The data matrix

David M. WILLIAMS
Department of Botany, The Natural History Museum,
Cromwell Road, London, SW7 5BD (United Kingdom)

Malte C. EBACH
Laboratoire Informatique et Systématique (LIS),
Université Pierre et Marie Curie, UMR 5143,
Paléobiodiversité et Paléoenvironnements, Équipe Systématique,
Recherche informatique et Structuration des Cladogrammes,
12 rue Cuvier, F-75005 Paris (France)
ebach@ccr.jussieu.fr

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

ABSTRACT
The data matrix can be generally viewed in three different ways for systematics and biogeography, which we refer to as *phylo-phenetics*, *phylogenetic systematics* (transformational) and *Cladistic*. Each views the matrix as a combination of manipulating points (taxa, areas and characters) and cells (character states), expressed in a table. All current methods, except Cladistics, treat the transformations or switches between cells and points as the basis for discovering and expressing relationships. We consider that Cladistics, however, treats a relationship between three cells, or points, as the smallest unit. We suggest that the difference between all three methods lies in the theory and application of the terms homologue and homology. It is shown that for most methods the data matrix is simply a phenetic device for optimising homologues rather than determining homologies and discovering relationships.

KEY WORDS
Cladistics,
data matrix,
homologues,
homology,
homoplasy,
paralogy.

RÉSUMÉ
La matrice de données.
La matrice de données peut être interprétée de trois manières différentes en systématique et biogéographie, baptisées ici *phylo-phénétique*, *systématique phylogénétique* (transformationnelle) et *Cladistique*. Chacune de ces interprétations considère la matrice comme une combinaison de lignes (taxons, aires et

caractères) et de cellules (états de caractère), exprimés sous forme de tableau. Toutes les méthodes habituelles, sauf la Cladistique, utilisent les transformations ou les changements entre les cellules comme source pour la découverte et la représentation des relations de parenté. Nous montrons que la Cladistique, en revanche, utilise les relations entre trois cellules comme l'unité élémentaire de relation. Nous suggérons que la différence entre ces trois méthodes se trouve dans la théorie et la pratique concernant les notions d'homologue et d'homologie. Nous montrons ici que, dans les deux premières interprétations citées, la matrice est simplement un outil phénétique destiné à optimiser des arrangements d'homologues plutôt qu'à déterminer des homologies et à découvrir des relations de parenté.

MOTS CLÉS
Cladistique,
matrice de données,
homologues,
homologie,
homoplasie,
paralogie.

INTRODUCTION

"Unfortunately, no one can be told what the Matrix is. You have to see it for yourself" (Morpheus in "*The Matrix*", Warner Bros. 1999).

"We managed to get rid of one pernicious black box – evolutionary systematics – but we've replaced it with another black box – the data matrix" (Patterson 1997).

Today, it is rare to see a taxonomic, systematic, phylogenetic or biogeographic study being undertaken without first constructing a data matrix. For the usual morphological matrix, entries are derived from reasoned determination of the similarities and differences between the parts of various organisms (see Rieppel & Kearney 2002). For the usual molecular matrix, composed of strings of nucleotides, the entries might be considered "ready-made" and alignment programs have been designed simply to order the sequences in the most optimal way (e.g., Wheeler 2002). These kinds of matrices relate to those described in Sokal & Sneath (1963: 123), who wrote: "We adopt the convention used in psychology of arranging data for such an undertaking in the form of an $n \times t$ matrix whose t columns represent the t fundamental entities to be grouped on the basis of resemblances and whose n rows are n unit characters". Later, Sneath & Sokal (1973: 114) summarised their earlier statement by describing a matrix as a table "whose t columns represent the OTU's [terminals] to be grouped on the basis of resemblances and whose n rows are n unit characters". These kinds of matrices are often said to contain raw data (the "unit characters"), in the sense that the entries are largely derived from some form of empirical investigation (Rieppel & Kearney 2002). In this sense, they relate to the tables of characters many taxonomists provide in their studies to allow the reader to contrast the various parts of specimens examined, and identify those parts considered similar from those parts considered different.

The data matrix used in phenetics, phylogenetic systematics, Cladistics and biogeography may be viewed in three fundamentally different ways, each corresponding to a different theory and methodology for the interpretation of the entries.

TERMINOLOGY

Consider the following data matrix of six characters (1-6) and four taxa (A-D; Fig. 1). The matrix may be defined as consisting as several interconnected elements. We introduce a new terminology in order to view the matrix as a functional unit that is used to build branching diagrams and to prevent confusion with existing terms (i.e. character-states, characters, etc.). The taxon rows and character columns are made up of data, herein termed *taxic points* and *character points* respectively, with the various character states represented by

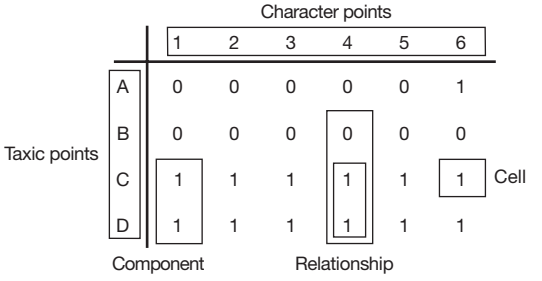

FIG. 1. — Diagram illustrating the terminology used for data matrices: characters and taxa (or areas) may be termed *character points* and *taxic points*, respectively. *Cells* indicate a particular entry, relative to a character and taxic point; *components* are determined by "matching" cells; *relationships* are determined by both the "matching" of similar cells with those that differ in the same column. A cell might be more conventionally understood as equivalent to a homologue, a part of an organism (Nelson 1989, 1994; Williams 2004; Williams & Humphries 2004); a component might be more conventionally understood as equivalent to a group based on shared homologues; a relationship might be usefully understood as the relation of homology determined from the homologues (Nelson 1989, 1994; Williams 2004; Williams & Humphries 2004).

A = 1:0, 2:0, 3:0, 4:0, 5:0, 6:1

B = 1:0, 2:0, 3:0, 4:0, 5:0, 6:0

C = 1:1, 2:1, 3:1, 4:1, 5:1, 6:1

D = 1:1, 2:1, 3:1, 4:1, 5:1, 6:1

1 = A:0, B:0, C:1, D:1

2 = A:0, B:0, C:1, D:1

3 = A:0, B:0, C:1, D:1

4 = A:0, B:0, C:1, D:1

5 = A:0, B:0, C:1, D:1

6 = A:1, B:0, C:1, D:1

FIG. 2. — Upper and lower boxes list all character and taxic points from Figure 1 respectively.

the morphological 0s and 1s and the molecular nucleotides (A's, C's, G's and T's), herein termed *cells*. Thus, in Figure 1, *character point* 1 and *taxic point* A are represented by the *cell* 0. In addition, each column (character point) implies a grouping among taxa (taxic point), which we refer to as *components* (*sensu* Nelson 1979: 3, "Cladistic components are branch points. A particular branch point is defined by the branch tips (terminals or terms) to which it leads"). Thus, for character point 1 and taxic points A-D, there is a group CD, represented by its component. In addition, character points imply certain relationships. For character point 1 for taxic points A-D, the relationship is AB(CD); that is, C and D are more closely related to each other than either are to A or B. This statement of relationship can be further simplified into two 3-item statements, A(CD) and B(CD). Here we mean that the data suggest the relationship to be true, whereas, with more data, the relationship may indeed be found false. Thus, from the three kinds of entries in a matrix (cells, components, relationships), we view the *relationship*, rather than the cell or component, as the basic unit of systematics.

In a world in which species and populations (herein *taxa*) are not viewed as wholes, that is as taxa that consist of the inter-relationships of their characteristics (homologies) and their relationships to other things (perception) (see Brady 1998), but rather as collections of genotypic and phenotypic characteristics, the data matrix functions (usually) as a simple table of binary variables, in the case of non-molecular data. The rows of cells that represent each point are the non-molecular or molecular (DNA) characters derived from specimens that are said to represent a property or an attribute of each taxon. Each row does not attempt to represent any kind of relationship, or even grouping, but rather a simple coding of taxic points (Figure 2 is a list of all character and taxic points that can be derived from Figure 1).

For the purposes of further analysing the matrix, cells between rows may be treated as *switches*, having the ability to change from one state to another (Figure 3A lists all possible switches for the data in Figure 1). As cells represent unique individual codings and can potentially "metamorphose", that is, turn into other cells, they can only *switch states* between different *character points*. In this two-dimensional

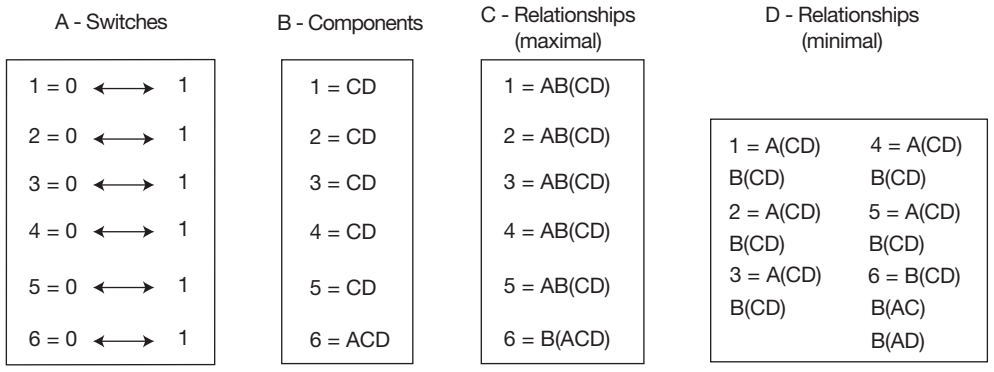

FIG. 3. — Matrix "operations": **A**, list of all switches between character points derivable from Figure 1; **B**, list of all components derivable from Figure 1; **C**, list of all possible maximal relationships (see text) derivable from Figure 1; **D**, list of all possible minimal relationships (see text) derivable from Figure 1.

data matrix, cell 0 or A (for nucleotides) can only replace cell 1 or G (for nucleotides), respectively.

A set of aligned DNA sequences and phenetic morphological matrices, when treated as two-dimensional structures, are herein termed *phylo-phenetic*.

From a transformational point of view, the matrix is effectively viewed as a three-dimensional table where the cells are treated as hypothesised (possible) transformations between character points. Thus character point 1 could actually imply a cell of 1:0 that may transform into character point 1, cell 1:1. In reality, a two-dimensional $q \times r$ matrix can only represent one state at a time, e.g., 0 or 1, and not both at the same time. Only two identical cells are needed to determine a component (Figure 3B lists the possible components for the data in Figure 1). The transformation between cells determines the taxic component, not the actual "grouping" of cells. Morphological data matrices that use optimisation procedures (such as parsimony) are herein termed *transformational*.

Both the phylo-phenetic and transformational matrices are usually processed with parsimony, maximum likelihood (ML) or phenetic computer programs, to determine optimal groupings.

Consider the cells to be points, rather than what constitutes a point. Taxic points C and D are related to each other by cells 1:1, its component. The cell points determine a particular group (cell component) by their presence, rather than by any conceivable state prior to transformation (e.g., 1:0 to 1:1, etc.). The taxic component (CD) is based on the cell component (11), and not on any switch because the cell points are equivalent to scored states, not switches or transformations of each other (Fig. 3B). The actual changes between cell points are not considered to be directional (see Brady 1998). That is, they are not grouped according to any imagined direction of transformation.

However, the component (11) is only a "part" of the data (the 00 is a part too), hence to express a specific relationship, both aspects of the data require consideration. Thus, the *relationship* is AB(CD). This relationship can be thought of as "maximal" (after Nelson & Platnick 1981), in that it includes all the taxic points (A-D). The simplest unit involves only three taxa, thus one might refer to these as "minimal" relationships, such as A(CD) and B(CD) (after Nelson & Platnick 1991) (Figure 3C lists all possible "maximal" relationships for the data in Figure 1; Figure 3D lists all possible "minimal" relationships for the data in Figure 1). Morphological and DNA matrices that follow the cell component approach, as expressed in terms of relationships, are herein termed *Cladistic* (we adopt the upper case to distinguish it from the phylogenetic systematics version, often called cladistics).

A cell might be more conventionally understood as equivalent to a homologue, a part of an organism (Nelson 1989, 1994; Williams 2004; Williams &

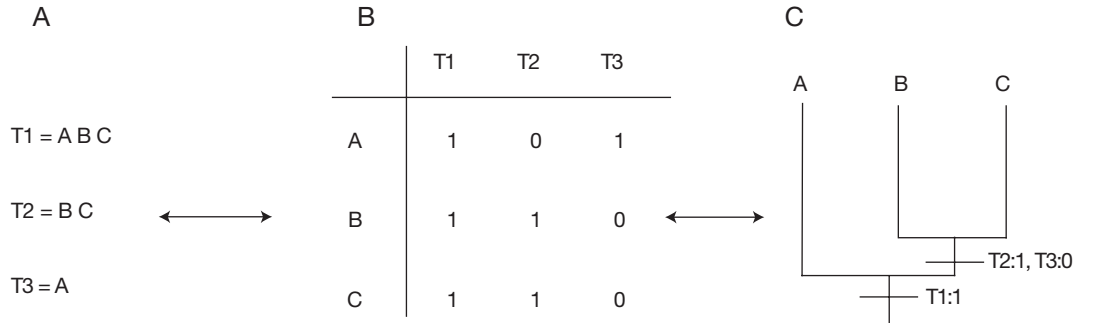

FIG. 4. — An example of a phylo-phenetic approach used in biogeography, similar to Parsimony Analysis of Endemicity (PAE) analysis: **A**, list of taxa T1-T3 and the areas they occupy (A, B and C); **B**, matrix in which similarities are grouped; **C**, the resulting phylophenogram in which areas B and C are grouped together by the component BC derived from T2 and (possibly) T3. Note that taxa T2 and T3 do not occur in the same areas and therefore do not share any direct relationship.

Humphries 2004); an informative internal component might be more conventionally understood as equivalent to a group (Nelson 1979), based on shared homologues; a relationship might be usefully understood as the relation of homology, based on shared homologues (Nelson 1989, 1994; Williams 2004; Williams & Humphries 2004).

REPRESENTING RELATIONSHIPS

The task of systematic and biogeographical methods is to discover the relationships among taxa and areas, respectively. The scoring of character-states in a matrix has become common practice in systematics as a basis for determining relationships, but in cladistic biogeography the relationships (the taxic cladogram) exist prior to analysis. For example consider an areagram, AB(CD) based on the taxon cladogram T1(T2, T3). The widespread basal taxon (T1) occurs in two areas indicated by an underscore, namely AB. The areagram is found by replacing the taxa with the areas in which they are found (Rosen 1978). Therefore the cladogram T1(T2, T3), has T1 occurring in two areas (A and B), T2 occurring only in area C and T3 occurring only in area D. The single areagram therefore contains only two relationships A(CD) and B(CD).

There are three ways in which the relationship of widespread terminal nodes can be resolved.

The problem may be treated phylo-phenetically by ignoring the known relationships (cladogram) and coding the presence and absence of taxa T1-T3 in the areas A to C (Fig. 4). In phylo-phenetic matrices the absences and presences of cells are treated as potential switches. Absence in this sense is not a state of presence and *vice versa*. If the taxon is absent then it is simply not there and if it is present the cell is simply scored to represent that fact. One method that utilises phylo-phenetic matrices is Parsimony Analysis of Endemicity (PAE) (Rosen 1988a, b but see Rosen 1985: fig. 4). PAE discovers groups based on the similarity of switches and not on relationships (i.e. T2 and T3 are not present in the same areas).

The transformational approach codes absence (cell 0) and presence (cell 1) of areas for all nodes. The approach differs from PAE, as the relationships between areas (areagram) are already known. The matrix resulting from a transformational approach will have an extra set of points representing the internal nodes from the areagram. The internal nodes function as components; that is, each component contains all areas that are represented at the terminals. The methods commonly associated with transformational approaches in biogeography are Assumption 0 (Zandee & Roos 1987), all versions of Brooks Parsimony Analysis (BPA, mBPA, secondary BPA, etc.) (see Brooks 1981 onwards, especially Brooks *et al.* 2004) and DIVA (Ronquist 1997). The Assumption 0, BPA or DIVA matrices

adjust the transformations between absence and presence thus generating a hypothesis based on optimization, in this case implemented with either parsimony or compatibility.

The areagram A(B, CD) is coded as four taxic points and five character points (Fig. 5) (with the underlined areas indicating widespread terminal components). The result, A(B(CD)), when compared with the original areagram, contains the extra component (CD), as coding protocols do not allow more than one taxon to occur on any single component.

BPA and Assumption 0 appear to resolve widespread taxa, but in fact the matrix can only assign one taxon per terminal. The optimisation determines the grouping with the areagram. Nelson's (1984: 288, fig. 14.12) opaque example, represented by the areagram AC(C, ABC), cannot be resolved with the transformation paradigm. BPA finds the areagram B(AC), a relationship not represented in the original areagram. The generated result is caused by both the inflexible matrix and optimisation between two non-transformational states, in this case the absence and presence of taxa (see Ebach *et al.* 2003).

The Cladistic approach does not actually need a data matrix as the relationships are already stated; only the terminals remain unresolved. Therefore the areagram A(B, CD) is no more than the combination of B(CD) and A(CD). The combination of A(CD) and B(CD) allows a result to be discovered, namely AB(CD). The Cladistic approach discovers the only relationship present in Nelson's opaque example by breaking down the areagram into the smallest units of relationship, termed *area homologies* (Humphries & Ebach 2004). AC(C, ABC) contains only one area relationship and therefore only one area homology, namely A(BC).

If a matrix based on phylo-phenetic or transformational approaches is unable to discover relationships, then how can the systematist and biogeographer be sure that the matrix truly represents relationships? There is no denying that the matrix represents the "data", but as shown by PAE and BPA the "data" are purely two dimensional and concern only the switching and transformation between absence and presence of cells, as determined by particular computer programs, not the actual relationships. In biogeographical methods it is clear that the matrix does not represent relationships and that existing relationships, as in Nelson's opaque example, cannot be extracted by using crude (some would say naive) binary coding methods. The same could be said for systematic matrices that represent genetic and morphological data.

INSIDE THE MATRIX

From the perspective of the usual kind of "Cladistic" data matrix, the entries are said to be "shared derived characters". "Shared derived characters" indicate (or hypothesize) particular monophyletic groups. Consider matrix 1 (Table 1; after Nelson 1996: fig. 1, 2004: fig. 6.4; see also Williams & Ebach 2005). A-D are taxa with three character points. If the entries are evidence of relationships, indicated by the 1 cells ("shared and derived"), and the lack of evidence, indicated by the 0 cells ("shared and primitive"), then conclusions might be straightforward. Here it would seem more appropriate to re-write the matrix characters in tree-form representing the relationships exactly, such that the three "characters" are AD(BC), AC(BD) and AB(CD) (Siebert & Williams 1999), which, when combined, unambiguously provide evidence for the solution A(BCD) (Nelson 1996; see also Williams & Ebach 2005). The re-writing effectively converts an ordinary (phenetic) matrix into a Cladistic matrix. Therefore, even when the entries in a matrix are said to be "shared derived characters", they actually function as phenetic characters, devoid (apparently) of any meaning.

From the perspective of the usual kind of "Cladistic" matrix, there may be several ways of representing the data.

The matrix in Table 1 consists of three character points. If the components considered as grouping statements (see Fig. 1), the data consist of three positive (1 cells) components (BC, BD and CD) and three negative components (AD, AC, AB). The positive components might be interpreted as specific homology (relationship) statements, in that BC are grouped relative to A, BD are grouped relative to A and CD are grouped relative to A, based on the evidence (the 1 cells). The negative components are the residual statements related by common shared

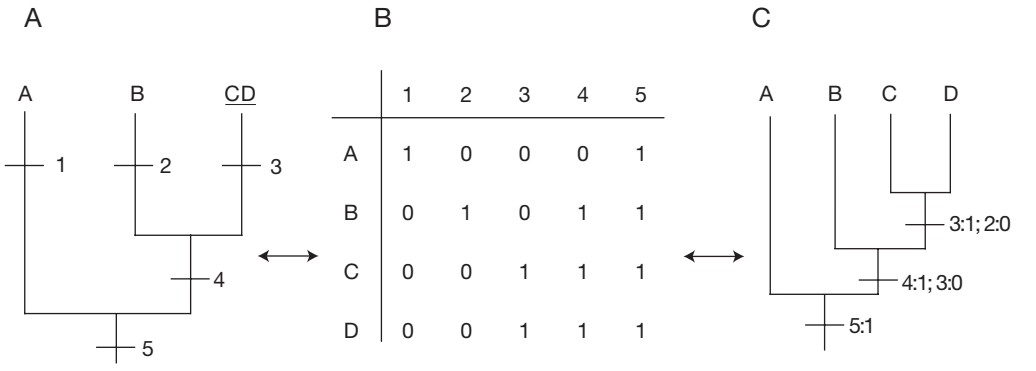

FIG. 5. — An example of a transformational approach in biogeography, similar to Brooks Parsimony Analysis (BPA). Both internal and terminal nodes of an areagram are coded for presence and absence of areas A-D; **A**, cladogram of relationships among areas A-D. Nodes 1-3 specify terminals, nodes 4-5 specify taxic relationships; **B**, the data matrix derived from all five nodes of the cladogram; **C**, the cladogram found after parsimony analysis of the matrix in B. The cladogram specifies that C is more closely related to D than any of the other areas even though there is no evidence of that relationship in the original areagram.

absence. Thus, negative components are not grouping or homology statements at all. Of the 12 data entries, only six constitute evidence. As components, there are only three items of evidence: BC, BD, CD. As relationships there are maximally three, AD(BC), AC(BD) and AB(CD) (Nelson & Platnick 1981) and minimally six, A(BC), D(BC), A(BD), C(BD), A(CD), and B(CD) (Nelson & Platnick 1991).

In terms of transformations, to polarize the characters one might consider the use of either outgroups or ontogeny to provide appropriate values. Values may be either positive or negative. The matrix in Table 2 includes an all-zero outgroup. The data consist of three positive 1 cell components: BC, BD and CD, which collectively relate to A through A(BCD), as BC + BD + CD = A(BCD).

The matrix in Table 3 has an outgroup with all positive values, 1 cells, effectively rendering the 0 cell (absent) entries "informative". The data consist of three "positive" 0 cell components: AD, AC and AB, which together relate no taxa, as AD + AC + AB = ABCD.

The matrix in Table 4 consists of three characters "polarised" by an outgroup comprised entirely of question marks (literally meaning either 0 or 1 cells). These data consist of three positive (1 cells) components (BC, BD and CD) and three negative components (AD, AC, AB). Together the positive components relate A(BCD), as BC + BD + CD = A(BCD) and the negative components relate AD + AC + AB = (ABCD), together relating A(BCD) + (ABCD) = A(BCD).

The matrix in Table 5 consists of three multistate characters "polarised" by an outgroup with all 0 cells. The difference between multistate character points and binary character points is in the "uninformative" portion, the 0 cell, as it is now deemed informative. Consider feathers as one of the character points. The 0 cells might be taken to mean either feathers lost, or feathers absent. "Feathers absent" relates the data (feathers) to all life that lacks feathers, which can be represented by a binary character. This, in turn, might suggest that we can postulate a taxon with feathers (birds, for example) but do not know of its relationships to the rest of life more precisely. 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. The increased complexity suggests that we may discover two taxa, birds (with feathers) and birds + lizards (lizard scales + feathers), rather than birds and life minus birds, the latter a non-group. Yet the complexity might extend to ignorance in not knowing which (the feathers or the lizard scales) comprises the subset and therefore the homologue relative to the set lizard scales plus feathers. In this case it might be seen that both lizard scales and feathers are the data, in as

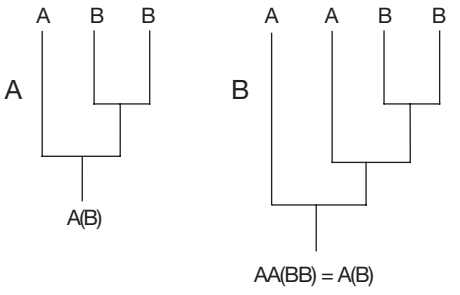

FIG. 6. — Diagrammatic representation of two uninformative areagrams specifying no relationships; **A**, cladogram A(BB); **B**, Cladogram A(A(BB)).

much as they are real observations. Here, the data are reduced to their phenetic aspect and lack any suggestion of relationships. In this case, these data consist of six positive (1 and 2 cells) components (BC, BD, CD, and AD, AC, AB). Together they do not relate any taxic points at all.

The matrix in Table 6 consists of a binary representation of the multistate character points from the matrix in Table 5, or binary representation of paired homologues (Scotland 2000). Thus if character point 1 is feathers and lizard scales, character point 1a is the "feather" homologue and character point 1b the "lizard scale" homologue. These data yield six positive and six negative components. In other words, each possible homology statement is supported by both a positive and a negative component. Together they relate no taxa at all.

With the addition of an all-zero outgroup (Table 7) these data also yield six positive components, identical to the multistate characters. Together they relate no taxic points at all.

The results are summarized in Table 8.

HOMOLOGUES, HOMOPLASY AND PARALOGY

The matrix entries, by themselves, do not yield homology by way of informative relationships (Fig. 1). Instead, the matrix simply represents the homologues, along with any potential switches (molecular data) and transformations (morphological data), invoked by one or more methods and their respective computer programmes.

Homologues are the independent cells (Fig. 1), the parts (character states) that constitute a character or taxic point. The cells act as homologues within a character or taxon that may be grouped with similar cells (phylo-phenetics) or as transformations of different cells (synapomorphies). In each case the similarity of the cells is measured either by similar taxic content (synapomorphy of two sister taxa) or by the number of cell transformations per node (steps). Phylo-phenetic and transformational methods operate under the principle that the similarity of homologues generated by some transformational model constitutes a mark of relationship. In either case, the character points are forced to comply with a particular model of cell switching or cell transformation (as implemented in various computer programmes). Each character point provides a different set of similarities, switches and transformations; with the absence of confirmed similarity (synapomorphy), the homologues are still placed on the tree or cladogram as *homoplasies*. Therefore, the probability of similarity between cells and points (branch lengths, etc.) makes homoplasy a statistical issue rather than one of relationship, or even non-relationship, as evidenced by claims that "Homoplasy *increases* phylogenetic structure" (Källersjö *et al.* 1999). Ultimately, the relationships are the result of the program rather than part of the data, imposed rather than discovered.

CONCLUSION

The Cladistic approach is concerned with homology (monophyly or cell of the character point relationships). As cells are treated as indicators of relationships and not as individual switches or transformations, a series of cell statements determines the relationship. A character point that contains A 1:0, B 1:1 and C 1:1, for instance, is not understood as a transformation between cells 0 and 1, but as the relationship implied between taxic points B and C relative to A, information derived from the data. If there is conflict in another character point, for instance A 2:1, B 2:1 and C: 2:0, the Cladis-

tic approach focuses on the relationship between taxic points A and B. Under phylo-phenetics and transformation, both characters points cannot be "true" and one or the other will eventually become homoplastic. Under the Cladistic approach both character points above imply relationships, even if the two overall do not imply the same relationship. Homoplasy, then, is artefact and does not exist using the Cladistic approach to represent data. However, paralogy certainly does contain homoplasy (*sensu* Nelson & Ladiges 1996; Ebach & Williams 2004).

Similarities and relationships clearly are two different things. Consider the two areagrams A(B,B) and A(A(B,B)) (Fig. 6). Although the areagrams are similar and may both be reduced to A(B), no specific relationship is implied. After all, A(B) can only mean A is related to B in some unspecific way. The relationship between cells of the same character point however also resembles a paralogous area relationship, such as 0(11). The relationships between cells cannot be paralogous as each is a unique instance or occurrence rather than a redundant repetition as is the case with paralogous areas.

The transformational and phylo-phenetic approaches are based on probabilities of transformations and switches based on either the ML (maximum likelihood) model or the "parsimony" model. Yet, parsimony, when combined with transformations, is an "optimisation model" that no longer acts as an objective Occam's razor but rather as a subjective pre-determined transformational mechanism.

Consider the cladogram A(B(C(D(EF)))), where taxa A, B, C and D are members of one genus and taxa E and F members of another. The cladogram indicates that the genus A-D is non-monophyletic as taxon D is more closely related to another genus (EF) than it is to A-C. As trees obtained from phylo-phenetics and optimization are based on a series of switches and/or transformations of homologues (cells) rather than on actual homologies (relationships), they cannot determine if the tree is monophyletic because it applies an assumed or inferred evolutionary history in which we can say that taxa E and F have "changed" from one genus to another. Under such circumstances all non-monophyletic groups remain "monophyletic". The use of such models is dubious, as existing taxonomic groups cannot be tested for monophyly; rather they are forced to abide by some *a priori* evolutionary model. The only way for such models to determine monophyly is to assume that the tree has been constructed from relationships. Clearly generating a tree under one model and interpreting it under the Cladistic paradigm introduces double standards into comparative biology.

The data matrix has restrained, and at times prevented, systematists and biogeographers from expressing taxic and area relationships precisely. It seems that once the matrix is abandoned and the data are treated as relationships then more complex problems can be tackled. Rather than enhancing our systematics enquiries, much hard won valuable information – evolutionary relationships – can acquire more precise meaning.

It is time to abandon the matrix.

> "Today, a matrix and a computer analysis are absolutely necessary…, if you don't provide it, the referees and editors will demand it"
> (Patterson 1997).
>
> "The Matrix is the world that has been pulled over your eyes to blind you from the truth"
> (Morpheus in "*The Matrix*", Warner Bros. 1999).

Acknowledgements
We thank René Zaragüeta Bagils (UPMC, Paris) and Gareth Nelson (University of Melbourne) for their helpful comments.

REFERENCES

BRADY R. H. 1998. — The idea in nature: rereading Goethe's organics, *in* SEAMON D. & ZAJONC A. (eds), *Goethe's Way of Science: A Phenomenology of Nature*. SUNY Press, Albany NY: 83-111.

BROOKS D. R. 1981. — Hennig's parasitological method: a proposed solution. *Systematic Zoology* 30: 229-249.

BROOKS D. R., DOWLING A. P. G., VAN VELLER M. G. P. & HOBERG E. P. 2004. — Ending a decade of deception: a valiant failure, a not-so-valiant failure and a success story. *Cladistics* 20: 32-46.

EBACH M. C., HUMPHRIES C. J. & WILLIAMS D. M. 2003. — Phylogenetic biogeography deconstructed. *Journal of Biogeography* 30: 1285-1296.

EBACH M. C. & WILLIAMS D. M. 2004. — Congruence and language. *Taxon* 53: 113-118.

HUMPHRIES C. J. & EBACH M. C. 2004. — Biogeography on a dynamic earth, *in* LOMOLINO M. & BROWN J. (eds), *The Foundations of Biogeography*. Sinauer Press, Sunderland, Massachusetts: 67-86.

KÄLLERSJÖ M., ALBERT V. A. & FARRIS J. S. 1999. — Homoplasy *increases* phylogenetic structure. *Cladistics* 15: 91-93.

NELSON G. 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. 1984. — Cladistic and biogeography, *in* DUNCAN T. & STUESSY T. F. (eds), *Cladistics: Perspectives on the Reconstruction of Evolutionary History.* Columbia University Press, Columbia: 273-293.

NELSON G. 1989. — Cladistics and evolutionary models. *Cladistics* 5: 275-289.

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

NELSON G. 1996. — *Nullius in verba.* Published by the author, 14 p. (reprinted in Williams & Ebach 2005).

NELSON G. 2004. — Cladistics: its arrested development, *in* WILLIAMS D. M. & FOREY P. L. (eds), *Milestones in Systematics.* CRC Press, Boca Raton, Florida: 127-147.

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

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

NELSON G. & PLATNICK N. I. 1991. — Three-taxon statements: a more precise use of parsimony? *Cladistics* 7: 351-366.

PATTERSON C. 1997. — *Molecules and Morphology, Ten Years On.* Unpublished address to Conference on Molecules and Morphology in Systematics, Paris. Natural History Museum archives, London.

RIEPPEL O. & KEARNEY M. 2002. — Similarity. *Biological Journal of the Linnean Society* 75: 59-82.

RONQUIST F. 1997. — Dispersal-vicariance analysis: a new approach to quantification historical biogeography. *Systematic Biology* 46: 195-203.

ROSEN D. E. 1978. — Vicariant patterns and historical explanation in biogeography. *Systematic Zoology* 27: 159-188.

ROSEN B. R. 1985. — Long-term geographical controls on regional diversity. *Journal of the Open University Geological Society* 6: 25-30.

ROSEN B. R. 1988a. — Tectonics from fossils? Analysis of reef-coral and sea-urchin distributions from the late Cretaceous to Recent, using a new method, *in* AUDLEY-CHARLES M. G. & HALLAM A. (eds), Gondwana and Tethys. *Geological Society Special Publications* 37, The Geological Society, Oxford University Press, Oxford: 275-306.

ROSEN B. R. 1988b. — From fossils to earth history: applied historical biogeography, *in* MYERS A. A. & GILLER P. S. (eds), *Analytical Biogeography: an Integrated Approach to the Study of Animal and Plant Distributions.* Chapman and Hall, London; New York: 437-481.

SCOTLAND R. W. 2000. — Homology and systematics: coding characters for phylogenetic analysis, *in* SCOTLAND R. W. & PENNINGTON R. T. (eds), *Homology and Systematics*. Taylor and Francis, London: 145-182.

SIEBERT D. J. & WILLIAMS D. M. 1999. — Recycled. *Cladistics* 14: 339-347.

SNEATH P. H. A. & SOKAL R. R. 1973. — *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* W. H. Freeman, San Francisco, 573 p.

SOKAL R. R. & SNEATH P. H. A. 1963. — *Principles of Numerical Taxonomy.* W. H. Freeman, San Francisco, xvi + 359 p.

WHEELER W. 2002. — Optimization alignments: down, up, error, improvements, *in* DE SALLE R., GIRIBET G. & WHEELER W. (eds), *Techniques in Molecular Systematics and Evolution*. Birkhäuser Verlag, Basel: 5-69.

WILLIAMS D. M. 2004. — Homology and homologues, cladistics and phenetics: 150 years of progress, *in* WILLIAMS D. M. & FOREY P. L. (eds), *Milestones in Systematics*. CRC Press, Boca Raton, Florida: 191-224.

WILLIAMS D. M. & EBACH M. C. 2005. — Drowning by numbers: re-reading Nelson's *Nullius in Verba*. *Botanical Review* 71: 415-447.

WILLIAMS D. M. & HUMPHRIES C. J. 2004. — Homology and character evolution, *in* STUESSY T., HÖRANDL E. & MAYER V. (eds), *Deep Morphology: Toward a Renaissance of Morphology in Plant Systematics*. Koeltz, Königstein: 119-130.

ZANDEE M. & ROOS M. 1987. — Component compatibility in historical biogeography. *Cladistics* 3: 305-332.

Submitted on 27 September 2005;
accepted on 4 May 2006.

APPENDIX

TABLE 1. — **1**, Matrix of four taxa (A-D) and three characters (1-3); **2**, list of characters and their components. Positive components are derived from the 1-entries, Negative components are derived from the 0-entries.

1

	1	2	3
A	0	0	0
B	1	1	0
C	1	0	1
D	0	1	1

2

Character	Positive	Negative
1	BC	AD
2	BD	AC
3	CD	AB

TABLE 2. — **1**, Matrix of four taxa (A-D) and outgroup (with all zero values) and three characters (1-3); **2**, list of characters and their components. Positive components are derived from the 1-entries, Negative components are derived from the 0-entries.

1

	1	2	3
OUT	0	0	0
A	0	0	0
B	1	1	0
C	1	0	1
D	0	1	1

2

Character	Positive	Negative
1	BC	–
2	BD	–
3	CD	–

TABLE 3. — **1**, Matrix of four taxa (A-D) and outgroup (with all positive values) and three characters (1-3); **2**, list of characters and their components. Positive components are derived from the 1-entries, Negative components are derived from the 0-entries.

1

	1	2	3
OUT	1	1	1
A	0	0	0
B	1	1	0
C	1	0	1
D	0	1	1

2

Character	"Positive"	"Negative"
1	AD	–
2	AC	–
3	AB	–

TABLE 4. — **1**, Matrix of four taxa (A-D) and outgroup (with all unknown values) and three characters (1-3); **2**, list of characters and their components. Positive components are derived from the 1-entries, Negative components are derived from the 0-entries.

1

	1	2	3
OUT	?	?	?
A	0	0	0
B	1	1	0
C	1	0	1
D	0	1	1

2

Character	Positive	Negative
1	BC	AD
2	BD	AC
3	CD	AB

TABLE 5. — **1**, Matrix of four taxa (A-D) and outgroup (with all zero values) and three multi-state characters (1-3); **2**, list of characters and their components. Positive components are derived from the 1- and 2-entries.

1

	1	2	3
OUT	0	0	0
A	2	2	2
B	1	1	2
C	1	2	1
D	2	1	1

2

Component	Positive	Negative
1	AD	–
2	AC	–
3	AB	–
4	BC	–
5	BD	–
6	CD	–

TABLE 6. — **1**, Matrix of four taxa (A-D) six binary characters (1-3) representing the multi-state characters from Table 5, hence binary characters 1a and 1b represent multi-state character 1 from Table 5; **2**, list of characters and their components. Positive components are derived from the 1-entries, Negative components are derived from the 0-entries.

1

	1a	1b	2a	2b	3a	3b
A	0	1	0	1	0	1
B	1	0	1	0	0	1
C	1	0	0	1	1	0
D	0	1	1	0	1	0

2

Component	Positive	Negative
1	AD	AD
2	AC	AC
3	AB	AB
4	BC	BC
5	BD	BD
6	CD	CD

TABLE 7. — **1**, Matrix of four taxa (A-D) six binary characters (1-3) representing the multi-state characters from Table 5, with the addition of an all-zero outgroup; **2**, list of characters and their components. Positive components are derived from the 1-entries, Negative components are derived from the 0-entries.

1

	1a	1b	2a	2b	3a	3b
OUT	0	0	0	0	0	0
A	0	1	0	1	0	1
B	1	0	1	0	0	1
C	1	0	0	1	1	0
D	0	1	1	0	1	0

2

Component	Positive	Negative
1	AD	–
2	AC	–
3	AB	–
4	BC	–
5	BD	–
6	CD	–

TABLE 8. — Summary of results from Tables 1-7. Column 3, "Root" represents the outgroup.

Data set	Characters	"Root"	Positive Components	Negative Components
Table 1	Binary	None	3	3
Table 2	Binary	Positive	3	0
Table 3	Binary	Negative	3	0
Table 4	Binary	Positive + Negative	3	3
Table 5	Multi-state	Positive	6	0
Table 6	Multi-state	None	6	6
Table 7	Binary	Positive	6	0