Optimization of character coding and a stepwise execution of cladistic analyses

J. C. von Vaupel Klein

Vaupel Klein, J.C. von. Optimization of character coding and a stepwise execution of cladistic analyses. Zool. Med. Leiden 83 (20), 9.vii.2009: 741-758, figs 1-3.— ISSN 0024-0672.
J.C. von Vaupel Klein, Division of Systematic Zoology at Leiden University [retired; current postal address: Beetslaan 32, NL-3723 DX Bilthoven, The Netherlands] (j.c.von.vaupel.klein@biology.leidenuniv.nl / jcvvk@xs4all.nl).

Key words: a/p coding; absence/presence coding; binarization; character coding; character linkage; cladistic analysis; datamatrix reduction; homoplasy bias; optimization; phylogeny reconstruction; redundancy; stepwise analysis; successive outgroup comparison.
The absence/presence routine of character coding is examined in regard to minimizing its inherent redundancy effects with the purpose of optimizing the structuring of a comprehensive, binary datamatrix. Cladistic analytical procedures are next evaluated with respect to the successive use of such a datamatrix at different hierarchical levels. It is concluded that performing a stepwise analysis has various advantages over the more often employed techniques, i.e., the 'total analysis' routines and the 'partitioned' approaches.

Introduction

Reconstructing the historical course of evolution makes a grand goal but is by no means an easy task, as systematists are well aware. Since Hennig's (1950) landmark study of phylogenetic systematics became available in English (Hennig, 1966), systematics has earned its place among the natural sciences by adopting a strict and objectively verifiable protocol for reconstructing The Natural System. In the first 30-odd years following that date, i.e., well into the 1990s, the methods of analysing a datamatrix and constructing cladograms constituted the main focus of phylogenetics, and optimization of the various routines was vigorously pursued (*e.g.*, Felsenstein, 1982; Huelsenbeck & Hillis, 1993). Although this part of cladistic pattern recognition has by no means been tried exhaustively yet, we may nonetheless note a shift in focus towards the data that form the basis of that matrix, i.e., the proper way of coding the characters recognized, from the early 1990s (*e.g.*, Hauser & Presch, 1991; Wilkinson, 1992, 1995; Slowinski, 1993; Meier, 1994) until the present (*e.g.*, Goloboff et al., 2006; Lawing et al., 2008). In particular the First Biennial International Conference of The Systematics Association held at Oxford in 1997 (Scotland & Pennington, 2000) may be acknowledged for having truly boosted theoretical developments in the realm of character coding. However, also in this field no consensus has been reached so far and consequently no dominant approach has emerged yet. Nevertheless, both a proper way of representing character states in a primary datamatrix and a proper analysis of the structure of that matrix constitute crucial stages in the analytical procedure. Only if those stages can be optimized, we may expect an optimal use of the information contained in the distribution of the character states over the taxa as well as an optimal representation of that information in the eventual cladogram. Hence, the absence of reliable, universal formats for these operations still comprises an impediment to producing objective, reproducible results in cladistics, which obviously

detracts from the confidence we may have in the application of its fundamental objectivity. Thus, in the present paper I want to discuss an aspect of character coding, i.e., alleged redundancy, that might be solved or at least mitigated through a stepwise execution of the analytical and synthetical phases employed in finding the (natural) order hidden in the datamatrix.

Coding and use of characters in the analytical sequence

The phylogenetic method of historical pattern analysis has been in use for over 40 years now, and many results have been published. Yet, there are no signs of a general agreement among systematists with regard to the detailed implementation of the various steps in the analytical procedure. As a consequence, it is necessary to first recite the sequence at issue in order to make clear precisely in which stages the points here raised are supposed to fit.

The key word in taxonomy is variation, i.e., variation among the members of natural, monophyletic groups and conceived in any applicable sense, morphological or otherwise. Following careful analysis of the taxa at issue with regard to the variation observed, the attributes or features found, whether morphological, physiological, ecological, ethological, genetic, chemical, molecular, or otherwise, are examined in order to recognize meaningful characters that can be used in phylogeny reconstruction. In this first truly taxonomic step, allegedly homologous, individual features are aggregated to form characters, which are defined as sets of character states that are linked through *a priori* hypotheses of homology. The compilation of a datamatrix of use in the construction of a hierarchical scheme, to be interpreted as describing the historical pathways along successive speciation events, next requires that the characters and their states be properly coded. In doing so, the character states are aligned into transformation series, the binary characters are defined *per se*, and the multistate characters can either be used as they are (but see below), or be broken down to series of binary characters. What follows is the analytical procedure *sensu stricto* in which the resulting transformation series with more than two elements are, at some stage, ordered by acknowledging the significance of a certain sequence, *e.g.*, 0-1-2 should indeed be placed in that order, whether as 0-1-2 or as 2-1-0, while binary characters were already ordered by definition, as these can only yield 0-1 or 1-0. In a subsequent stage the series are also polarized through outgroup comparison, which recognizes the character states comprised as either plesiomorphous, or apomorphous at a certain level. Based on the distributions of their apomorphous character states, the taxa are next clustered hierarchically onto the branches of an essentially dichotomous cladogram that takes into account their connection with other taxa in series of sister group relationships, which are based on the synapomorphous possession of character states. The final, *a posteriori* step in the procedure then involves recognizing true homologies on the one hand, and relegating the *a priori* hypotheses apparently describing non-homologous character states to *ad hoc* statements of homoplasy or character reversal, on the other. After the analytic and synthetic routines have been completed, the ultimate stage encompasses interpreting the cladogram as (an approximation of) a historically correct phylogenetic tree, at least through the addition of a time scale on the vertical, *y*-axis and, possibly but not necessarily, by plotting some measure of similarity on the horizontal, *x*-coordinate.

We all know this sequence, and it has been adequately described in textbooks of phylogenetic inference, like Wiley (1981), Forey et al. (1992), and Kitching et al. (1998) to name only the more prominent. Yet, I recite this procedure here in full both as a reminder of all the steps involved, and in order to expose the vulnerability of any cladistic analysis. Each of those successive steps in the analytical sequence, namely, is equally crucial and an improper execution of any step will introduce information that is bound to be erroneous to some unkown extent, hence flawing the finally resulting cladogram to an unpredictable degree and in an equally unpredictable direction (or directions). Indeed, most steps can be approached *via* more than one method, and in by far the majority of cases the choice of an alternative method for any step will yield an incongruent cladogram, hence a different hypothesis about the true historical course of evolution. Thus, to determine which method may be considered the 'best' in every stage, is a matter that has to be taken most seriously.

Binary *versus* multistate characters

The primary issue to be discussed herein, is how to incorporate multistate characters into a datamatrix in a maximally pure and unbiased way. This concerns all three kinds of multistate characters, i.e., the continuous characters, the meristic ones, and the so-called classes. The two former categories, continuous and meristic, can also be characterized as quantitative, whereas the classes, just like the purely binary characters, may be acknowledged as being qualitative in nature.

Those truly binary characters are not at issue here: they comprise features that can be described in full by only two states, i.e., 0 and 1, like the presence or absence of an attribute, *e.g.*, an external shell, or else the straight or twisted structure of, *e.g.*, a spine. Data like these may be incorporated in the primary matrix as such, by simply coding '0' for absent and '1' for present, or, *e.g.*, '0' for not twisted and '1' for twisted. On the contrary, what concerns us here are characters described as so-called multistate classes, i.e., characters with states to which values deviating from 0 or 1 can be attributed in a, often well considered but nonetheless arbitrary, way.

In this respect, it is evident that in a cladistic analysis, such in contrast to the situation in a phenetic approach, continuous or meristic characters usually cannot be included in the datamatrix as they are: they have to be converted to 'classes' first. In the case of continuous characters, like the length of a wing varying, *e.g.*, between 1.31 and 3.09 mm, the states in the resulting transformation series will have to be defined as falling into, *e.g.*, classes 0 = wings absent, 1 = 1.00-1.99 mm, 2 = 2.00-2.99 mm, and 3 = 3.00-3.99 mm, or, of course, any alternative scheme that would be more relevant in the case at issue.

Where meristic characters are concerned, like the number of spines on a given part of the body, varying, *e.g.*, between 0 and 140, classes could be defined in a similar way by (at least to some degree arbitrarily) dividing the range of 0-140 into a relevant number of partitions, like: 0, 1-35, 36-70, 71-105, 106-140; or by any other scheme that may be interpreted as adequately representing the variation observed. The classes so recognized may then be coded as, *e.g.*, 0, 1, 2, 3, 4. A comparable yet fundamentally different situation may be recognized in a case of, *e.g.*, a small number of spines with discrete positions: if 1-4 spines are present, i.e., on positions I-IV, and if cases of '3 spines in total' may be distinguished in the absence of a fourth spine either in position no. II, or no. III,

it is more relevant to code: spine in position II absent or present = 0 - 1, spine in position III absent or present = 0 - 1, etc., rather than simply resorting to 1 spine = 1, 2 spines = 2, 3 spines = 3, and 4 spines = 4, since available information about the configuration of the spines would then be lost. This may immediately be apparent from the observations (see table 1).

Table 1. Example of two taxa with an equal number of spines that are, however in part present at different positions

Taxon / Position of spine	I	II	III	IV
Taxon A, 3 spines present	s	-	s	s
Taxon B, 3 spines present	s	s	-	s

The final category of multistate characters is that in which classes are inevitable from the start, as the various states cannot otherwise be represented in a numerical way. For instance, when colour is used as a character, this may, *e.g.*, comprise the colours red, yellow, and blue, which can be represented as classes through coding as, *e.g.*, red = 1, yellow = 2, blue = 3 (and 0 may be used for, *e.g.*, 'other colour').

It may thus be evident from the above, that all basically non-binary characters in a cladistic analysis will have to be represented somehow as classes, whether in a primary or in a secondary sense.

Binary representation of multistate characters

The justification, or even necessity, of representing multistate characters in discrete units already emerges from the basic principle of phylogenetic systematics as formulated by Hennig (1966) himself. Character states can be either plesiomorphous or apomorphous, but nothing in between: they can have no such status as 'largely apomorphous' or 'for 0.33 plesiomorphous' or anything like it. Character states will invariably have to be recognized as either plesiomorphous, or apomorphous (at least at the hierarchical level at issue), implying that only two possible ratings could be assigned to any character state. In binary characters these are either 0 or 1, but in transformation series with more than two elements there is the obvious possibility of assigning, next to 0 and 1, also states coded as 2, 3, 4, etc. It is here that, in a later stage, the ordering as referred to above is to be applied.

Thus, although in linear or branched transformation series of three and more elements any element can only be apomorphous at a single level, the possibility of coding states with values above 1 remains intact. Yet, high values may flaw the analytical procedure to some degree by gaining disproportionate preponderance in comparison with characters that have values limited to 0 and 1. In addition, any character state being incorporated in a multistate transformation series, as, *e.g.*, '3' in a series from 0 to 5, is prone to be influenced by the behaviour of the other elements in the series and thus may not be completely 'free' to manifest itself as apomorphous at a certain level in the analytical procedure. This is because, the outcome of an analysis will depend on the distribution of all character states throughout the matrix and, though character states 'linked' together in a single, multistate character will tend to have a greater influence on the final result as a group, it is generally considered that this may at the same time obscure their own, individual merits to some (though hardly quantifiable) degree. Moreover, any 'greater influence as a group' is fundamentally undesirable, since any such influence has an *a priori* chance of directing the final result, thus reducing the independent, objective character of the analysis.

It has thus been considered that the purest way of representing multistate characters would be to break these down into series of binary characters. In executing the binarization, each state is recognized as a separate character that can either be present, or absent. In the above example of three colours, this then leads to:

not red *vs.* red 0 - 1
not yellow *vs.* yellow 0 - 1
not blue *vs.* blue 0 - 1

Thus, the equivalence of the colours is guaranteed and each may be valued according to its own merits and behaviour in the course of the pattern analysis, literally as an equal towards the other states occurring for that character. Hence, as we may presume, this way the fundamental objectivity of this part of the procedure can be maximized.

The implementation of a/p character coding

In an attempt at formalizing the process of the coding of characters initially recognized as multistate classes in the form of binary sequences, a seminal paper was produced by Pleijel (1995). In that study, the author has investigated various ways of implementing such an operation and the ultimate conclusion he reached, acknowledged the fundamental superiority of the so-called absence/presence (or a/p) type of coding. In this protocol, all individual states into which a transformation series may be dismembered are treated as potential apomorphies (cf. Pleijel, 1995: 315), thus guaranteeing the maximization of the chances for those character states to be recognized according to their true, historical status at any relevant level in the final cladogram.

Pleijel's (1995: 310, his fig. 1) paradigm comprised a feature X that is found in five conditions, or expressions: (1) absent; (2) round and black; (3) round and striped; (4) square and black; and (5) square and striped, which he coded in a binary way by implementing this a/p procedure (type 'D' in his paper). The scheme advocated to such end was, using his own example (compare also fig. 1 herein), to code the five possible states as follows:

(1) Feature X: absent (0) / present (1)
(2) Rounded shape of feature X: absent (0) / present (1)
(3) Square shape of feature X: absent (0) / present (1)
(4) Black pigmentation of feature X: absent (0) / present (1)
(5) Striped pigmentation of feature X: absent (0) / present (1)

Soon after the publication of Pleijel's (1995) paper, systematists became intrigued by this method, to which the impact of the paper at the 1997 Oxford conference may testify: in the proceedings (Scotland & Pennington, 2000), seven out of ten papers (including the Introduction) cite his article, published hardly two years before the congress took place. Attending myself at the Oxford meeting, I can state that various participants only learned about the method described by Pleijel (1995) either shortly before, or even at the conference: but even so some had quickly reworked their presentations according to the a/p coding scheme in an attempt at ameliorating, or at least better corroborating, their results and conclusions.

Advantages of a/p character coding

The most prominent advantage of the transcription scheme recommended by Pleijel (1995) unquestionably is the possibility to assess the phylogenetic significance of each individual character state according to its own behaviour in the analytical procedure. As each character state is entered as a separate character, this will allegedly ensure its maximal freedom to emerge as a(n) (syn)apomorphy in the resulting cladogram, provided, of course, that such a status is embodied in the datamatrix as a whole.

The basic principle is, that all character states are essentially equal: compare the example of the colours above. Why should blue have a higher value (3) than either yellow (2), or red (1)? In what way does the arbitrary assignment of those values influence the resulting cladogram? With a/p coding, such questions cannot even be asked, for the various states are treated as equal from the start. Although this by no means can imply that a/p coding would invariably emerge as the 'best performer' among coding schemes (cf. Forey & Kitching, 2000; Hawkins, 2000), it certainly may be recognized as carrying the least, even minimal inherent bias, as no (not even inadvertent) weight is assigned to any character state in particular. With a/p coding, it thus would seem, the judgment of the investigator would tend to be minimized and hence the result based on a maximal influence of the character state distributions over the datamatrix *per se*.

Another, also quite convenient trait of a/p coding is that, as Pleijel (1995: 312) observes '... the problem with inapplicable character states disappears; ...'. Indeed, where simply 'absence' or 'presence' are coded, without any *a priori* interpretation, the true status of each state at every level may be expected to eventually emerge from the analysis by itself. This means that no special requirements are necessary to deal with missing entries or inapplicable data (compare, *e.g.,* table 2).

Table 2. Datamatrices corresponding to the configuration depicted in fig. 1a: a, the initial matrix; and, b, the reduced matrix as adapted following determination of the first, i.e., basal dichotomy in the ingroup, (A)-(B-G). As character (1) is no longer variable in the new ingroup (B-G), it carries no phylogenetic information relevant for the structure of that group and has hence been omitted. Note there is no difference in polarity of the states in (a) and (b), as the new outgroup (A) in matrix (b) shows the same character states as the original outgroup (OG) in matrix (a).

Taxon / Character	(1) X	(2) Ro	(3) Sq	(4) Bl	(5) St
a, Initial datamatrix:					
OG	0	0	0	0	0
A	0	0	0	0	0
B	1	0	1	1	0
C	1	0	1	0	1
D	1	1	0	1	0
E	1	1	0	1	0
F	1	1	0	0	1
G	1	1	0	0	1
b, Secondary datamatrix:					
A (= new OG)		0	0	0	0
B		0	1	1	0
C		0	1	0	1
D		1	0	1	0
E		1	0	1	0
F		1	0	0	1
G		1	0	0	1

X, feature X; Ro, rounded; Sq, square; Bl, black; St, striped

Disadvantages of a/p coding examined

When advocating a/p coding as an optimized way of binarizing multistate characters (i.e., transformation series with more than two elements), Pleijel (1995) already admitted there remain three serious problems with this kind of coding, the first of which being considered the most important: (a) an effect of

redundancy, i.e., more elements are introduced into the datamatrix than absolutely necessary. Indeed, the presence of feature X will already be apparent from two of the characters nos. (2)-(5) scoring '1', whereas the absence of X will immediately be revealed by four '0's for those same characters. This obviously means that scoring 'Feature X absent or present' as a separate character, seems superfluous. In addition, two more undesired effects may occur: (b) the phenomenon of character linkage, i.e., if dissected that far, the various elements of a transformation series cannot be regarded as independent variables any longer: presence of a square shape automatically implies that 'rounded shape' scores '0', and then also one of the colours will score '1', the other '0'. Finally, (c) the effect of homoplasy bias, i.e., the disproportionate weight binary coded multistate characters may get over 'purely' binary characters in a mixed matrix: the above five-state character makes four or five binary characters, whereas a truly binary character will never make more than one character, by its very nature. The recoded multistate characters will thus tend to outweigh the purely binary characters by the sheer numbers of their states, all coded as separate characters. Next, any redundant information as noted under (a), above, may enlarge the already inevitable effect of imbalance in numbers, again at the expense of the (effective) significance of the purely qualitative, binary characters.

These effects were also immediately recognized at the 1997 Oxford meeting, and participants agreed with the author that the scheme inherently suffered from a certain redundancy: if X is present, one shape and one colour will each score '1', so adding another '1' for the mere presence of X does not convey any additional information. Thus, the schedule most users of a/p coding soon adopted, consisted of deleting character no. (1) in the above series, i.e., the primary recognition of the presence or absence of feature X as such. This 'reduced a/p coding' would allegedly subvent the redundancy acknowledged.

As the author already pointed out himself (Pleijel, 1995), there could be chances that the disadvantages noted above may unduly affect the results of cladistic analyses, if not properly dealt with. So, this is why I herein suggest a correction towards the actual application of a/p coding as described, in order to minimize the redundancy effect, viz., by performing cladistic analyses in a stepwise manner. Presumably, this will generally imply that all three negative aspects of the procedure may be considered to potentially become restrained within reasonable limits, since the effects mentioned above eventually all come down to some sort of redundancy, i.e., to inadvertent weighting.

Performing pattern analysis stepwise

Cladistic pattern analysis is usually performed with the aid of computerized algorithms embedded in computer programs and, as a rule, in one go: the 'total analysis' approach. The datamatrix as a whole is analysed to find the 'best' structure, which, according to the program at issue, is indicated as the 'most parsimonious' solution, or the solution 'of best fit', or an equivalent term. Usually, a large but restricted number of possible cladograms is probed and from those, the one(s) requiring the least character transformations is/are presented as the result(s). Whether initially rooted or unrooted (the latter yielding a network only), eventually all results will be rooted to produce one or more 'maximally parsimonious' cladograms, from which (if more than one) the investigator has to choose the one that is judged most appropriate (for instance, based on additional, qualitative arguments).

Alternatively, the (super)datamatrix can be analysed in parts, the so-called 'partitioned' approach, in which, *e.g.*, morphological data are examined separately from molecular data, and the resulting trees will have to be united into a single, integrated cladogram through the use of 'consensus' techniques.

Afterwards, the 'most parsimonious' cladogram eventually chosen is usually supported, or corroborated in a statistical sense, by indicating (often branch by branch) the robustness of the procedure (i.e., mostly based on the percentages of original hypotheses of homology that have not been discarded) by referring to a 'consistency index' and/or a 'retention index', or else through the application of additional techniques like 'jackknifing', 'bootstrapping', and the like. These routines all purport to demonstrate that the result ultimately accepted indeed represents the best choice from the, usually many, possible options, and all basically resort to the use of the initial datamatrix.

Disregarding the technical (i.e., mathematical) details of the algorithms and of the analytical procedure, the process of rooting, whether beforehand or afterwards, usually involves applying the criterion of outgroup comparison: the character states in the ingroup are compared to the states present in the outgroup, which are by definition considered plesiomorphous, whereby the alternative states are labelled as apomorphous for the purpose of the analysis. In the majority of cases, so it would seem, those initial labels 'plesiomorphous' and 'apomorphous' are retained in the course of the entire analysis, which effectively means: in structuring the cladogram as a whole.

My prime concern with these methods is, that they all apparently use the initial datamatrix to find the underlying phylogenetic structure throughout the entire cladogram. However, each and every character state is, in fact, of phylogenetic relevance only at a single level, i.e., exactly at the level where that state once developed as an evolutionary novelty and thus now constitutes a synapomorphy for the taxa that have since evolved from the common ancestor that developed that apomorphous state. At all other levels in the cladogram, that state does not convey phylogenetic information and thus is there, at best, neutral with regard to the performance of the analytical procedure.

In the scenario of a stepwise analysis, however, the initial datamatrix is only used to find the first dichotomy that follows the root. To resolve the structure of the higher levels of the tree, the matrix is adapted (a) by removing those characters that, from that point onwards, are no longer variable; and also (b) by a renewed polarization of the remaining characters and their states according to that, now accepted, first dichotomy – thereby employing the (newly found) sister group(s) as (an) outgroup(s); and finally (c) by removing the initial (or subsequent) outgroup 'after use'.

Fig. 1a, based on the various states of feature X of Pleijel (1995), shows a suite of seven ingroup taxa, A-G, plus an outgroup OG. According to classical outgoup comparison, the presence of feature X constitutes an apomorphous state at the split (A) *vs.* (B-G). The most parsimonious solution to explain the distribution of feature X over the cladogram is to assume that X developed in the common ancestor of clade (B-G) and hence constitutes a synapomorphy for that group, forming an argument for its monophyly.

From the above it follows, that character (1) in the scheme of Pleijel (1995), 'Feature X absent or present', is only informative there, and makes a redundancy when analysing the structure of clade (B-G), since all members of that clade by definition possess

feature X. Thus, it would be desirable to omit character (1) in the analysis of that clade. On the other hand, the presence of X as such provides one argument to recognize clade (B-G), whence it constitutes by no means a redundant character at level (A)-(B-G).

This is where the advantage of a stepwise procedure becomes clear: character (1), X absent/present, is to be included in the primary datamatrix in which all character states have been coded as separate characters, but once the first dichotomy in the structure of the ingroup has been determined, this has to be recognized as the basis of the (here already rooted) cladogram. Then the datamatrix needs to be 'cleaned' from redundant characters, i.e., those characters that are invariable from that level upward must be deleted to yield a reduced matrix, and the former outgroup has to be discarded as well. Next, a successive outgroup comparison has to be performed in which the remaining terminal clade (B-G) figures as ingroup and the sister group, (A) now is taken into account as the outgroup. Generally, several characters may have to be repolarized, as now the character states of the new outgoup A are by definition considered plesiomorphous (though this is not always at issue, see below). A point that should also be noted is, that the OG-comparison in the program will have to decide for each character state which of the conditions '0' or '1' is to be considered plesiomorphous in any given case: since '0' is always coded for absence and '1' for presence, '0' is not automatically interpreted as plesiomorphous, or '1' as apomorphous.

The above means that the original matrix for feature X as in table 2a, will be reduced to the scheme in table 2b: here only characters (2)-(5) are included, while

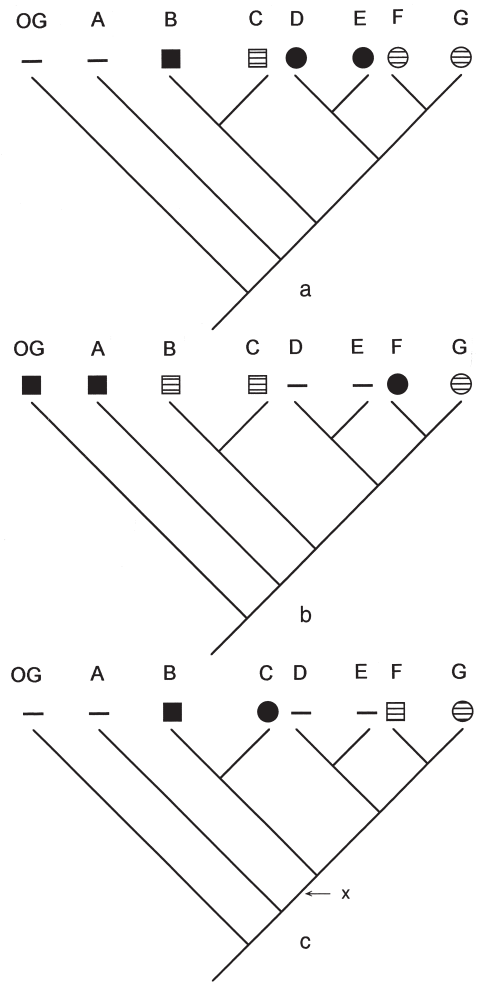

Fig. 1. Some hypothetical examples of alternative possibilities with regard to the apomorphous or plesiomorphous status of the expressions of 'Feature X'; the shape of the cladograms has been kept congruent for easy comparison. The absence of X as such is plesiomorphous only in (a), apomorphous only in (b), and plesio- as well as apomorphous in (c), i.e., 'absence' is not homologous throughout the cladogram. Likewise, the striped pigmentation would be synapomorphous in (c), while partly homoplastic and partly synapomorphous in (a) and (b). The black pigmentation would be plesiomorphous in (a), partly plesiomorphous with a reversal in (b), and symplesiomorphous in (c). The shapes are shown as homoplastic in (c), whereas in (a) and (b) 'rounded' would be synapomorphous *versus* 'square' being symplesiomorphous.

(1) now has been removed, as has group OG. The fact that no repolarization has been performed is due solely to the fact, that also in the new outgroup 'Feature X' is absent and hence its derivatives, (2)-(5) score an '0' there. The possession of the shapes, square or rounded, nor of the colours, black or striped, can be polarized in (B-G) through mere outgroup comparison, since in outgroup A feature X is absent, and thus has neither a shape, nor a colour. However, the initial datamatrix already implied that same condition, and retaining character (1) would make no difference in this respect. Obviously, determining the structure of clade (B-G) in this example will depend on other characters and their states, and the true status of both the shapes and the colours of X will emerge afterwards by interpreting the distribution found according to the most parsimonious hypothesis of character transformation. Thus, with respect to the status of 'Feature X present', this state was assigned as apomorphous in the beginning and would have retained that status throughout the analysis, if not character (1) had been removed after the first dichotomy had been determined: it would have been uninformative and hence redundant from level (A)-(B-G) onward.

In another hypothetical scenario, depicted in fig. 1b, the situation is different: here the absence of X constitutes an apomorphous loss, because X as such was already present as a plesiomorphous state in the whole ingroup (A-G) according to the initial outgroup comparison. Here, the respective matrices will be shaped as in table 3a-c. As regards the absence of X, this state thus was qualified as apomorphous in the initial matrix and will retain that status during several successive stages in the stepwise procedure.

Table 3. Datamatrices corresponding to the configuration depicted in fig. 1b: a, the initial matrix, with 'Feature X present' as the plesiomorphous state; b, the reduced matrix as adapted following determination of the first, i.e., basal dichotomy in the ingroup, (A)-(B-G) by omitting OG, but retaining character (1) as this is variable here; and, c, the partial matrix for groups (B-C) and (D-G). Again, there is no difference in polarity of the states in (a) and (b); in (c) outgroup comparison is only possible for the pigmentation characters, but an interpretation whether 'black' in F constitutes a new apomorphy or a reversal, or is due to retaining the ancestral condition from (OG-A), cannot reliably be established; OG comparison is not directly possible for shape 'rounded', which is, however, provisionally considered apomorphous as in (b), while the absence of feature X in (D-E) is interpreted as an apomorphous loss.

Taxon / Character	(1) X	(2) Ro	(3) Sq	(4) Bl	(5) St
a, Initial datamatrix:					
OG	1	0	1	1	0
A	1	0	1	1	0
B	1	0	1	0	1
C	1	0	1	0	1
D	0	0	0	0	0
E	0	0	0	0	0
F	1	1	0	1	0
G	1	1	0	0	1
b, Secondary datamatrix:					
A (= new OG)	1	0	1	1	0
B	1	0	1	0	1
C	1	0	1	0	1
D	0	0	0	0	0
E	0	0	0	0	0
F	1	1	0	1	0
G	1	1	0	0	1
b, Tertiary, partial datamatrix for (D-G) with (B-C) as outgroup, and *vice versa*:					
B-C (= new OG)	1	0	1	0	1
D	0	0	0	0	0
E	0	0	0	0	0
F	1	1	0	1	0
G	1	1	0	0	1

X, feature X; Ro, rounded; Sq, square; Bl, black; St, striped

Fig. 1c presents a condition in which the absence of feature X presumably comprises both a plesiomorphous state (as in OG and A), and an apomorphous loss, as in

Table 4. Datamatrices corresponding to the configuration depicted in fig. 1c: a, the initial matrix; b, the reduced matrix as adapted following determination of the first, i.e., basal dichotomy in the ingroup, (A)-(B-G); and, c, the partial matrix for groups (B-C) and (D-G). Also here, there is no shift in polarity of the states in (a) and (b). Based on outgroup comparison alone, neither the shapes nor the pigmentation can be decided to be apo- or plesiomorphous in (B-G). The absence of feature X in (D-E) is considered apomorphous, though, whether as a truly new development or as a reversal to the ancestral state without feature X.

Taxon / Character	(1) X	(2) Ro	(3) Sq	(4) Bl	(5) St
a, Initial datamatrix:					
OG	0	0	0	0	0
A	0	0	0	0	0
B	1	0	1	1	0
C	1	1	0	1	0
D	0	0	0	0	0
E	0	0	0	0	0
F	1	0	1	0	1
G	1	1	0	0	1
b, Secondary datamatrix:					
A (= new OG)	0	0	0	0	0
B	1	0	1	1	0
C	1	1	0	1	0
D	0	0	0	0	0
E	0	0	0	0	0
F	1	0	1	0	1
G	1	1	0	0	1
b, Tertiary datamatrix for (D-G) with (B-C) as outgroup, and *vice versa*:					
B (= new OG *p.p.*)	1	0	1	1	0
C (= new OG *p.p.*)	1	1	0	1	0
D	0	0	0	0	0
E	0	0	0	0	0
F	1	0	1	0	1
G	1	1	0	0	1

X, feature X; Ro, rounded; Sq, square; Bl, black; St, striped; *p.p., pro parte*

clade (D-E). The corresponding matrices are given in table 4a-c. With respect to the absence of X, that state initially coded as plesiomorphous, but repolarized as apomorphous from level (B-C)-(D-G) onward: in this case, character (1) has not been removed after step no. 1, since it obviously carries phylogenetic information of use in structuring clade (D-G). Thus, it may be clear that 'Feature X absent', though initially qualified as plesiomorphous, nonetheless appears as a synapomorphy in the course of the continued analysis. Hence, it would have been unjustified to omit character (1) from the primary matrix, as the possibility for that state to emerge as a(n) (syn) apomorphy may not be excluded beforehand.

Performing stepwise analysis does not mean that all bias can be avoided: in feature X, either the shape, or the pigmentation will presumably represent a synapomorphy, leaving the other trait as a (pair of) homoplastic character state(s): compare fig. 2a-b, to consider only the simplest possible scenarios. Which of the pairs represents a synapomorphy and which must be considered homoplastic *a posteriori*, can only emerge from the analysis by taking into account the other characters in the matrix (here not shown or discussed) and implementing the principle of parsimony. If the 'shape' states cluster together, at least one of these (square or rounded) will be interpretable as a synapomorphy, the other one as a symplesiomorphy. The same will hold true, *mutatis mutandis*, for the colours: either black will represent the ancestral condition, or striped. It should be stipulated explicitly here, that only data based on actual observations can be admitted to the analysis, which means that possibilities of ancestral conditions of shape being something else than rounded or square, *e.g.*, 'triangular', or of pigmentation deviating from black or striped, like 'blank' (fig. 3a-b), cannot be taken into account.

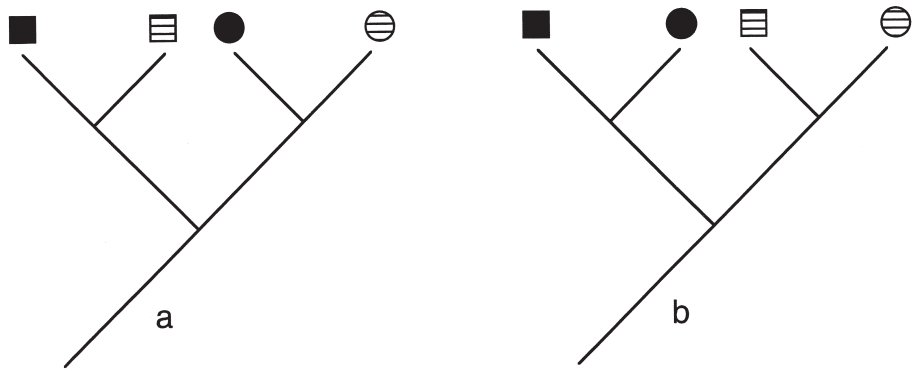

Fig. 2. Homoplasy in pigmentation (a) *versus* homoplasy in shape (b) of feature X. See text for further explanation.

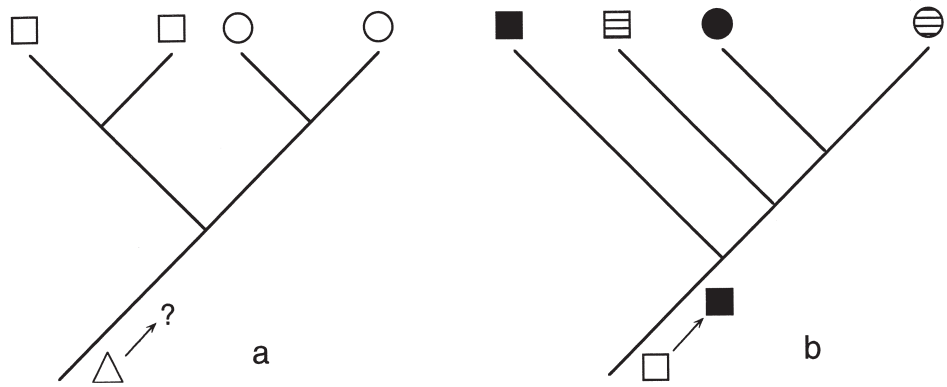

Fig. 3. No allowance should be made in the datamatrix for possibilities that the ancestral state of shape may have been different from either 'square', or 'rounded', *e.g.*, 'triangular', or that of pigmentation having been, *e.g.*, 'blank', if there are no concrete indications for assuming the situations as depicted.

Redundancy and the stepwise procedure

From the above (fig. 1a-c, tables 2-4), it is apparent that character (1) 'Feature X absent/present' is only redundant at certain stages in the analysis of certain configurations, like in the second stage documented in fig. 1a and table 2. Hence, omitting character (1) from the initial matrix is not an option: but neither is it an option to retain character (1) redundantly throughout the analysis in, *e.g.*, this same case of fig. 1a, table 2. The stepwise execution of the analytical procedure, however, at least reduces the problem, hence mitigates the effect of redundancy, by removing character (1) as soon as the phylogenetic information contained in its states is no longer relevant for the structuring of the remaining clade(s), and thus should be acknowledged as offering an at least partial solution to the disadvantage of redundancy that the a/p coding protocol inherently carries with it. Obviously, only a single binarized multistate character is at

issue, but in a large datamatrix the reduction will be more substantial, hence also the effect of subventing part of the otherwise accumulating redundancy.

That omitting 'Feature X absent/present' from the initial datamatrix is not a viable option, even more prominently emerges from the cases shown in fig. 1b-c, tables 3-4: there even the absence of feature X eventually emerges in whole (fig. 1b) or in part (fig. 1c) as a synapomorphy.

Thus, the redundancy noted when transcribing the states of character X with the use of the full a/p coding, will only hold true for the ingroup, i.e., for those taxa indeed possessing feature X. In a larger dataset, by contrast, i.e., a matrix in which taxa without X are included (and in which the possession of X thus is a synapomorphy for the ingroup), recognition of the mere presence of X, in whatever form, is by no means to be qualified as redundant. This obviously means that scoring the presence or absence of X as a separate character, will only be redundant from a certain level onward, being the level at which the mere presence of X constitutes a plesiomorphy. Removing character (1), X absent or present, from the matrix in successive analyses of already recognized clades, thus subvents the problem of redundancy to some degree. Though not entirely eradicating redundant information at all levels, at least the effect may be qualified as having been reduced.

The same may hold true, *mutatis mutandis*, for the other two phenomena noted, i.e., character linkage and homoplasy bias: also the influence of those will be reduced when successive datamatrices are cleaned from uninformative, superfluous data.

These effects thus constitute advantages of employing a stepwise protocol in the cladistic analytical procedure. However, the practice of performing a procedure in discrete steps is not unproblematic, for it requires adapting the datamatrix manually at each successive node, i.e., before the analysis of every subsequently recognized clade.

Problems in performing an analysis stepwise

The main problem encountered when performing a cladistic analysis stepwise allegedly would be, that such an operation will have to be performed partly by hand: not something scientists in this computer age will be particularly fond of. At each step, viz., the datamatrix will have to be adapted, and this has to be done manually. On the other hand, the newly required outgroup comparison(s), i.e., with the sister group(s) of the newly determined clade(s), may well be performed with the usual algorithms embedded in the existing, computerized programs for phylogenetic analysis. Scientists thus should: (a) consider whether or not in the cases they wish to analyse, the advantages thus provided will justify the effort required; and (b) whether or not such an iterative routine could eventually be designed, and next be included in available programs. One would be inclined to think this might certainly be worth attempting, or at the very least, considering.

Discussion

Despite the promising start of absence/presence coding, Pleijel & Dahlgren (1998) noted, three years after Pleijel (1995) demonstrated the value of the routine, that although a/p coding was by then in common use, it was hardly ever implemented consistently.

Most users, as they remarked, appeared to mix the method with other coding techniques when composing their datamatrices. Notwithstanding the fact that Kitching et al. (1998: 30) state that the method of absence/presence coding yields data that would tend to remain stable at any level of analysis, they also note that a/p coding is only seldom used, and then certainly not exclusively. They consider that the apparent hesitation of authors to apply this type of coding throughout may be due to linkage problems the users experience in their analytical procedures (Kitching et al., 1998: 36). The technique has, however, been referred to extensively on a theoretical level in various contributions in Scotland & Pennington (2000). Hawkins (2000) included a/p coding in a survey of a large number of coding criteria, while Forey & Kitching (2000) gave a detailed evaluation in comparison with various other types of character coding. More recently, Struck et al. (2006) mentioned the method in passing but did not resort to it further in their own analyses. Other aspects of implementing multistate characters are still being examined and ameliorated nowadays, as in, *e.g.*, the sophisticated approach to continuous characters developed by Goloboff et al. (2006) and the improvement of dealing with meristic characters as detailed by Lawing et al. (2008). Nevertheless, the format of character coding in general seems not to have reached a truly final stage, yet.

In regard of the alleged influence of the use of coding values higher than the binary 0 and 1 as with, *e.g.*, continuous and meristic characters as well as with the classes presented in the example of the three colours, above, it may be surmised that such an influence could be real if the character at issue is used in an ordered state but is less so, or even non-existent, when used unordered, i.e., without a particular sequence in the values. This is because, in the ordered state a meaning is attributed to the value of the coding as such, which is absent in case of an unordered use. [Note that polarization should have no such influence, since then, at any time, only two of the states are opposed to each other, viz., as plesiomorphous *versus* apomorphous.] However, it should also be considered that (a) in a phenetic analysis a value of, *e.g.*, 3 is more influential than a '1', for purely arithmetic reasons; (b) it is hardly ever made sufficiently clear in the various cladistic analytical routines that such an arithmetic influence has been satisfactorily subvented; and such may even be unlikely (c) in view of the acknowledged, serious problem posed by the phenomenon known as 'long-branch attraction'. In cladistic pattern recognition, long-branch attraction (cf. Bergsten, 2005) stands for the unwarranted grouping of two or more long branches in a cladogram as sister groups, based on false hypotheses of homology. This is likewise interpreted as an artifact originating largely from arithmetic influences for which no satisfactory solution seems to have been found until now. Hence, trying to avoid high values in coding through the binarization of multistate characters remains a sensible option when processing data in an analytical protocol, in an attempt at staying on the 'safe side', i.e., evoking as little bias as possible.

With respect to the problems of redundancy, character linkage, and homoplasy bias, it may be noted that these may be real on a theoretical level, but their influence in practice is hard to estimate and might as well be limited. In phylogenetics, it is generally acknowledged that the coordinate evidence embodied in the truly homologous character states will, under the regimen of maximized parsimony, overrule the scattered information contained in the states that result from homoplasy or character reversal. So, the additive directional effect of all phylogenetically relevant information from the datama-

trix will eventually reveal the true significance that has to be granted to the different manifestations of feature X in various parts of the cladogram. Yet, these states can only perform optimally if allowed to freely express their own information that (in part) determines the correct evolutionary direction, i.e., without being impeded by linkage to other characters. We thus need to consider the possibilities for distribution of the states of a character and next decide in which way the matrix might or might not be adapted, so as to avoid that some potential distributions would be restricted in playing their full part in shaping the historically correct tree. The adaptation alluded to may involve the coding protocol, or the analytical procedure, or both. If we wish to give a/p coding its full credit, we may thus have to use the stepwise protocol as argued in the above.

A further issue may be recognized in the choice of the outgroup: ideally an outgroup ought to be constituted by the sister group of the ingroup, i.e., by that group that is allegedly closest to the group to be analysed. There is, of course, the general theoretical problem that, in order to determine which group is the sister group of the group at issue, a cladistic analysis would be needed to determine this, for which, however, another outgroup will be required, again ideally comprising the sister group of that (now larger) group — and so on and so forth. However, in a stepwise procedure the first step will involve finding the basic dichotomy of the clade analysed, whence at subsequent (= higher) levels in the cladogram (= lower taxonomic levels) sister groups will emerge more or less automatically from the procedure. It may thus be presumed that this protocol at least embodies an approach that is as close to the ideal situation as possible.

Though in the above explicit reference has been made to a stepwise execution of cladistic analyses, this does not seem to be a very popular kind of procedure. While it has been advocated before (*e.g.*, Estabrook et al., 1977; Vaupel Klein, 1984, 1987) this was primarily conceived within a framework of attempts at optimizing character compatibility methods (cf. Le Quesne, 1979, 1982). As, however, compatibility seems to be hardly employed any more, those references may no longer be relevant nowadays. In an overview of current phylogenetics software listed at http://en.wikipedia.org/wiki/List_of_phylogenetics_software [accessed 22.iii.2009], a total of 29 programs figures, from 'BEAST' to 'Xrate', most of which are based on either 'parsimony', 'maximum likelihood', 'distance criteria', 'Bayesian inference', or various other clustering methods, whereas I was unable to find an explicit reference to compatibility. Yet, compatibility is still taken into account as a possible mode of analysis, as Felsenstein (2004) quite recently gave the method ample attention in his overview of methods for phylogenetic inference.

In trying to execute an analysis stepwise from the start, it may be noted that in the 'Integrated Approach' I have used earlier (Vaupel Klein, 1984) and then advocated, the first step comprised a phenetic analysis on (undirected) similarity, only in order to find some basic structure in the dataset. Many systematists will perform a similar step intuitively, by inspecting the group of taxa they wish to analyse, thereby just relying on their experience. After all, one has to choose from a vast collection those taxa to be either included in, or excluded from the analysis to be performed. The choice of a taxon to be analysed and to determine its closest relatives constitutes, in fact, a 'step no. 0' in every analysis of cladistic relationships.

Yet, the fact that it seems hard to find examples of stepwise analyses does not mean that these would not be performed. For example, Fransen (2002) used a stepwise

method to analyse the phylogeny of the genus *Pontonia* (Crustacea: Decapoda: Palaemonidae) by first searching for monophyletic groups (genera) within a subfamily and next analysing the relationships within the constituting genera using (a member of) the most closely related genus as an outgroup. Analysing relations at family level will often require other characters than analysing the internal structure of genera, so not infrequently those analyses will be separated: this may be considered an implicitly stepwise approach, and it often will be even more implicit as the familial and generic analyses may be published in separate papers, even by different authors: the steps are there, but not in a direct, obvious connection with each other.

Though Pleijel (1995: 311) indicated the possibilities for performing stepwise analyses (i.e., in his paragraph on 'Hierarchical Character Linkage'), he did not further develop this issue into a (partial) solution for the problems he noticed.

The only explicitly iterative procedure in use today seems to consist of the approach through 'successive character weighting' descibed by Farris (1969). Though probably not often applied in its basic form any more, the principle has been included in the 'implied weights' routine of Goloboff (1993), which now forms part of the application 'Tree Analysis Using New Technology' (T.N.T.) by Goloboff et al. (2003). As far as I can assess, this currently would constitute the only procedure in which an iterative routine has been implemented: that, in other words, adapts the influence of characters and their states according to their performance in the analysis.

However, the 'total' approaches, i.e., in which a datamatrix as a whole or in part ('partitioned' routine) is analysed in one large move, seem to hold a dominant position among the various protocols available. In order to avoid any confusion: 'total' thus is not just meant here as referring to 'total evidence' (all data lumped together) *vis-à-vis* the partitioned way of analysing a datamatrix, with, *e.g.,* morphological and molecular data being examined separately and only the results being combined. The qualification 'total' herein rather contrasts that 'single stroke' analysis with a stepwise procedure. Now in a survey of volume 23 of *Cladistics* (2007), 23 papers were found to deal with specific analyses of real taxa, and all used either the truly 'total' approach or a 'partitioned' routine combined with consensus techniques (*e.g.*, Richter et al., 2007): no stepwise procedure could be detected. The same was found in screening volume 57 of *Systematic Biology* (2008): in 29 analyses referring to actual taxonomic groups, again all appeared to have been performed with either total routines, or partitioned approaches, and, again, not a single one in a truly stepwise manner. Apart from the architecture of the procedure, the only study in which some manual manipulation of data was explicitly stated to have been included, was found to be a paper by Dunlap et al. (2007) on bioluminescent symbioses.

Yet, as I hope to have demonstrated in the above, the principle of building up a cladogram node by node and branch by branch, and adapting the datamatrix at each successive step to achieve that, remains an option that would deserve more attention than it apparently receives today.

Conclusion

So, in conclusion, by performing a phylogenetic analysis stepwise, redundancy can be minimized: no more expressions of a multistate character require coding in any of

the (partial) datamatrices than absolutely necessary. This also implies that linkage is minimized, for no more is there a character (like the original character (1), cf. Pleijel, 1995) that 'automatically' gets code '1' if one of the other characters ((2)-(5) of Pleijel, 1995) is present. Homoplasy bias is reduced as well, since no basically uninformative character is present that, as an inadvertent side-effect, could corroborate (false) information content in characters that are, at the level at issue, still considered informative in the analysis.

Yet, no relevant phylogenetic information is ever obscured, or its expression hampered beyond methodological necessity. In working this way, we may thus minimize the disadvantages of the a/p character coding scheme by minimizing redundancy, whilst making maximal use of the advantages of the system: just by performing the phylogenetic pattern analysis stepwise instead of in one whole.

Acknowledgements

I take great pleasure in dedicating this paper to the honour of my former *collega proximus*, Dr A.C. van Bruggen, on the occasion of his 80th birthday. From 1969 until his retirement in 1994, we have been working closely together at what was then the Division of Systematic Zoology at Leiden University, especially in the *curriculum* for undergraduates. I am sure that his broad knowledge of the animal kingdom and of taxonomy in general, as well as his great dedication towards teaching the very essence of biology, will be remembered by many generations of his former students.

In addition, I would like to thank Prof. Dr E. Gittenberger and Dr M. Zandee, both formerly of Leiden University, for commenting upon a (much) earlier draft of a part of the present text. Mr M. Brittijn, former staff artist at the Department of Biology of the same university, skilfully took care of the illustrations. Finally, I am grateful to Dr C.H.J.M. Fransen of the National Museum of Natural History in Leiden and to two anonymous referees for valuable suggestions in the final stage of preparing this paper.

References

Bergsten, J., 2005. A review of long-branch attraction.— Cladistics 21: 163-193.

Dunlap, P.V., J.C. Ast, S. Kimura, A. Fukui, T. Yoshino & H. Endo, 2007. Phylogenetic analysis of host-symbiont specificity and codivergence in bioluminescent symbioses.— Cladistics 23: 507-532.

Estabrook, G.F., J.G. Strauch, Jr. & K.L. Fiala, 1977. An application of compatibility analysis to the Blackiths' data on orthopteroid insects.— Syst. Zool. 26: 269-276.

Farris, J.S., 1969. A successive approximations approach to character weighting.— Syst. Zool. 18: 374-385.

Felsenstein, J., 1982. Numerical methods for inferring evolutionary trees.— Quart. Rev. Biol. 57(4): 379-404.

Felsenstein, J., 2004. Inferring phylogenies: i-xx, 1-664.— Sinauer Associates, Sunderland, Massachusetts.

Forey, P.L., C.J. Humphries, I.L. Kitching, R.W. Scotland, D.J. Siebert & D.M. Williams, 1992. Cladistics. A practical course in systematics.— Syst. Assoc. Publ. 10: i-xi, 1-191.

Forey, P.L. & I.J. Kitching, 2000. Experiments in coding multistate characters. In: R.[W.] Scotland & R.T. Pennington, Homology and systematics. Coding characters for phylogenetic analysis.— Syst. Assoc. spec. Vol. Ser. 58: 54-80.

Fransen, C.H.J.M., 2002. Taxonomy, phylogeny, historical biogeography, and historical ecology of the genus *Pontonia* Latreille (Crustacea: Decapoda: Caridea: Palaemonidae): i-xiv, 1-433.— Ph.D. Thesis, Leiden University, Leiden. [Scientific part also as: Zool. Verh., Leiden 336. National Museum of Natural History Leiden.]

Goloboff, P.A., 1993. Estimating character weights during tree search.— Cladistics 9: 83-91.
Goloboff, P.A., J.S. Farris & K.C. Nixon, 2003. T.N.T. Tree Analysis Using New Technology.— Program available at: http://www.zmuc.dk/public/phylogeny.
Goloboff, P.A., C.I. Mattoni & A.S. Quinteros, 2006. Continuous characters analyzed as such.— Cladistics 22: 589-601.
Hauser, D.L. & W. Presch, 1991. The effect of ordered characters on phylogenetic reconstruction.— Cladistics 7: 243-265.
Hawkins, J.A., 2000. A survey of primary homology assessment: different botanists perceive and define characters in different ways. In: R.[W.] Scotland & R.T. Pennington, Homology and systematics. Coding characters for phylogenetic analysis.— Syst. Assoc. spec. Vol. Ser. 58: 22-53.
Hennig, W., 1950. Grundzüge einer Theorie der phylogenetischen Systematik: i-ii, 1-370.— Deutsches Entomologisches Institut, Berlin-Friedrichshagen / Deutscher Zentralverlag, Berlin.
Hennig, W., 1966. Phylogenetic systematics: 1-263.— University of Illinois Press, Urbana, Illinois.
Huelsenbeck, J.P. & D.M. Hillis, 1993. Success of phylogenetic methods in the four-taxon case.— Syst. Biol. 42(3): 247-264.
Kitching, I.J., P.L. Forey, C.J. Humphries & D.M. Williams, 1998. Cladistics. Second edition. The theory and practice of parsimony analysis.— Syst. Assoc. Publ. 11: i-xiii, 1-228.
Lawing, A.M., J.M. Meik & W.E. Schargel, 2008. Coding meristic characters for phylogenetic analysis: a comparison of step-matrix-gap-weighting and generalized frequency coding.— Syst. Biol. 57(1): 167-173.
Meier, R., 1994. On the inappropriateness of presence/absence coding, for non-additive multistate characters in computerized cladistic analyses.— Zool. Anz. 232: 201-209.
Pleijel, F., 1995. On character coding for phylogeny reconstruction.— Cladistics 11: 309-315.
Pleijel, F. & T. Dahlgren, 1998. Position and delineation of Chrysopetalidae and Hesionidae (Annelida, Polychaeta, Phyllodocida).— Cladistics 14: 129-150.
Quesne, W. le, 1979. Compatibility analysis and the uniquely evolved character concept.— Syst. Zool. 28(1): 92-94.
Quesne, W. le, 1982. Compatibility analysis and its applications. In: C. Patterson (ed.), Methods of phylogenetic reconstruction.— Zool. J. Linn. Soc., Lond. 74: 267-275.
Richter, S., J. Olesen & W.C. Wheeler, 2007. Phylogeny of Branchiopoda (Crustacea) based on a combined analysis of morphological data and six molecular loci.— Cladistics 23: 301-336.
Scotland, R.[W.] & R.T. Pennington, 2000. Homology and systematics. Coding characters for phylogenetic analysis.— Syst. Assoc. spec. Vol. Ser. 58: i-vii, 1-213.
Slowinski, J.B., 1993. 'Unordered' versus 'ordered' characters.— Syst. Biol. 42: 155-165.
Struck, T.H., G. Purschke & K.M. Halanych, 2006. Phylogeny of Eunicida (Annelida) and exploring data congruence using a partition addition bootstrap alteration (PABA) approach.— Syst. Biol. 55(1): 1-20.
Vaupel Klein, J.C. von, 1984. A primer of a phylogenetic approach to the taxonomy of the genus *Euchirella* (Copepoda, Calanoida).— Crustaceana (Suppl.) 9: 1-194.
Vaupel Klein, J.C. von, 1987. Phylogenetic analysis and its foundations. In: P. Hovenkamp et al. (eds.), Systematics and evolution: a matter of diversity: 159-172.— Utrecht University Press, Utrecht.
Wiley, E.O., 1981. Phylogenetics. The theory and practice of phylogenetic systematics: i-xv, 1-439.— John Wiley, New York.
Wilkinson, M., 1992. Ordered versus unordered characters.— Cladistics 8: 375-385.
Wilkinson, M., 1995. A comparison of two methods of character construction.— Cladistics 11: 297-308.

Received: 31.iii.2009
Accepted: 4.v.2009
Edited: A.S.H. Breure