signal is filtered out from the genetic distance. The latter contains not only apomorphies but also autapomorphies and analogies, the result of 'multiple hits', therefore distance-matrix methods can correctly be labelled as being 'phenetic' in the same way as morphological methods that only compare similarities (e.g. Williams 1992).

4 Distance methods use the wrong information

Since only apomorphies are the 'phylogenetic signal' preserved in sequences, then neither trivial characters (autapomorphies of terminal taxa), that increase genetic distance when compared with plesiomorphic (conserved) character states, nor plesiomorphies should be used to calculate evolutionary distances. However, neither classical phenetic nor distance-matrix methods are capable of discerning between apomorphies and plesiomorphies.

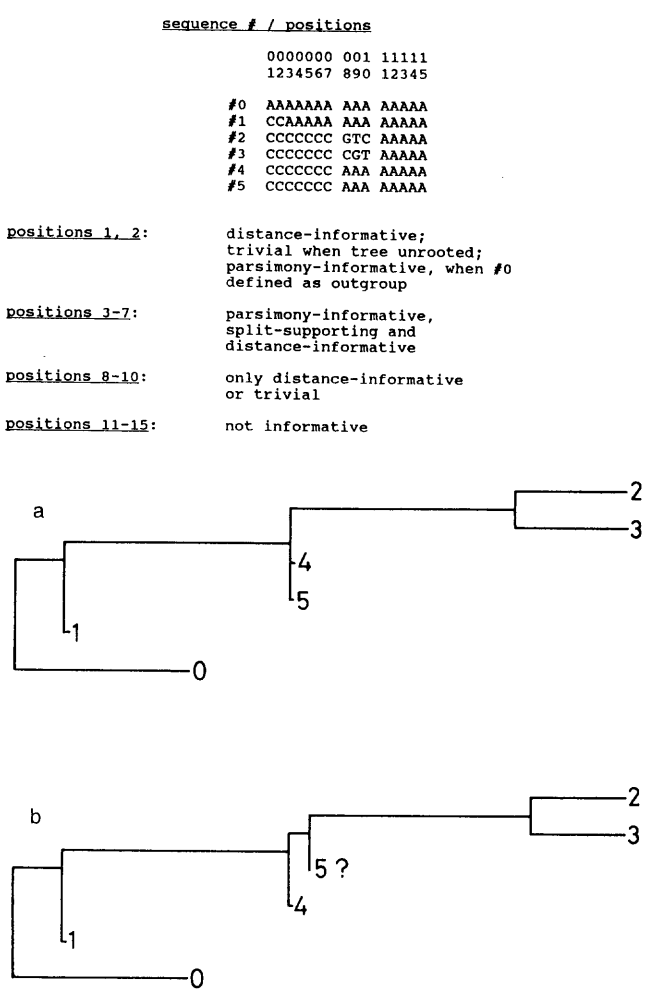


Fig. 3. Constructed evolution of a sequence, using #0 as outgroup. The information content of positions is explained. Trees were calculated with neighbour-joining algorithms (computer program MEGA) using p-distances (a) or the Jukes-Cantor parameter (b). Note wrong grouping of sequences #2 and #3, and artificial separation of #4 and #5 in b (program error)

This can easily be seen with model data (Fig. 3). The evolution constructed in Figure 3 begins with two substitutions (positions 1, 2), preserved in sequence no. 1 and descendants, followed by a further five substitutions (positions 3–7). Positions 8–10 contain, for 2 sequences, non-homologous substitutions (trivial characters) as they occur in nature in rapidly evolving species. Neighbour-joining (using MEGA) suggests a close relationship of sequences nos. 2 and 3, which is not supported by any apomorphy of the data. This erroneous tree results from the effect of trivial positions. The result is that plesiomorphies ('A' in positions 8–10) have the effect of false synapomorphies.

Maximum-parsimony methods are filters for trivial information and produce, in this case, the correct tree (0(1(2,3,4,5))).

5 Split decomposition of groundpatterns

Figure 4 shows a constructed sequence evolution with ancestor and descendant sequences. The result of a split-decomposition analysis shows a straight line connecting ancestor sequences to descendant sequences. The method produces unrooted trees. However, in contrast to diagrams calculated only from species

```
MATRIX
OUTGR       CCAAAAAAAAAA
ANC_0       AAAAAAAAAAAA
ANC_1       AAAAAAAAAATT
SPEC_1      AAAAAAAAAGTT
ANC_2       AAAAAAAAATTT
ANC_3       AAAAAAAACTTT
ANC_4       AAAAAAAATTTT
SPEC_3_1    AAAAAAAGCTTT
SPEC_3_2    AAAAAAGACTTT
SPEC_4      AAAAAAACTTTT
ANC_5       AAAAAAATTTTT
SPEC_5_1    AAAAAACTTTTT
SPEC_5_2    AAAAAGATTTTT
```

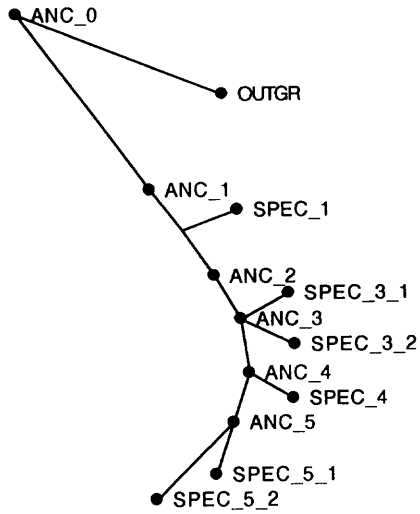


Fig. 4. Model sequence evolution (ANC = ancestor sequence, SPEC = descendant sequence) and SPLITSTREE analysis of these data (Hamming-distances, drawn to scale)

sequences (terminal taxa), this type of diagram has to be read with a strict polarity, from ancestor zero to the last (phylogenetic youngest) ancestor sequence.

Node sequences obtained from parsimony analysis of real data should behave in the same way. This was tested with an alignment of 12SrDNA sequences. These sequences were first analysed by Ballard et al. 1992, who obtained a peculiar phylogeny for arthropod groups. A CLUSTALV alignment corrected by hand to increase homology was used for maximum-parsimony (MP) analysis with PAUP. The node sequences reconstructed for one of the shortest trees were used for the following example (for more details about this data set, see Wägele and Stanjek 1995). Basal nodes of the corresponding tree (Fig. 5) represent groundpatterns of taxa with an encaptic order (Articulata (Arthropoda (Euarthropoda (Mandibulata (Tracheata (Insecta)))))). The node sequences (groundpatterns) should lie on a straight line representing the ancestor-descendant relationship between the corresponding ancestral species (see Fig. 4). SPLITSTREE analysis of these node sequences (Fig. 5) proves that the latter are not perfect groundpatterns. The diagram is a network with distances between nodes in the above-mentioned order. The largest distance separates molluscs and the Articulata. But there are additional splits, e.g. separating groundpatterns of Articulata + Arthropoda from the remaining sequences. Such splits do not of course represent phylogeny, since most of the remaining taxa are, in reality, included in the Arthropoda (except molluscs). The additional splits can have several causes: 1. Back mutations; 2. Chance similarities (analogies); or 3. Imperfect reconstruction of groundpatterns.

Factors 1 and 2 occur in nature. Factor 3 is an artefact of the method: if a taxon is represented by only few species, the reconstructed groundpattern must be very incomplete. The more species of a taxon one considers, the better the approach to the basal node (see also effect studied by Zharkikh and Li 1993). This is the reason why an increase in the number of sequences enhances the probability of obtaining the correct tree with the MP method (e.g. Nei 1991).

This effect is illustrated drastically with Figure 6: the node sequence of the ' Myriapoda' is based on two sequences of *Cormocephalus aurantipes* and one sequence of *Allothereura* sp. Addition of this node sequence to the groundpatterns of Figure 5 produces a slightly modified diagram (any addition will increase the total length of the diagram; relative distances decrease). As expected, the Myriapoda turned out to be derived from the groundpattern of the Tracheata. However, if we add only the groundpattern for two sequences of *Cormocephalus aurantipes*, the diagram (Fig. 6b) changes its shape drastically: the distance to *Cormocephalus* is much larger and there are several convergences with other groundpatterns; an obvious derivation from the Tracheata-sequence is not visible. A single fly sequence (*Musca* sp.) would even appear between the nodes for Euarthropoda and Mandibulata.

Thus, split decomposition is a useful tool for visualizing the quality of data and studying the placement of species within a system of node sequences. Pairwise percentage substitutions between higher taxa have been studied before (e.g. Carmean et al. 1992), but published histograms are not as complete and illustrative as split-networks.

It must be stressed again that the composition of groundpattern sequences depends on the species sequences used for their reconstruction. The addition of a terminal-taxon sequence to a 'groundpattern-tree' will only show its position within the parsimony tree from which the nodes were calculated. A problem of symmetrical split-decomposition analysis is that deeper ramifications become nearly invisible when a larger number of highly derived sequences are added: only the longer branches remain. This visualizes an important effect of conventional methods: if only very divergent and relatively short sequences are compared, the background noise gives false signals and the relationships are difficult to reconstruct.

Whenever groundpatterns are available for monophyla, these can be used in the same way as terminal species to reduce the long-branch problems. A goal of future sequence comparisons must be the reconstruction of groundpatterns for higher taxa, as strived for by Smith (1992) for higher taxa of the Echinoder-

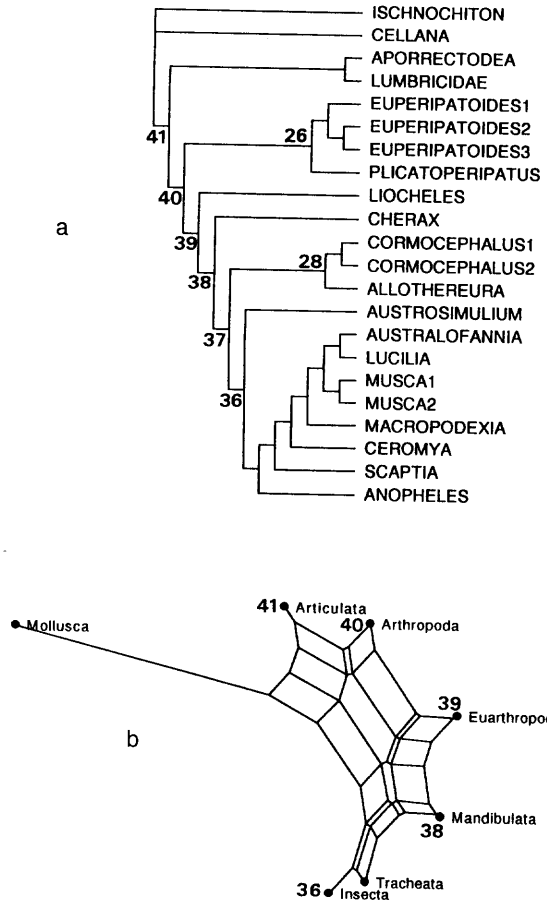


Fig. 5. a. Maximum-parsimony analysis using PAUP for the same data used for Figure 2, with the settings heuristic search, nearest-neighbour interchange branch swapping. One of six shortest trees (tree length 806); the only variations seen in other trees occur within the Diptera (after node 36). b. SPLITSTREE analysis of nodes obtained from the MP-tree seen in Figure 6 (341 sites; of these 158 constant; Hamming-distances, drawn to scale). Articulata = node 41, Arthropoda = node 40, Euarthropoda = node 39, Mandibulata = node 38, Tracheata = node 37, Insecta = node 36

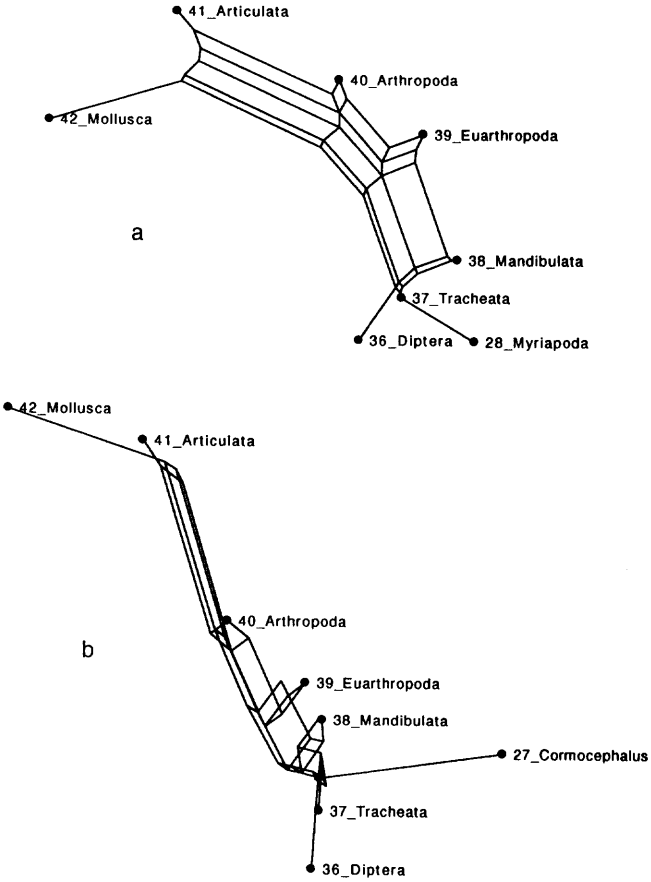


Fig. 6. a. Analysis as in Figure 5, with addition of node sequence 28 (' Myriapoda'). The node inserts correctly in the node 'Tracheata'; b. As (a), but using a terminal sequence ('*Cormocephalus*') instead of the node sequence 'Myriapoda'. Note distortions produced by convergences and wrong insertion of the node '*Cormocephalus*' between 'Mandibulata' and 'Tracheata'

mata. Groundpatterns can guarantee a more reliable study of relationships of the large groups of Metazoa.

6 Discussion

Several of the hitherto published results based on sequences are not consistent with morphological data. This does not necessarily mean that textbook knowledge on animal evolution is always obsolete or erroneous. Examples have been listed in wägele 1994, Wägele and Wetzel (1994). There are numerous other cases: Field et al. (1988), obtained in their trees of diverse metazoans annelids or brine shrimps among echinoderms and chordates; Kojima et al. (1993) found a sistergroup-relationship between annelids and molluscs; Ali et al. (1991) proposed that the Plathelminthes are closer to crustaceans than the Nematoda; Ratto and Christen (1990) had a sequence of the Bivalvia inserting between echinoderms and chordates. The following are probably the main causes of erroneous results: 1. Aligned sequences may have too few apomorphies, i.e. the data set is not informative; 2. Alignment did not establish the correct homology of several sites; 3. The method used for inferring phylogenies did not discern between apomorphies, plesiomorphies and analogies; or 4. Trivial characters affected distances.

Low information content has often been attributed to short sequences (e.g. Sourdis and Krimbas 1987; Faith 1990; Halanych 1991; Steele et al. 1991; Hillis and Huelsenbeck 1992). However, recommendations concerning minimum sequence length are not useful, since the number of informative positions contained in a data set varies with the degree of relationship of the taxa and the substitution rates. The only relevant information is whether a set of aligned sequences contains distinctly more apomorphies than analogies or not.

Since molecular data are not additive and contain analogies, evolutionary distances can not be detected precisely (e.g. Mindell 1991), they can only be estimated. As long as accumulation of secondary substitutions on a branch is slow and the number of synapomorphies high, distance methods can calculate the same result as parsimony methods. However, if the number of terminal autapomorphies is large, maximum parsimony has the advantage that a bifurcation must be supported by a more complex distribution of a character: the presence of a character state in ancestral nodes is required. This is more than just the similarity of terminal sequences, and therefore data are analysed with higher discriminatory power

(e.g. Cracraft and Helm-Bychowski 1991), since only nucleotides fulfilling the synapomorphy criterion are used.

Comparisons of the reliability of methods for estimating phylogenetic distances are usually based on computer simulations. The underlying models of sequence evolution essentially consider different probabilities for a substitution of a nucleotide by another one, depending on sequence position and rate of substitution. Several indications for the circumstances under which the methods give reliable results have been gained (e.g. Tateno et al. 1982; Sourdis and Nei 1988; Li and Nei 1990; Nei 1991; Kim 1993; Schöniger and von Haeseler 1993; Zharkikh and Li 1993). However, to quote Williams (1992): '...the underlying theoretical justification behind their [= computer programs] use is often not fully discussed...'. Sourdis and Nei (1988), for example, stated (p. 298) that 'the relatively poor performance of the MP method for these cases is due to the fact that information from singular sites is not used in this method'. However, to use all sites for the justification of a bifurcation means to include the wrong information (analogies, plesiomorphies, autapomorphies of terminal taxa). Hillis et al. (1993) and Huelsenbeck and Hillis (1993) warned that simulations might often just yield the predictions of the model used (see also critique in Cracraft and Helm-Bychowski 1991). There are no universal recommendations of methods for the applications in the real world (e.g. Schöniger and von Haeseler 1993). The choice of distance parameters depends on the substitution rate within different populations at certain historical times, and the bias caused by stabilizing selection on certain sequence sites (e.g. Eigen and Winkler-Oswatitsch 1990). To find out if synapomorphies could be masked by secondary substitutions, knowledge about substitution rates and time of divergence is necessary, the 'biology of the sequences' must be known *a priori* (Yang 1994). In practice these details are not known.

The problems caused by secondary substitutions also affect parsimony methods and distance methods (e.g. Fitch 1977; Farris 1986; Nei 1987; Felsenstein 1988; Hein 1989; Li and Nei 1990; Li and Graur 1991; Williams 1992; Zharkikh and Li 1992; Schöniger and von Haeseler 1993). Examples of the failure of MP are well known. When, as a result of different rates of substitution, the number of analogies (convergencies) is high, the 'long branches attract' (e.g. Fitch 1977; Felsenstein 1978; Lake 1987; Cedergren et al. 1988; Li and Gouy 1991). The phenomenon is easily explained: the number of convergencies with synapomorphy-like distribution is in these cases higher than the number of true synapomorphies. The same constellation occurs when an 'interior' branch separating two basal nodes from which long branches stem is very short; even when substitution rates are equal, the long branches will attract (case studied by Takezaki and Nei 1994). Recently, Yang (1994) listed further disadvantages of the MP method.

The law behind all failures of the MP method is simple: as soon as autapomorphies of a monophylum become undetectable due to secondary substitutions in the corresponding sequence positions or due to many convergencies in outgroups, phylogeny cannot be reconstructed correctly. Symmetrical split decomposition can help to decide whether the number of incompatible sites is high, meaning that the data set is not informative enough. Then phylogeny can not be reconstructed with any method using that data set because the phylogenetic signal is not preserved. Analysis of asymmetrical splits will allow estimation of which functional ingroups are well supported by the data, but further work is still necessary to find algorithms that allow calculation of probabilities.

7 Conclusions

Several of the following conclusions are partly scattered over published literature. They are repeated here for the sake of completeness together with new ideas:

- The logical rules for phylogeny reconstruction discovered by Hennig (1950, 1966) are also valid for analysis of sequences.
- The 'phylogenetic signal' in a data set is the apomorphy. Bifurcations of a phylogeny can be reconstructed as long as states of sequence poisitions that fulfill the synapomorphy condition dominate quantitatively over analogies.
- No method should be able to reconstruct phylogeny when the information (synapomorphies) is not present in the sequences. High substitution rates or long evolutionary time can have the effect that the information is lost by secondary substitution of synapomorphic sites.
- The probability that synapomorphies are present in sequences increases with the length of the sequence and decreases with increasing phylogenetic distance of the taxa.
- Apomorphies and analogies are parsimony-informative characters. In terms of the effect of methodology, conventional methods do not distinguish between phylogenetic information (apomorphies) and analogies. The effect of analogies can only be neutralized with distance parameters when the history of the sequences is known or guessed correctly. Parsimony methods allow identification and summing of putative apomorphies that support a bifurcation, but only *a posteriori*. Distance-matrix methods also use trivial characters (autapomorphies) — an important source of mistakes.
- Symmetrical split decomposition allows *a priori* illustration of the network of incompatible distances produced by apomorphies vs. analogies. Data sets with low phylogenetic signal can be discarded prior to phylogenetic analysis.
- Putative apomorphies can be identified *a priori* with patterns of sequence positions that support asymmetrical splits. An estimation of the information content of the data set is possible.
- Deep ramifications of a tree (= phylogenetically old bifurcations) are discovered with more reliability when groundpatterns of terminal taxa are considered instead of the terminal taxa alone, and thus 'long-branch problems' can be reduced.
- The quality of a reconstructed groundpattern of a monophylum increases with the number of species considered. The evolutionary noise contained in groundpatterns can be visualized with a split-decomposition analysis.

Acknowledgements

The author is grateful to Prof. Dr A. Dress (Faculty of Mathematics, Universität Bielefeld) for an introduction to the theory of the split decomposition method.

Zusammenfassung

Identifikation von Apomorphien und die Bedeutung von Grundmustern in der molekularen Systematik

Allein potentielle Apomorphien sind das 'phylogenetische Signal', das in Sequenzdaten enthalten ist. Es wird ein Verfahren vorgeschlagen, daß die *a priori*-Identifikation von potentiellen Apomorphien ermöglicht, wobei die Sequenzposition als Merkmal genutzt wird. Es wird aufgezeigt, daß Distanzmatrix-Methoden triviale Sequenzpositionen verwenden. Das Konzept des asymmetrischen Splits wird vorgestellt. Grundmuster ('Knotensequenzen') sollten rekonstruiert werden, ehe Taxa höheren phylogenetischen Alters verglichen werden. Das 'Rauschen' in den Grundmustern kann mit der Split-Zerlegung in Form eines Distanz-Netzwerks dargestellt werden.

References

Ali, P.O.; Simpson, A.J.G.; Allen, R, Waters, A.P.; Humphries, C.J.; Johnston, D.A.; Rollinson, D., 1991: Sequence of a small subunit rRNA gene of *Schistosoma mansoni* and its use in phylogenetic analysis. Mol. Biochem. Parasitol. **46,** 201–208.

Ax, P., 1988: Systematik in der Biologie. Stuttgart: UTB–G. Fischer.

Ballard, J.W.; Olsen, G.J.; Faith, D.P.; Odgers, W.A.; Rowell, D.M.; Atkinson, P.W., 1992: Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods. Science **258,** 1345–1348.

Bandelt, H.J.; Dress, A.W., 1992: Split decomposition: A new and useful approach to phylogenetic analysis of distance data. Molec. Phylog. Evol. **1,** 242–252.

Cabot, E.L.; Beckenbach, A.T., 1989: Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. CABIOS. Appl. Notes Biosci. **5,** 233–234.

Carmean, D.; Kimsey, L.S.; Berbee, M.L., 1992: 18S DNA sequences and the holometabolous insects. Molec. Phylog. Evol. **1,** 270–278.

Cedergren, R.; Gray, M.W.; Abel, Y.; Sankoff, D., 1988: The evolutionary relationships among known life forms. J. Mol. Evol. **28,** 98–112.

Cracraft, J.; Helm-Bychowski, K., 1991: Parsimony and phylogenetic inference using DNA sequences: some methodological strategies In: Miyamoto, M. M.; Cracraft, J. (eds), Phylogenetic Analysis of DNA Sequences. New York: Oxford Univ. Press. pp. 184–220.

Day, W.H.; McMorris, F.R., 1993: Discovering consensus molecular sequences In: Opitz, O.; Lausen, B.; Klar, R. (eds), Information and Classification. Berlin: Springer Verlag. pp. 393–402.

Eigen, M.; Winkler-Oswatitsch, R., 1990: Statistical geometry on sequence space. Methods Enzymol. **183,** 505–530.

Faith, D.P., 1990: Chance marsupial relationships. Nature **345,** 393–394.

Farris, J.S., 1986: On the boundaries of phylogenetic systematics. Cladistics **2,** 14–27.

–, Kluge, A.G.; Eckardt, M.J., 1970: A numerical approach to phylogenetic systematics. Syst. Zool. **19,** 172–189.

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

–, 1988: Phylogenies from molecular sequences: inference and reliability. Ann. Rev. Genet. **22,** 521–565.

Field, K.G.; Olsen, G.J.; Lane, D.J., 1988: Molecular phylogeny of the animal kingdom. Science **239,** 748–752.

Fitch, W.M., 1977: The phyletic interpretation of macromolecular sequence information: simple methods. In: Hecht, M.K.; Goody, P.C.; Hecht, B.M. (eds), Major Patterns in Vertebrate Evolution. New York: Plenum Press. pp. 169–204.

Halanych, K.M., 1991: 5s ribosomal RNA sequences inappropriate for phylogenetic reconstruction. Mol. Biol. Evol. **8,** 249–253.

Hein, J., 1989: A tree reconstruction method that is economical in the number of pairwise comparisons used. Mol. Biol. Evol. **6,** 669–684.

Hennig, W., 1950: Grundzüge einer Theorie der phylogenetischen Systematik. Berlin: Deutscher Zentralverlag.

–, 1966: Phylogenetic Systematics. Urbana: University of Illinois Press.

Higgins, D.G.; Sharp, P.M., 1988: CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73,** 237–244.

Hillis, D.M.; Bull, J.J.; White, M.E.; Badgett, M.R.; Molineux, I.J., 1993: Experimental approaches to phylogenetic analysis. Syst. Biol. **42,** 90–92.

–, Huelsenbeck, J.P., 1992: Signal, noise, and reliability in molecular phylogenetic analyses. J. Heredity **83,** 189–195.

Huelsenbeck, J.P.; Hillis, D.M., 1993: Success of phylogenetic methods in the four-taxon case. Syst. Biol. **42,** 247–264.

Huson, D; Wetzel, R., 1994: SlitsTree, a MacIntosh-application for analysing and visualizing sequence data. Bielefeld: Shareware, FSP Mathematisierung, University of Bielefeld.

Kim, J., 1993: Improving the accuracy of phylogenetic estimation by combining different methods. Syst. Biol. **42,** 331–340.

Kluge, A.G.; Farris, J.S., 1969: Quantitative phyletics and the evolution of anurans. Syst. Zool. **18,** 1–32.

Kojima, S.; Hashimoto, T.; Hasegawa, M.; Murata, S.; Ohta, S.; Seki, H.; Okada, N., 1993: Close phylogenetic relationship between Vestimentifera (tube worms) and Annelida revealed by the amino acid sequence of elongation factor 1a. J. Mol. Evol. **37,** 66–70.

Kumar, S.; Tamura, K.; Nei, M., 1993: MEGA: Molecular Evolutionary Genetics Analysis, Vers. 1.0. University Park: Pennsylvania State University.

Lake, J.A., 1987: A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol. Biol. Evol. **4,** 167–191.

Li, J.; Nei, M., 1990: Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7,** 82–102.

Li, W.H.; Gouy, M., 1991: Statistical methods for testing molecular phylogenies In: Miyamoto, M. M.; Cracraft, J. (eds), Phylogenetic Analysis of DNA Sequences. New York: OUP. pp. 249–277.

–, Graur, D., 1991: Fundamentals of Molecular Evolution. Sunderland: Sinauer Assoc.

Lundberg, J.G., 1972: Wagner networks and ancestors. Syst. Zool. **21,** 398–413.

Mindell, D.P., 1991: Aligning DNA sequences: homology and phylogenetic weighting. In: Miyamoto, M. M.; Cracraft, J. (eds), Phylogenetic Analysis of DNA Sequences. New York: OUP. pp. 73–89.

Nei, M., 1987: Molecular Evolutionary Genetics. New York: Columbia University Press.

–, 1991: Relative efficiencies of different tree-making methods for molecular data In: Miyamoto, M. M.; Cracraft, J. (eds), Phylogenetic Analysis of DNA Sequences. New York: OUP. pp. 90–128.

Ratto, A.; Christen, R., 1990: Phylogénie moléculaire des échinodermes déduite de séquences partielles des ARN ribosomiques 28S. C.R. Acad. Sci. Paris **310,** 169–173.

Schöniger, M.; von Haeseler, A., 1993: A simple method to improve the reliability of tree reconstructions. Mol. Biol. Evol. **10,** 471–483.

Smith, A.B., 1992: Echinoderm phylogeny: morphology and molecules approach accord. Trends Ecol. Evol. **7,** 224–229.

Sourdis, J.; Krimbas, C., 1987: Accuracy of phylogenetic trees estimated from DNA sequence data. Mol. Biol. Evol. **4,** 159–166.

–, Nei, M., 1988: Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5,** 298–311.

Steele, K.P.; Holsinger, K.E.; Jansen, R.K.; Taylor, D.W., 1991: Assessing the reliability of 5S RNA sequence data for phylogenetic analysis in green plants. Mol. Biol. Evol. **8,** 240–248.

Swofford, D.L., 1991: PAUP: Phylogenetic Analysis Using Parsimony, Vers. 3.1. Champaign: Illinois Natural History Survey.

–, Maddison, W.P., 1987: Reconstructing ancestral character states under Wagner Parsimony. Mathem. Biosci. **87,** 199–229.

–, Olsen, G.J., 1990: Phylogeny reconstruction In: Hillis, D.M.; Moritz, C. (eds), Molecular Systematics. Sunderland: Sinauer Assoc. pp. 411–501.

Takezaki, N.; Nei, M., 1994: Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. J. Mol. Evol. **39,** 210–218.

Tateno, Y.; Nei, M.; Tajima, F., 1982: Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. **18,** 387–404.

Wägele, J.W., 1994: Review of methodological problems of 'computer

cladistics' exemplified with a case study on isopod phylogeny (Crustacea: Isopoda). Z. zool. Syst. Evolut. -forsch. **32,** 81–107.

–, 1995: On the information content of characters in comparative morphology and molecular systematics. J. Zoo. Syst. Evol. Research **33,** 42–46.

–, Stanjek, G., 1995: arthropod phylogeny inferred from partial 12srrna revisited: monophyly of the tracheata depends on sequence alignment. J. Zoo. Syst. Evol. Research **33,** 75–80.

–, Wetzel, R., 1994: Nucleic acid sequence data are not per se reliable for inference of phylogenies. J. Nat. Hist. **28,** 749–761.

Williams, D.M., 1992: DNA analysis: theory, methods In: Forey, P.L.; Humphries, C.J.; Kitching, I.J.; Scotland, R.W.; Siebert, D.J.; Williams, D.M. (eds), Cladistics — a Practical Course in Systematics. Oxford: Clarendon Press. pp. 89–123.

Yang, Z., 1994: Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39,** 105–111.

Zharkikh, A.; Li, W.H., 1992: Statistical properties of boostrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. J. Mol. Evol. **35,** 356–366.

–, –, 1993: Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock. Syst. Biol. **42,** 113–125.

Author's address: Johann-Wolfgang Wägele, Fakultät für Biologie, Universität Bielefeld, POB 100131, D-33501 Bielefeld, Germany