# Hierarchical representation of hypotheses of homology

**Nathanaël CAO**
CNRS, UMR 5143, Équipe Paléodiversité,
systématique et évolution des embryophytes,
Université Pierre et Marie Curie-Paris 6,
Laboratoire Paléobotanique et Paléoécologie,
12 rue Cuvier, F-75231 Paris cedex 05 (France)
ncao@snv.jussieu.fr

**René ZARAGÜETA BAGILS**
**Régine VIGNES-LEBBE**
CNRS, UMR 5143, Équipe Systématique,
recherche informatique et structuration des cladogrammes,
Université Pierre et Marie Curie-Paris 6,
Laboratoire Informatique et Systématique,
12 rue Cuvier, F-75231 Paris cedex 05 (France)
rzb@ccr.jussieu.fr
vignes@ccr.jussieu.fr

Cao N., Zaragüeta Bagils R. & Vignes-Lebbe R. 2007. — Hierarchical representation of hypotheses of homology. *Geodiversitas* 29 (1): 5-15.

ABSTRACT
Tables (usually called data matrices) are currently used to represent hypotheses of homology in cladistic analysis because they are easily readable and concise. Williams & Ebach (2006) have recently criticized their use. They argue that tables are unable to represent nested relationships and the underlying homologies. We agree with their point of view and we supplement their argumentation. We identify the formal reasons of the inadequacy of this kind of representation in systematics and cladistic biogeography. We propose a solution that has the advantage of being easily understandable and concise as well. In order to guarantee consistency in the representation proposed herein, we provide precise definitions of the concepts of "component", "character", "character-state", and "homologue".

KEY WORDS
Data matrix,
homology,
character,
classificatory structure,
hierarchy,
formalization,
cladistic analysis.

**RÉSUMÉ**
*Représentation hiérarchique des hypothèses d'homologie.*
La représentation des hypothèses d'homologie sous forme de tableaux (appelés couramment matrices de données) est classiquement utilisée en analyse cladistique grâce à la grande lisibilité et concision. Williams & Ebach (2006) en ont récemment fait la critique. Ils fondent leur argumentation sur le fait que les tableaux sont incapables de représenter des relations de parenté, c'est-à-dire les homologies sous-jacentes. Nous sommes d'accord avec leur point de vue et complétons ici leur argumentation. Nous identifions les raisons formelles de l'inadéquation de cette forme de représentation pour les hypothèses d'homologies en systématique et biogéographie cladistique. Nous proposons également une solution qui a pour avantage de rester claire et concise. Afin de garantir la cohérence de la nouvelle représentation, nous précisons les concepts de « composante », « caractère », « état de caractère » et « homologue ».

**MOTS CLÉS**
Matrice de données,
homologie,
caractère,
structure classificatoire,
hiérarchie,
formalisation,
analyse cladistique.

## INTRODUCTION

Recently, Williams & Ebach (2006) criticized the use of data matrices in phylogenetics (cladistics) and historical biogeography. Their main point is that data matrices are unable to represent hypotheses of homology. We agree with these authors. However, although they point out what we understand as the main problem in phylogenetics, i.e. the discovery of relationships with a device that cannot represent them, they do not assess the proper reasons for the impossibility of representing relationships by a two dimensional table, nor do they give a relevant alternative for coding hypotheses of homology.

Here we provide formal reasons for the failure of data matrices and present a simple, relevant alternative representation of hypotheses of homology. Before introducing our new representation, formalization of a number of concepts is required in order to clarify the rationale of cladistic analysis.

## FORMALIZATION VERSUS TERMINOLOGY

Before discussing some terminological points introduced by Williams & Ebach, we describe some formal entities and structures.

### Formalization

Formalization is the expression of a message in a language that conveys no ambiguity (Lebbe 1991). In general, the language used is mathematics or logic.

#### *Classificatory structures*

A classificatory structure is a covering set of classes (Celeux *et al.* 1989: 65; Diday 1991). This means that, given a set of individuals to classify, each individual belongs to at least one class and that there are no empty classes. There exist a number of kinds of classificatory structures. We will focus on two of them: partitions and hierarchies.

A partition is a classificatory structure that verifies an additional property: the intersection between two classes is always empty (Barthélemy & Guénoche 1988). Partitions admit a constraint of order. That is, any two classes can be sorted following one or several criteria.

A hierarchy is a classificatory structure that verifies three conditions (Barthélemy & Guénoche 1988):
1. There is a class that contains all the individuals. This class is equivalent to the root of a rooted tree.
2. There is a class that contains a single individual, for each individual. These classes are called singletons.
3. The intersection between two classes is either empty or is one of the classes; that is, classes either

do not intersect or one is included in the other.

Only classes different from the root and the singletons may be informative in systematics and biogeography.

Hierarchies and rooted trees are mathematically isomorphic (Diday 1991). For the purposes of this paper, we consider them simply equivalent. Note that hierarchies and unrooted trees are not isomorphic: an unrooted tree does not define a hierarchy.

In order to formally explain the difference between the transformational and the taxic approaches, we also need to reassess the definition of a methodological term.

*Analysis*

Cladistic analysis, in phylogenetics and biogeography, is not called analysis by chance. The analytical method has roots very deep in time (Descartes 1637). It has a precise definition, requirements and constraints. In his "*Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences, plus la Dioptrique, les Météores et la Géométrie qui sont des essais de cette méthode*", René Descartes gives the methodological principles that let him redefine correct formal reasoning. Among them, his second principle states that one has "to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution" and the third, "to conduct my thoughts in such order that, by commencing with objects the simplest and easiest to know, I might ascend by little and little, and, as it were, step by step, to the knowledge of the more complex". We can reformulate the analytical method as follows: when facing a problem that is too complex, the only way to solve it is to decompose it into a set of simpler problems. If the complex problem is how to "find the relationships of a set of taxa", the simpler problem is how to decompose it? The only way that has gained some success is to decompose it into a suite of characters. This means that the parts that compose taxic relationships are characters (Nelson 1994a), structured by hypotheses of homology.

We will not discuss here why all systematicists dealing with phylogenetics search for hierarchies of taxa. Nevertheless, if characters are the parts of taxa relationships, it follows that characters have a hierarchical structure. All systematicists use Descartes' analytical method to accomplish their task. Some method of historical biogeography (e.g., BPA, PAE, etc.) use the same *modus operandi*: biogeographical areas are decomposed into hierarchical distributions of taxa structured by a concept of area homology.

However, there is a constraint when using this method. As stated by Descartes' second principle, the analytical method uses an ascending reasoning, i.e. what is known about the whole (taxa or areas) is nothing more than the combination, addition or congruence of the solutions to the partial problems. In other words, what is known about the relationships of taxa is no more than the combination of our knowledge concerning characters. As a consequence, nothing can be learned about characters (or taxa in biogeography) from the relationships of taxa (or areas in biogeography). Nevertheless, this is exactly what parsimony analysis (and almost any other matrix-based phylogenetic or biogeographic method) does: characters lack a hierarchical structure, and their combination always leads to an unrooted tree. Outgroups are added in order to obtain the hierarchical structure and to "polarize" characters. This is a violation of the Cartesian analytical method, since something is learned about the characters (the parts) from the taxa (the whole). Nevertheless, everything that is known about taxa is nothing more that the combination or addition of what is hypothesized about their parts, i.e. characters. The need for this inconsistent, circular reasoning follows from the use of the data matrix. Indeed, the data matrix cannot represent what it is supposed to represent, i.e. hypotheses of relationships (Williams & Ebach 2006). Before discussing this point, we will clarify our use of "character".

*Representation*

Systematics is concerned with organisms and taxa, their description, discrimination, identification and phylogenetic and biogeographical relationships. The concepts of taxa, character and state belong to systematics. However, as each set of problems is conceptually different, these concepts have different meanings. There is a requirement of consistency: different meanings of the same concepts cannot be

incompatible. We will focus in the differences needed for identification and classification purposes.

Identification is the procedure that assigns a specimen to a previously established classification (Pankhurst 1978; Lebbe & Vignes 1991). Most identification methods proceed by a recursive partitioning of the initial taxonomic sampling in order to isolate a candidate by discriminating classes: its rationale is "divisive". Identification proceeds by recursively excluding parts of the candidates. On the other hand, the goal of phylogenetics is to propose a natural classification. Its rationale is "agglomerative" (Jardine & Sibson 1971; Lecointre & Le Guyader 2001). Systematics and cladistic biogeography is concerned with the initial problem of grouping two observations, organisms, taxa or distributions as being more closely related than they are to a third (Nelson 1979, 1994a; Platnick 1979; Nelson & Platnick 1981; Kitching *et al.* 1998; Humphries & Parenti 1999). The rationale of cladistics leads to the construction of hierarchies, or rooted trees, whereas in order to assign a specimen to a taxon, identification methods choose a class in a pre-established partition. Note that a particular pathway through an identification key (e.g., a printed key) may take a hierarchical appearance; however, the ensemble of the possible pathways of an identification key (the key) is not supposed to be hierarchical. The differences between both procedures and the issue of giving a particular formal structure to each of these concepts are essential to the relevance of the results. The structure of characters used in identification, i.e. descriptors plus attributes (Vignes-Lebbe 2000), and in phylogenetics, i.e. hypotheses of homology (Nelson & Platnick 1981) or relationships (Nelson 1994a, b), is different. Identification characters are best represented as partitions, while cladistic relationships are best represented as hierarchies.

Is the data matrix able to represent cladistic relationships, i.e. hierarchies? The data matrix is nothing more than a two-dimension table. It is not a matrix, in the mathematical sense. It represents, or it is supposed to represent, hypotheses of homology. Thus, ironically, the data "matrix" contains no data (Vignes 1991). Data refers to what can be measured. Therefore, it is linked to a single object and explains only what has been observed. In the data matrix, several observations are represented by the same symbol. Thus, they are not data, but concepts (Frege 1971; Lebbe 1991; Vignes 1991). Concepts express more than what has been observed, but are linked to more than the object. Hennig (1968) already made the distinction in phylogenetics between semaphoronts, from which data can be observed, and holomorphs, the concepts that have semaphoronts as instances or the "lowest taxonomic level". What does the "data" "matrix" represent? As pointed out by Williams & Ebach, the "data" "matrix" represents classes of cells coded with the same symbol. The "data" "matrix" does not represent homology, i.e. the hierarchical relationship between classes, but the classes of homologues themselves. These classes define a partition, a classificatory structure suitable for identification purposes but not for Cladistics *sensu* Williams & Ebach (2006). For this reason, no method, except three-item analysis and some versions of compatibility analysis (implicit in Le Quesne 1969, 1979) finds the target, i.e. a hierarchy of taxa. All current methods find unrooted trees, which are irrelevant, as they were never the target of investigation. In order to root the trees, the addition of several outgroups is required, and hence the analytical rationale violated. Note that even with the addition of outgroups, homologies will not necessarily have a hierarchical structure. They do not even represent partitions. Characters "polarized" by outgroup rooting, rather than having a formal structure, have a narrative one.

The influence of the "data" "matrix" and the confusion between identification and phylogenetics (but see Kitching *et al.* 1998: 27), i.e. between partitions and hierarchies, is so deep that most authors have tried to adapt their definitions of characters to justify the use of partitions. Colless (1985) considers a character as a mutually exclusive set of attributes. However, mutually exclusive sets of attributes define partitions, not hierarchies. Pimentel & Riggins (1987) define cladistic characters as linear morphoclines. Again, morphoclines or series of transformations (Hennig 1968; Pogue & Mickevich 1990) define ordered partitions, not hierarchies, with a constraint of order. Cladistic programs may use *ordered* or *additive* characters, which applies to partitions and not to hierarchies.

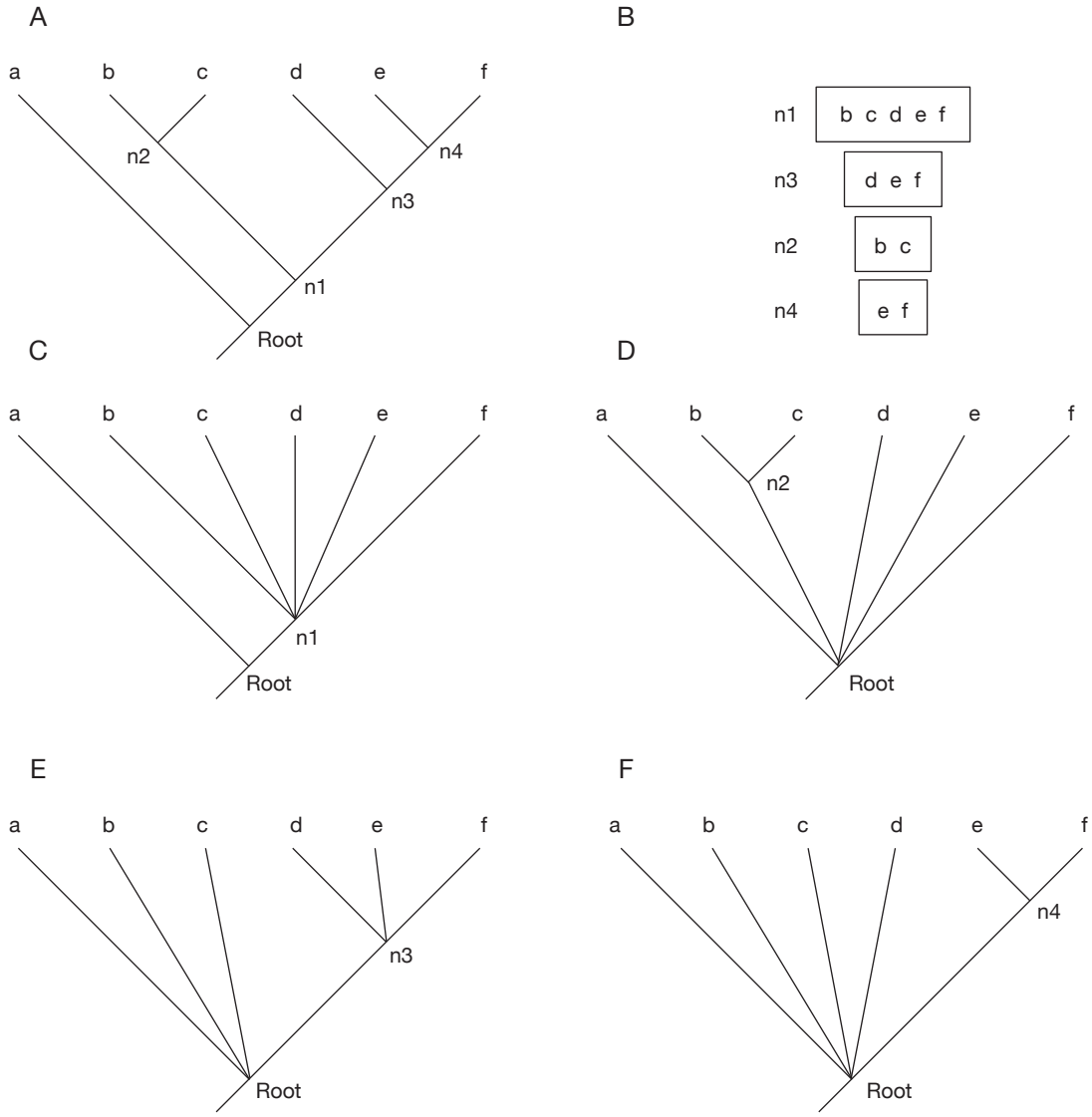



FIG. 1. — **A**, a hierarchical character of six OTUs; **B**, states of the character shown in A; **C**-**F**, components of the character showed in A; the nodes of a tree correspond to the informative classes of its isomorphic hierarchy. Thus, a rooted tree can be separated into a series of components (trees with a single internal node) or classes (states). However, when interpreted as an ensemble of classes (states), part of the information is lost. When coded as a matrix, almost all of the information is lost; for example, state n1 cannot be coded in a matrix as a state of the same character.

Darlu & Tassy (1993) define a character as "any observable attribute in an organism", a definition that does not define relationships, but belongs to a "divisive" rationale. Some recent work (Wagner 2001; Wiens 2001) gives a certain number of definitions of characters, none of which produces hierarchies. More evidence of the confusion between identification and phylogenetics is the consideration of the diagnostic properties of synapomorphies (see Kitching *et al.* 1998: 24, for examples). Some authors have

even judged the relevance of phylogenetic methods considering their "diagnostic performance" (Farris 2000). None of them has succeeded in explaining why we search for hierarchies of taxa and why the methods they use are unable to find them. If multistate characters can be decomposed into a series of components, it follows that components are also hierarchies. We redefine components as hierarchies having a single internal node (Fig. 1).

## Terminology

Williams & Ebach (2006: 410) introduce, under their section "Terminology", some new terms to describe a data matrix. They use these terms to redefine some concepts. However, for some of the concepts they use, we think that the difficulty is formal or logical rather than terminological. We focus here on the notion of "component" to illustrate it.

### *Components*

Following Nelson (1979: 5), "cladistic components are branch points. A particular branch point is defined by the branch tips (terminals, or terms) to which it leads".

Wilkinson (1994: 344) considers that "Components (= clades, monophyletic groups, holophyletic groups, clusters) are statements of relationships that apply, through inclusion or exclusion, to all taxa under consideration". The first part of the sentence (clades, monophyletic groups, holophyletic groups or clusters) seems to deal with classes, the second (statements of relationships) clearly with hierarchies (he also states that "components are the most inclusive unambiguous single statements of cladistic information that might be shared by a set of fundamental trees" and the third (inclusion or exclusion) with partitions. The rest of our paper considers components either as classes or as partitions.

Kitching *et al.* (1998: 203) define a component as "a group of taxa as determined by the branching pattern of a cladogram. For example, in a group comprising three taxa A, B, and C, where B and C are more closely related to each other than either is to A, there are two components, ABC and BC".

Williams & Humphries (2003) analyzed the asymmetry that exists between a data matrix and a cladogram. The asymmetry comes from the difference that exists between components, represented by binary characters in a data matrix, and a cladogram, represented by a single multistate character in a data matrix. This difference is only detected by the conversion of both data matrix and characters into three-item statements. The conversion from a cladogram to a data matrix, called component coding, means that components can be associated to columns of a data matrix, or character points *sensu* Williams & Ebach (2006):

"Component coding follows each internal node in a cladogram to all its tips and enters those data with scores of 1 in the matrix. All remaining taxa not supported by that node are scored with a 0. If a cladogram has more than one node, then one component is coded for each node" (Williams & Humphries 2003).

It results from their statement that components are synonyms of binary characters ("[…] three binary characters (three components) […]") and they are represented as hierarchies: A(BCDE), AB(CDE), ABC(DE) (Williams & Humphries 2003: table 2). In fact, a correct hierarchical representation should be ([A]([B][C][D][E])), ([A][B]([C][D][E])), ([A][B][C]([D][E])). The classes represented by the brackets ("[*x*]") are singletons. In order to simplify the notion, we will not represent singletons. However, we consider essential the representation of the root, i.e. the class that contains all the individuals: (A(BCDE)), (AB(CDE)), (ABC(DE)).

Williams & Humphries (2003) and Williams & Ebach (2006) extend the use of component to the data matrix. In their description of the data matrix, Williams & Ebach (2006) define components as the class of the terminal taxa that have the same "character-state". This use implies a new meaning of component: the representation of a class of a primary homology hypothesis (*sensu* de Pinna 1991) in a data matrix. Williams & Ebach (2006: 411) state that they follow Nelson (1979) when they refer to components. In their table 5 they refer to positive and negative components without any allusion to a cladogram. Nevertheless, if components are "branching points", they should be linked to cladograms.

Nelson's definition, however, contains some ambiguity. In its first part, components are considered branching points [of trees]. Branching points may

be identified by a rooted tree with a single chosen node with all the less inclusive terminals collapsed on to the chosen node (branch point) and the remaining terminals collapsed at the root. In the second part of the same definition, Nelson states that a branching point is defined by "the branch tips (terminals, or terms) to which it leads". Then, a branching point is not a tree with a single internal node but a single class, or group.

Since Nelson's definition, components have either been considered hierarchies, partitions or groups (i.e. classes) of terminal taxa. Only the former requires a cladogram . The reason for the confusion generated is, in our opinion, the intrinsic ambiguity of what the data matrix is actually representing, compared to what it is supposed to represent. In order to clarify the concept of component, we need to introduce some formal, or mathematical, classificatory notions.

*Character points*

Williams & Ebach (2006) call *character points* the columns of a data matrix. They do not call them simply characters because, from their point of view, characters are relationships, while the columns of a data matrix are not. Whereas we agree with their viewpoint that characters and homology are relationships and that the data matrix cannot represent them (see below). We also agree with Williams & Ebach that homology defines a relationship between homologues. What, then, are the homologues? Williams & Ebach (2006: 412) state that: "A cell might be more conventionally understood as equivalent to a homologue, a part of an organism", following Nelson (1994a) and Williams (2004). They also maintain that "a relationship might be usefully understood as the relation of homology, based on shared homologues" (2006: 413). Later they declare that "if our knowledge is rather better and we are able to say that the homologues of feathers are lizard scales then this might be construed as a multistate character point, with increased complexity".

Let us admit that homologues are part of organisms. Are organisms formally individuals? For Hennig organisms, seen as holomorphs, are classes (Hennig 1968: 86). He states that organisms can be seen as "the lowest taxonomic unit of a group", whereas the formal individuals, the "elements of systematics" are the semaphoronts, i.e. an organism during a very short period of its life. Are homologues parts of semaphoronts or parts of holomorphs? In order to deal with these different approaches, operational taxonomic units (OTUs) were defined (Sokal & Sneath 1963). OTUs allow an operational treatment of classes, i.e. holomorphs, as if they were formal individuals, i.e. semaphoronts. Homologues are parts of holomorphs and cannot be seen as observations or specimens. This allows the representation in a hierarchy of multistate relationships of homology and the use of any taxon as a terminal. Homologues, then, can be seen as the parts of the OTUs, whatever their nomenclatural or taxonomic rank may be, related by homology. Our definition does not contradict Williams & Ebach approach, because their taxic point can refer to any taxa, not only to organisms. If a cell represents a homologue, it can instantiate any OTU. A homologue may not only be a part of an organism but a part of any OTU.

## PROPOSITION

We have shown that characters necessarily need to have a hierarchical structure, as a requirement of consistency of the method. We have also shown that a matrix cannot represent hierarchical hypotheses of primary homology (de Pinna 1991). Can a different representation be found? Hierarchies are usually represented in phylogenetics and biogeography with Venn diagrams (e.g., Hennig 1968: figs 18, 19). Venn diagrams can be transcribed to a series of taxon names in parentheses, where each pair of parentheses indicates a class. However, using parentheses has a drawback: a character coded in 50 taxa, drawn in parentheses, will occupy a lot of space and not be easily readable. One hundred hypotheses of relationships for 50 terminal taxa may require many pages to be written, leading to an incomprehensible message. A system where the names of taxa are replaced by a code, e.g., a single letter, as in the computer program MATRIX (Nelson & Ladiges 1993), could be imagined. From a programmer's viewpoint, parsing such a file is easier. Nevertheless, understanding hypotheses of

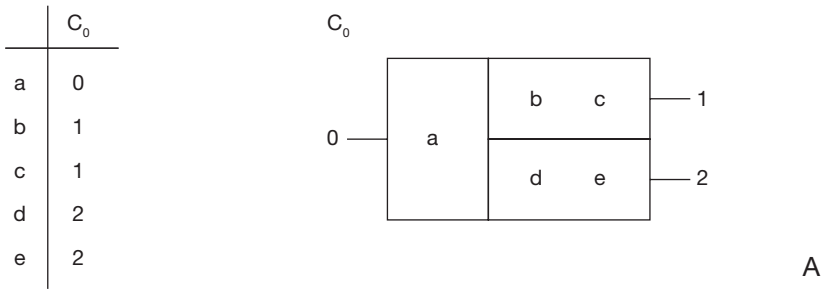
homology becomes extremely difficult with such a representation. Another way of representing primary hypotheses of homology could be by directly drawing the characters as rooted trees. This straightforward representation is easily readable and explicit. But again, whereas a data matrix can summarize a vast quantity of information (even if it is an extremely inadequate way) in a compact way, a huge number of trees will require a lot of space. We propose a new representation of hierarchies that, while correcting the inadequacy of the data matrix, preserves its advantages, mainly of concision.

Our proposition intends to guarantee all the advantages of the representation of conjectures of primary homology in the form of a table, while enhancing it by expressing the relationships among homologues. The way relationships are usually expressed is by applying the same symbols, e.g., numbers, or, for aligned molecular sequences, the symbols "A", "T"/"U", "C", "G", and "gap", to the taxa presenting them. Each symbol represents a character-state or the intension of a class (a "component" in the terminology used by Williams & Ebach 2006). The hierarchical structure of characters is built by placing each state (or "positive components") inside another, more general state or into the root (negative components *sensu* Williams & Ebach 2006). If this hierarchy of components is explicit in the data matrix, then the representation would become relevant, while remaining concise and readable by anyone who is used to reading such tables. We propose to add an additional row with the hierarchy of character states given in parentheses (Fig. 2). This simple improvement of the data matrix converts it into a representation for the analysis of hierarchical relationships. It also allows the representation of states that cannot be coded in a matrix, that is, states with no instances (Fig. 2). This is because our representation constitutes an isomorphism with hierarchies. Note that a state, as defined here, is an informative class of a hierarchical character.

Our proposition, as simple as it may appear, has some important consequences. The line we propose to add to the matrix, i.e. the hierarchy of states, represents only a hypothesis of homology, without reference to the homologues. In the original matrix, only the different classes of homologues are represented. Our solution clearly distinguishes, and explicitly separates, the extension of the structure of concepts (the matrix) and its intension (the subordination of characters states). As a result, homologies (relationships) are coded separate from groups of homologues – the components of Williams & Ebach (2006: fig. 1). With our enhancement, the systematist, rather than the algorithms or outgroups, decide on the hypotheses of homology. Such hypotheses can be changed should new knowledge become available, independently of the homologues; they can also modify the hypotheses of primary homology without changing the representation of homologues, e.g., after an analysis and a re-examination of the hypothesis. This is achieved simply by changing the hierarchy of character-states in the enhanced matrix (Fig. 2). Our notation allows the coding of complex characters, i.e. characters with two or more states. Complex hypotheses of homology cannot be simply represented in the matrix. This is the case for cladistic biogeography. Biogeographic analysis uses cladograms of taxa as the primary hypothesis of area homology (Nelson & Platnick 1981; Ebach 2003). Each hypothesis of hierarchical relationships among areas corresponds to the paralogy-free subtrees extracted from the cladograms of taxa (Nelson & Ladiges 1996; Ebach *et al.* 2005).

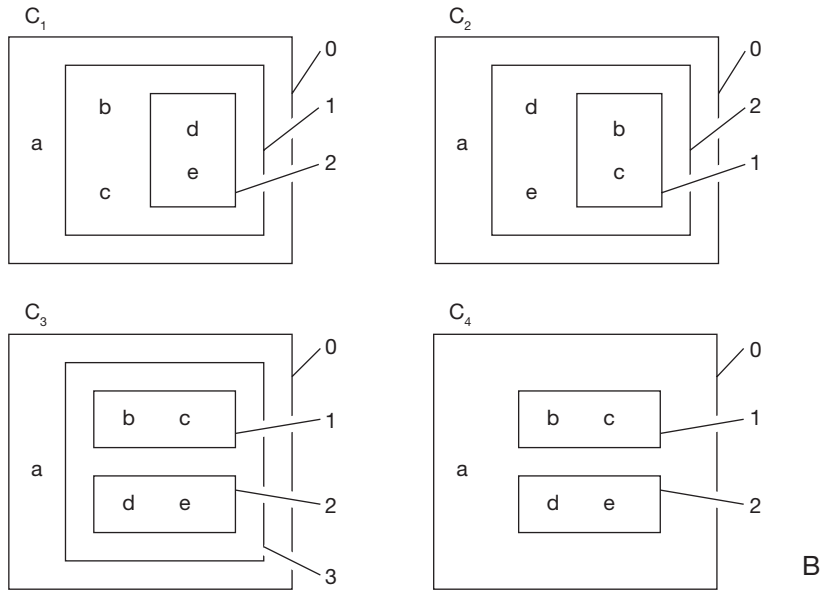
The matrix shown in Figure 2 illustrates that tables are unable to represent even relatively simple hypotheses of primary homology. The only solution for representing complex hypotheses of homology in a "data" "matrix" is to transform them into a series of additive binary characters. However, it has been shown (Nelson 1993) that coding the hypothesis as separate binary, independent, characters, is not equivalent to coding it as states of the same character, as asserted by some authors (Kluge 1993).

It is impossible to represent these complex trees in a matrix, but straightforward using the method presented herein. The hierarchical account allows the representation of any phylogenetic or biogeographic hypotheses. In biogeography, the impossibility of representing this kind of information has led to the development of methods, such as BPA (Wiley 1988), that are defective (Ebach & Humphries 2002). However, with the notation we propose




A

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| | (0(1(2))) | (0(2(1))) | (0(3(1)(2))) | (0(1)(2)) |
| a | 0 | 0 | 0 | 0 |
| b | 1 | 1 | 1 | 1 |
| c | 1 | 1 | 1 | 1 |
| d | 2 | 2 | 2 | 2 |
| e | 2 | 2 | 2 | 2 |




B

FIG. 2. — The inadequacy of the matrix for representing hierarchical relationships: **A**, the usual matrix coding where characters are represented as partitions (see the partition $C_0$ at right); this poor representation introduces a high amount of ambiguity often resulting in a parsimony analysis, in a high number of most parsimonious trees leading to unresolved relationships when a consensus method is applied; **B**, the formal coding proposed here: characters are represented as hierarchies; the hierarchical structure is given by the upper line, with the character-states in parentheses. Note that character $C_0$ is extremely ambiguous: four hierarchical representations of the same character are shown (characters $C_1$ to $C_4$ in B) implying very different assumptions concerning homology.

the source trees are coded directly, independent from the number of areas of distribution and the complexity of relationships.

## CONCLUSION

Cladistics is an analytical method of reasoning. There is a widely accepted consensus that the best representation of phylogenetic and biogeographic relationships is hierarchical. Our formalization of classificatory structures, as is generally done in the knowledge analysis field, and the analytical method, following Descartes, implies that characters (or relationships of distributions of taxa) must have a hierarchical structure in order to be consistent.

Cladistics can be seen as a method of combination of hierarchies. However, as shown by Williams & Ebach (2006), the "data" "matrix" is an inadequate way of representing hierarchies. We have shown that this is the main reason that explains why all current methods do not find hierarchies of taxa but, instead, unrooted trees.

Our proposal allows the representation of hierarchical hypotheses of taxon or area homology. It allows the explicit representation of the extension and the intension of the hypothesis, the latter being absent from current "data" "matrix" representation. Our proposition is of no help to systematists until it can be implemented in a computer program that is able to read and "understand" (i.e. that is able to preserve the sense of) the new representation. We are currently developing such a program that will manipulate, decompose and combine hierarchies or rooted trees using three-item analysis, the only cladistic method that can use our representation, for it is the only method to explicitly require hierarchical hypotheses of homology. In short, we will be able to go out of the matrix.

**Acknowledgements**
D. M. Williams and L. Parenti contributed to the improvement of the text. We acknowledge our colleagues in the SRISC (Systematics, Research in Informatics and Structure of Cladograms) and in the Systematics and Evolution of Embryophyte of the UMR 5143 for useful discussions. We specially acknowledge D. M. Williams and M. C. Ebach for identifying a fundamental problem in phylogenetic systematics and cladistic biogeography.

## REFERENCES

BARTHÉLEMY J.-P. & GUÉNOCHE A. 1988. — *Les arbres et les représentations des proximités*. Masson, Paris, 239 p.

CELEUX G., DIDAY E., GOVAERT G., LECHEVALLIER Y. & RALAMBONDRAINY H. 1989. — *Classification automatique des données*. Dunod Informatique; Bordas, Paris, 285 p.

COLLESS D. H. 1985. — On "character" and related terms. *Systematic Zoology* 34: 229-233.

DARLU P. & TASSY P. 1993. — *La reconstruction phylogénétique. Concepts et méthodes*. Masson, Paris, 245 p.

DE PINNA M. C. C. 1991. — Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7: 367-394.

DESCARTES R. 1637. — *Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences, plus la Dioptrique, les Météores et la Géométrie qui sont des essais de cette méthode*. De l'imprimerie de I. Maire, Leyde, 79 p.

DIDAY E. 1991. — Des objets de l'analyse des données à ceux de l'analyse des connaissances, *in* KODRATOFF Y. & DIDAY E. (eds), *Induction symbolique numérique à partir de données.* Cépadués, Toulouse, 460 p.

EBACH M. C. 2003. — Area cladistics. *Biologist* 50: 169-172.

EBACH M. C. & HUMPHRIES C. J. 2002. — Cladistic biogeography and the art of discovery. *Journal of Biogeography* 29: 427-444.

EBACH M. C., NEWMAN R. A., WILLIAMS D. W. & WALSH S. A. 2005. — Assumption 2: opaque to intuition? *Journal of Biogeography* 32: 781-787.

FARRIS J. S. 2000. — Diagnostic efficiency of three-taxon analysis. *Cladistics* 16: 403-410.

FREGE G. 1971. — *Écrits logiques et philosophiques*. Éditions du Seuil, Paris, 234 p.

HENNIG W. 1968. — *Elementos de una sistemática filogenética*. Editorial Universitaria de Buenos Aires, Buenos Aires, 353 p.

HUMPHRIES C. J. & PARENTI L. R. 1999. — *Cladistic Biogeography: Interpreting Patterns of Plant and Animal Distributions*. Second edition. Oxford University Press, Oxford, 187 p.

JARDINE N. & SIBSON D. 1971. — *Mathematical Taxonomy*. Wiley & Sons, Ltd., London, 286 p.

KITCHING I. J., FOREY P. L., HUMPHRIES C. J. & WILLIAMS D. W. 1998. — *Cladistics: the Theory and*

*Practice of Parsimony Analysis.* Second edition. The Oxford University Press, Oxford, 228 p.

KLUGE A. G. 1993. — Three-taxon transformation in phylogenetic inference: ambiguity and distortion as regards explanatory power. *Cladistics* 9: 246-259.

LE QUESNE W. J. 1969. — A method of selection of characters in numerical taxonomy. *Systematic Zoology* 18: 201-205.

LE QUESNE W. J. 1979. — Compatibility analysis and the uniquely derived character concept. *Systematic Zoology* 28: 92-94.

LEBBE J. 1991. — *Représentation des concepts en biologie et médecine. Introduction à l'analyse des connaissances et à l'identification assistée par ordinateur*. Thèse de doctorat, spécialité Sciences de la vie, Université Pierre et Marie Curie-Paris 6, Paris, France, 281 p.

LEBBE J. & VIGNES R. 1991. — Génération de graphes d'identification à partir de descriptions de concepts, *in* KODRATOFF Y. & DIDAY E. (eds), *Induction symbolique et numérique*. Cépadués, Toulouse: 193-239.

LECOINTRE G. & LE GUYADER H. 2001. — *Classification phylogénétique du vivant*. Belin, Paris, 543 p.

NELSON G. J. 1979. — Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's *Familles des Plantes* (1763-1764). *Systematic Zoology* 28: 1-21.

NELSON G. J. 1993. — Reply. *Cladistics* 9: 261-265.

NELSON G. J. 1994a. — Homology and systematics, *in* HALL B. K. (ed.), *The Hierarchical Basis of Comparative Biology*. Academic Press, San Diego: 101-149.

NELSON G. J. 1994b. — La systématique et l'homologie, *in* TASSY P. & LELIÈVRE H. (eds), *Caractères*. Société française de Systématique, Paris: 5-28.

NELSON G. J. & LADIGES P. Y. 1993. — *Matrix 1.0.* Published by the authors, New York; Melbourne.

NELSON G. J. & LADIGES P. Y. 1996. — Paralogy in cladistic biogeography and analysis of paralogy-free subtrees. *American Museum Novitates* 3167: 1-58.

NELSON G. J. & PLATNICK N. I. 1981. — *Systematics and Biogeography: Cladistics and Vicariance*. Columbia University Press, New York, 567 p.

PANKHURST R. J. 1978. — *Biological Identification. The Principles and Practice of Identification Methods in Biology*. Edward Arnold, London, 104 p.

PIMENTEL R. A. & RIGGINS R. 1987. — The nature of cladistic data. *Cladistics* 3: 201-209.

PLATNICK N. I. 1979. — Philosophy and the transformation of cladistics. *Systematic Zoology* 28: 537-546.

POGUE M. G. & MICKEVICH M. F. 1990. — Character definitions and character state delineation: the bête noire of phylogenetic inference. *Cladistics* 6: 319-361.

SOKAL R. R. & SNEATH P. H. A. 1963. — *Principles of Numerical Taxonomy*. W. H. Freeman, San Francisco, 573 p.

VIGNES R. 1991. — *Caractérisation automatique de groupes biologiques.* Thèse de doctorat, Université Pierre et Marie Curie-Paris 6, Paris, France, 268 p.

VIGNES-LEBBE R. 2000. — Caractère pour le biologiste, caractère pour l'informaticien, *in* BARRIEL V. & BOURGOIN T. (eds), *Biosystema 18*. Société française de Systématique, Paris: 61-70.

WAGNER G. P. (ed.) 2001. — *The Character Concept in Evolutionary Biology*. Academic Press, San Diego, 622 p.

WIENS J. J. 2001. — Character analysis in morphological phylogenetics: problems and solutions. *Systematic Biology* 50: 689-699.

WILEY E. O. 1988. — Parsimony analysis and vicariance biogeography. *Systematic Zoology* 37: 271-290.

WILKINSON M. 1994. — Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* 43: 343-368.

WILLIAMS D. W. 2004. — Homologues and homology, phenetics and cladistics: 150 years of progress, *in* WILLIAMS D. W. & FOREY P. L. (eds), *Milestones in Systematics*. CRC, The Systematics Association, London: 191-224.

WILLIAMS D. W. & EBACH M. C. 2006. — The data matrix. *Geodiversitas* 28 (3): 409-420.

WILLIAMS D. W. & HUMPHRIES C. J. 2003. — Component coding, three-item coding, and consensus methods. *Systematic Biology* 52: 259-271.

*Submitted on 20 October 2006;*
*accepted on 25 January 2007.*