Cladistics (1996) 12:265–272

FORUM

THREE STEPS OF HOMOLOGY ASSESSMENT

Andrew V. Z. Brower[1] and Valerie Schawaroch[1,2]

[1]*Department of Entomology, American Museum of Natural History, Central Park West at 79th Street, New York, NY, U.S.A. and* [2]*Department of Biology, City College, CUNY, New York, NY, U.S.A.*

Received for publication 2 January 1996; accepted 10 May 1996

Abstract — In 1991 de Pinna (Cladistics 7: 367–394) coined the term primary homology as the putative homology statements prior to tree reconstruction. However, some confusion still exists regarding the conjectural nature of homology and to the analysis of DNA sequences. By dividing de Pinna's term primary homology into topographical identity and character state identity, we emphasize the sequential refinement of putative homology statements. We discuss the problem of transformational versus taxic homology and explain the application of our terms to DNA sequence data.

Definitions

This paper is an elaboration of de Pinna's (1991) discussion of homology concepts, assessment and testing. That paper, as the first of its kind in *Cladistics*, is widely cited as the primary reference for many of the ideas we will address, although most insights into homology have long pedigrees back into the previous century. De Pinna (1991) followed Patterson (1982) (and many others cited by de Pinna) in equating homology with synapomorphy. This definition has the advantage of making explicit the relational aspect of homology: a homology is a condition shared by some subset of taxa in relation to the other members of a more inclusive set of taxa lacking that condition[1]. De Pinna's review further noted that discovery of homology involves two steps. His "primary homology" is the conjecture that similar individual characters are the same and represent evidence of grouping, based on similarity in structure and position (the morphological correspondence of Woodger, 1945). His "secondary homology" is primary homology that has been corroborated by other such homologies in cladistic analysis and characterizes a monophyletic group with respect to its sister taxon in a most parsimonious cladogram. Rieppel (1988) also reviewed these issues at some length, using "topographical correspondence" and "homology" the same way de Pinna used primary and secondary homology. Lipscomb (1992) cites numerous authors,

[1]Bock (1973) is often attributed with the insight that homology is a relational concept, but it seems that the idea has been around at least since Westwood's (1840) discussion of the relationship between affinity and analogy at different hierarchical levels.

0748-3007/96/030265+08/$18.00/0

including Wiley (1975), Platnick (1977), Gaffney (1979) and Eldredge and Cracraft (1980) as generally agreeing with this two-step process.

We agree with the distinction between conjectural homology assessments prior to cladistic analysis and corroborated homology assessments after cladistic analysis, but we feel that de Pinna's characterization of primary homology conflates two distinct operations (as does Rieppel's characterization of topographical correspondence). In this paper, we offer a refinement of these concepts. We restrict our discussion to the idea of primary homology as a binary notion (same versus different) and do not address contingent ideas about relations among character states, such as adjacency or polarity, which have been debated extensively in recent volumes of this journal (e.g. Pogue and Mickevich, 1990; Wheeler, 1990; Lipscomb, 1992).

It may be argued, as Platnick (1979) did, that character states are merely arbitrary groupings of characters, all of which represent modifications of other, more general characters. This view holds that the identification of "independent" characters with "alternate" states is in fact a state-ordering procedure, rendering character enumeration an implicit hypothesis of transformation. We do not find this an especially useful line of reasoning, however, because it implies the non-independence of characters, which many cladists consider problematic (Farris, 1983; Kluge, 1994). In any case, failure to incorporate transformational order among some levels in the "great chain of characters" (Platnick, 1979) prior to cladistic analysis is part and parcel of the systematic endeavor. As Platnick pointed out, if we knew the entire character transformation series, we would also know the taxonomic phylogeny, and our work would be complete. The focus of this paper is at a prior stage of analysis—transformation series and other hypotheses of connectedness among character states represent differential, a priori weighting schemes that might be applied to the completed data matrix during cladistic analysis, while our goal is to clarify concepts relating to building the data matrix.

To propose a primary homology, one must first discover comparable features among the taxa in question—establish "one-to-one comparison" of Woodger (1945) and Inglis (1966) among the characters. Borrowing from Jardine (1969), we will call the desired attribute of characters *topographic identity*. Jardine used the term "topographical homology" to describe characters that fulfill the criterion of correspondence of relative position (see also Rieppel, 1988). Only after these units of comparison are established may individual character states be hypothesized homologous, or hierarchically singular[2]. Our terminology is compared schematically to de Pinna's in Fig. 1. A convenient way to conceptualize this distinction is to imagine an empty character matrix. Identifying comparable characters in the study taxa by lining up the columns of the matrix is clearly a separate operation from filling in the individual cells. Obviously, we are aware that some particular physical features of particular organisms that we deem homologous with

[2]We are not especially concerned here with an explicit ontological definition for homology. Equating homology with synapomorphy only foists its metaphysical baggage onto the latter concept. However, we will claim that, since evolution has nothing to do with the successful discovery and interpretation of characters, evolutionary definitions are clearly superfluous. Practically speaking, homology may not require an operational definition more complicated than "similarity, as ordered by the cladistic method".

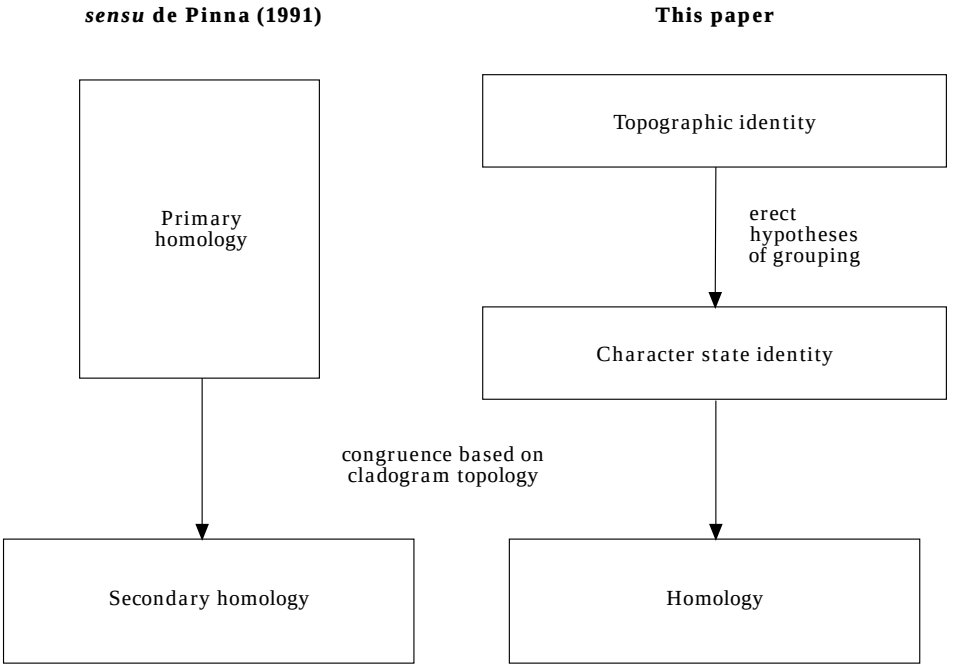

Fig. 1. Flow chart comparing de Pinna's (1991) homology assessment scheme to our modification. The distinction between topographical identity and character state identity, which are conflated under de Pinna's primary homology, is described in the text. Transformation series, step matrices, or other hypotheses of character state ordering may be incorporated in the cladistic analysis, after character state identity is hypothesized.

features of other organisms are difficult to relate by this criterion. However, the topographical identity we are referring to is logical, rather than empirical; all the features of our taxa listed in a given column in our data matrix are, by definition, states of the same, topographically identical character. All the states of a topographically identical character in different taxa are transformationally homologous (Platnick, 1979; Patterson, 1982), but, because the states are not ordered relative to one another at this stage of the operation, no hypothesis of cladistic grouping is put forth at this stage.

After characters are identified via topographical identity, the various character states in the study taxa are hypothesized to be identical or not, based on the standard test criteria for recognizing homology (see Remane, 1952; Patterson, 1982, 1988; de Pinna, 1991 for extensive discussion and references on criteria for recognizing homology). We will refer to this second step as *character state identity*. Character states among taxa are classified either identical or not, when entered in a column of the data matrix[3]. Because the nested relations among character states do not need to be designated a priori, all shared, identical character states rep-

[3]Again, "identity" refers to a logical relation, rather than an empirical one: regardless of the degree of similarity among features of organisms, their codification as data points in a matrix demands the all-or-none classification of sameness or difference (identity / non-identity).

resent conjectures of *potential* homology, and count as evidence in phylogenetic analysis, even if they are subsequently discovered to be symplesiomorphic. The adjudication of homology (=synapomorphy) is a result of cladogram construction.

Cladistic analysis treats each set of identical character states as an independent hypothesis of grouping, and summarizes this information into networks relating the taxa[4]. The parsimony criterion is employed to choose the network that maximizes the character state agreement among all the characters in the data matrix. In the most parsimonious network, some "identical" character states may be distributed among taxa that join the network at more than one node. Although these character states were hypothetical homologies at Step 2, the hypotheses are rejected due to lack of contiguity on the most parsimonious network as established by the weight of the evidence and are therefore deemed homoplastic.

Rooting a network polarizes the character states (polarity statements are an unnecessary assumption prior to cladogram construction; Farris, 1982; Nixon and Carpenter, 1993). Only after the network is rooted are heirarchical apomorphy relations established among character states borne by contiguous taxa. Monophyletic groups are identified by synapomorphies, while paraphyletic groups are identified by symplesiomorphies, which together with the transformed apomorphic states represent monophyletic groups at a more general hierarchical level. Categorization of character evidence into homoplasy, symplesiomorphy and synapomorphy is, of course, always contingent upon accumulation of more data: no synapomorphy is ever proven true or false (Gaffney, 1979; de Pinna, 1991). Cladograms are accepted tentatively, as the best explanation of the available data. Furthermore, while the topographic identity of the characters is assumed as background knowledge (Popper, 1959) during the character state scoring operation, and both topographic identity and character state identity are assumed in subsequent cladistic analysis, either may be re-examined at any time via Hennig's (1966) reciprocal illumination.

Note that our terms, *topographical identity* and *character state identity* do not invoke the notion of homology. This is not because of the problem of equating convergence with homology discussed by Rieppel (1988) and de Pinna (1991), but for a more basic reason. An important conclusion of the cladistic approach is that homology in the evolutionary sense (features shared due to common ancestry; e.g. Mayr, 1969) is not knowable as an intrinsic quality of features of organisms, but only as an inferred explanation for the results of systematic analysis. If homology is synapomorphy (Patterson, 1982), and synapomorphies are identified by rooting the cladogram, then homology cannot be identified prior to cladistic analysis. Such an argument is tied to the view that systematics is independent of (indeed,

[4]This discussion assumes that cladograms will be rooted with outgroups. While alternate character polarization criteria, such as ontogeny, may be employed for some data, outgroup rooting is the most flexible method for handling diverse data types. Molecular data in particular are not amenable to ontogenetic character polarization (Hillis, 1987). Because we are interested in incorporating character data from as many sources as possible, we advocate outgroup rooting as the preferred method. We also assume that character states are by default non-additive. However, while ad hoc hypotheses reduce the explanatory power of the resultant cladogram (Farris, 1983), use of character transformation series or other weighting schemes may be a valuable way to constrain/refine the inference of clades by incorporating additional background knowledge into the discovery operation.

logically prior to) theories of evolution that offer causes for observed patterns of character state distribution (Platnick, 1979; Nelson and Platnick, 1981; Brady, 1985). Given these considerations, we agree with Rieppel (1988) that use of the term "homology" is premature when applied to conjecturally identical character states.

Implications

Viewing homology discovery as the three step process described above helps avoid semantic confusion in several areas. For example, de Pinna (1991: 373) stated, "(s)imilarity or topographical correspondence is factual, while primary homology is already a statement of putative generality, an expectation that correspondences are part of a general pattern". It is evident, however, that topographical identity (correspondence) is not a fact, but a hypothesis or conjecture at a prior level of generality (Popper, 1959; see Jardine, 1969, and de Beer, 1971, for paradoxical examples of topographical non-identity). The fact-like status of a hypothesis of topographical identity is acquired when it is accepted as background knowledge for further conjecture of grouping via character state identity.

The problem of taxic versus transformational homology (Patterson, 1982; Rieppel, 1988; de Pinna, 1991) is also clarified by our terminology. All the states of a topographically identical character are considered transformationally homologous with one another, but no grouping of taxa is implied at this stage. Thus, transformational "homology" is logically prior to taxic homology. Only after character state identity has been hypothesized are exclusive groups implied. Again, we are not discussing relations between character states (transformation series, etc.), but the relations among taxa sharing the same character state. As discussed above, the topographical identities of a given study inevitably become character states of some more general character. All characters represent both synapomorphies at some level in the taxonomic hierarchy (Eldredge and Cracraft, 1980) and transformations of more general characters (Platnick, 1979; Nelson and Platnick, 1981). For example, comparisons of floral structure (as alternate character states) assume the presence of flowers (as the same character) in the taxa being compared, but presence of flowers is a synapomorphy of a more inclusive group. This point implies that the presence of a feature is topographically identical with its absence as alternate states of the same character. If one cared to score it, for example, absence of a placenta would be an equally valid (if uninformative) character state for a paramecium as for a kangaroo. Note that we need not be concerned with the implication that such states are homologous, because homology (apomorphy) is not invoked until after cladogram inference.

A third important area where the proposed terms are helpful is in considering homology of DNA sequences. The literature on sequence alignment (e.g. Doolittle, 1981; Mindell, 1991; Wheeler, 1994, 1995) suggests that aligning sequences is equivalent to de Pinna's (1991) primary homology assessment and represents a large but necessary assumption of background knowledge for subsequent cladistic analysis. In our terms, sequence alignment corresponds to topographical identity, while assessment of the status of individual nucleotide sites corresponds to character state identity. With only five discrete character states (A,

G, C, T, gap) there is little ambiguity in evaluating character state identity once an alignment is obtained. Thus, determining which characters are topographically identical among taxa is the challenging step of hypothesizing "primary homology" in sequence data. This is in contrast to assessment of many morphological characters, in which determining topographical identity is so uncontroversial as to seem "factual" (de Pinna, 1991, see above), while hypothesizing character state identity is more complex.

Patterson (1988) argued that statistically determined similarity is the arbiter of homology in DNA sequences, while congruence is the arbiter in morphology. This is wrong. Patterson's confusion arose from his view of DNA sequences as single characters represented as transformational strings, rather than series of aligned characters with clear relationships among topographically identical sites. Once characters are aligned, their positional relationship disappears from the calculus of identity and they become independent, unitary statements of similarity or dissimilarity with their topographically identical counterparts. DNA characters and morphological characters are treated exactly the same way once they are situated in the data matrix, and congruence is the arbiter for both.

This discussion is confined to systematic comparisons of DNA sequences only. Pairwise searches of an unknown gene product against computerized sequence libraries do not imply hierarchical grouping. The view that homology is inferred from cladistic analysis makes moot notions like Patterson's (1988: 620) statement, "(i)n molecular sequence comparisons the 'gray zone' between homology and nonhomology concerns similarity—and whether similarities (say 2.0 SDs above chance expectation) are or are not homologous". In our terms, such pairwise comparisons are merely topographically identical with some degrees of statistical significance above an expectation based on a particular null model (e.g. Doolittle, 1981). Of course, rejection of the null hypothesis does not provide any ontological legitimacy to the result, it merely shunts our belief from the data at hand onto the underlying model. Further, measures of overall similarity conflate the three categories of character state identity (synapomorphy, symplesiomorphy, homoplasy), and so are incompatible with the cladistic view of hierarchical relationships.

In summary, we have proposed the terms topographical identity and character state identity to replace de Pinna's primary homology. The distinction between characters as comparable categories and character states as hypotheses of grouping may help resolve some of the misunderstanding surrounding the operation of homology discovery in both morphological and molecular systematics.

Acknowledgments

We thank the members of our systematics discussion group (Donat Agosti, Rick Baker, Ranhy Bang, Jim Bonacum, Rob DeSalle, Celeste Durando, Paul Goldstein, Aloysius Phillips, Howard Rosenbaum and Peter Walsh) for discussion, and Jim Carpenter, Jim Miller, Michael McDonald and an anonymous reviewer for comments on the manuscript. AVZB is supported by a Kalbfleisch Postdoctoral Fellowship at the AMNH, and by NSF BSR-9220317 to J. S. Miller and R. DeSalle. VAS is supported by the Graduate Student Fellowship Program of the AMNH.

REFERENCES

BOCK, W. J. 1973. Philosophical foundations of classical evolutionary classification. Syst. Zool. 22: 375–392.

BRADY, R. H. 1985. On the independence of systematics. Cladistics 1: 113–126.

DE BEER, G. R. 1971. Homology: an unsolved problem. Oxford Biol. Readers 11: 3–16.

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

DOOLITTLE, R. F. 1981. Similar amino acid sequences: chance or common ancestry? Science 214: 149–159.

ELDREDGE, N. AND J. CRACRAFT. 1980. Phylogenetic Patterns and the Evolutionary Process. Columbia University Press, New York.

FARRIS, J. S. 1982. Outgroups and parsimony. Syst. Zool. 31: 328–334.

FARRIS, J. S. 1983. The logical basis of phylogenetic analysis. In: N. I. Platnick and V. A. Funk (eds). Advances in Cladistics, vol. 2. Columbia University Press, New York, pp. 7–36.

GAFFNEY, E. S. 1979. An introduction to the logic of phylogeny reconstruction. In: J. Cracraft and N. Eldredge (eds). Phylogenetic Analysis and Paleontology. New York, Columbia University Press. pp. 79–111.

HENNIG, W. 1966. Phylogenetic Systematics. University of Illinois Press, Urbana, IL.

HILLIS, D. M. 1987. Molecular versus morphological approaches to systematics. Annu. Rev. Ecol. Syst. 18: 23–42.

INGLIS, W. G. 1966. The observational basis of homology. Syst. Zool. 15: 219–228.

JARDINE, N. 1969. The observational and theoretical components of homology: a study on the morphology of the dermal skull-roofs of rhidpidistian fishes. Biol. J. Linn. Soc. 1: 327–361.

KLUGE, A. G. 1994. Moving targets and shell games. Cladistics 10: 403–413.

LIPSCOMB, D. L. 1992. Parsimony, homology and the analysis of multistate characters. Cladistics 8: 45–65.

MAYR, E. 1969. Principles of Systematic Zoology. McGraw-Hill, New York.

MINDELL, D. P. 1991. Aligning DNA sequences: homology and phylogenetic weighting. In: M. M. Miyamoto and J. Cracraft (eds). Phylogenetic Analysis of DNA Sequences. Oxford University Press, Oxford, pp. 73–89.

NELSON, G. AND N. PLATNICK. 1981. Systematics and Biogeography. Columbia University Press, New York.

NIXON, K. C. AND J. M. CARPENTER. 1993. On outgroups. Cladistics 9: 413–426.

PATTERSON, C. 1982. Morphological characters and homology. In: K. A. Joysey and A. E. Friday (eds). Problems of Phylogenetic Reconstruction. Academic Press, London and New York, pp. 21–74.

PATTERSON, C. 1988. Homology in classical and molecular biology. Mol. Biol. Evol. 5: 603–625.

PLATNICK, N. I. 1977. Cladograms, phylogenetic trees, and hypothesis testing. Syst. Zool 26: 438–442.

PLATNICK, N. I. 1979. Philosophy and the transformation of cladistics. Syst. Zool. 28: 537–546.

POGUE, M. G. AND M. F. MICKEVICH. 1990. Character definitions and character state delineation: the bête noire of phylogenetic inference. Cladistics 6: 319–361.

POPPER, K. R. 1959. The Logic of Scientific Discovery. Basic Books. New York.

REMANE, A. 1952. Die Grundlagen des naturlichen Systems, der vergleichenden Anatomie und der Phylogenetik. Geest und Portig, Leipzig.

RIEPPEL, O. C. 1988. Fundamentals of Comparative Biology. Birkhäuser Verlag, Basel.

WESTWOOD, J. O. 1840. Observations upon the relationships existing amongst natural objects, resulting from more or less perfect resemblance, usually termed affinity and analogy. Mag. Nat. Hist. 6 (N.S.): 141–144.

WHEELER, Q. D. 1990. Ontogeny and character phylogeny. Cladistics 6: 225–268.

WHEELER, W. C. 1994. Sources of ambiguity in nucleic acid sequence alignment. In: B. Schierwater, B. Streit, G. P. Wagner, and R. DeSalle (eds). Molecular Ecology and Evolution: Approaches and Applications. Birkhäuser Verlag, Basel, pp. 323–354.

WHEELER, W. C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44: 321–331.

WILEY, E. O. 1975. Karl R. Popper, systematics, and classification: a reply to Walter Bock and other evolutionary systematists. Syst. Zool. 24: 233–243.

WOODGER, J. H. 1945. On biological transformations. In: W. E. Le Gros Clark and P. B. Medawar (eds). Essays on Growth and Form Presented to D'arcy Thompson. Oxford University Press, Oxford. pp. 95–120.