# The Delimitation of Phylogenetic Characters

**Eric S. J. Harris**
Osher Research Center
Harvard Medical School
Cambridge, MA, USA
eric_harris@hms.harvard.edu

**Brent D. Mishler**
Department of Integrative Biology
University and Jepson Herbaria
University of California
Berkeley, CA, USA
bmishler@calmail.berkeley.edu

*The bud disappears in the bursting-forth of the blossom, and one might say that the former is refuted by the latter; similarly, when the fruit appears, the blossom is shown up in its turn as a false manifestation of the plant, and the fruit now emerges as the truth of it instead. These forms are not just distinguished from one another, they also supplant one another as mutually incompatible. Yet at the same time their fluid nature makes them moments of an organic unity in which they not only do not conflict, but in which each is as necessary as the other; and this mutual necessity alone constitutes the life of the whole.*

— Hegel, Preface to *Phenomenology of Spirit*

## Characters and Classification in Biology

The main difficulty of character delimitation in biology comes from the fact that it is difficult to find objective suture lines along which the analytical gaze of a biologist can cut and define a character (Lewontin 2001). Yet biologists need to find such sutures in order to progress with their research. The bud is isolated from the unity of the plant and distinguishing features noted, the blossom is examined, and so on at various levels of biological organization. Character delimitation plays a fundamental role in any biological research, thus the process of "carving nature at its joints" (Plato 1997, *Phaedrus* 265d–266a) needs careful examination if it is to be part of a rational scientific inference process.

Using information derived from an examination of the characteristics of objects or organisms under study, all biologists, indeed all scientists, need to develop classifications of the things they study. There are four desirable criteria for classifications: (1) *practicality*: names that are easy to apply and stable; (2) *information content*: names that index an optimal summarization of what is known; (3) *predictivity*: names that maximally predict unknown features; (4) *function in theories*: names that capture entities acting in, or resulting from, natural processes. These criteria sometimes seem contradictory, in which case debates erupt between pragmatists emphasizing criterion 1 and theoreticians emphasizing criterion 4 (e.g., debates over chemical classification in the 1960s and 1970s; debates over biological classification in the 1970s and 1980s). Ultimately, however, these criteria should not be contradictory, and should flow from criterion 4 to criterion 1, in the sense that representing an important natural process in a classification will lead to high predictivity, information content, and true practicality for users of the classification (Mishler 2009). The key to "carving nature at its joints" is to find the joints first. Molecular biologists need to recognize and name genes, functional regions of genes, and gene products. Ecologists need to characterize elements of food chains or nutrient cycles. Phylogenetic systematists need to name monophyletic groups.

Selecting a character requires denying the continuity of form of the organism. But there are several ways that this continuity can be denied and the selection process depends on what purpose the characters will be used for. These different purposes can have important consequences for the resulting inferences or analyses that use the characters as their basis. In order to make the process of character delimitation repeatable and explicit, the underlying biological and evolutionary processes must be examined carefully, and some general procedures must be adopted based on these processes. The purpose of this article is to address the problem of character analysis as it is manifested in the field of phylogenetic systematics. Any resolution to the problem of character delimitation, not only in phylogenetics but also biology more broadly, includes both a theoretical and a practical component. The theoretical

February 17, 2009; accepted October 10, 2009

Biological Theory 4(3) 2009, 230–234. 

aspects of phylogenetic character coding are addressed in this article. The practical issues of phylogenetic character delimitation are addressed in a separate paper (Harris and Mishler in preparation).

## "Individuals" Versus "Classes"

The problem of character delimitation is not unique to biology, of course. In fact, it can be thought of as a particular instantiation of the very general problem of assigning predicates to objects, of defining and describing things. The relation between an object and its predicates remains a problematic issue in the philosophy of language. The notion of "character" has been under discussion for many years (e.g., Wilkins 1668). Traditionally, objects have been considered to be definable by a set of necessary and sufficient conditions (Schwartz 1977). The traditional view has been that the definition of a term is made through reference to the essential properties of an object, and if something exhibits those essential properties then it is *ipso facto* the object specified.

However, some philosophers have argued that even general terms are defined demonstratively through specific reference to a particular object or group of objects, that is, by ostension (e.g., Kripke 1980). These philosophers of language have emphasized the importance of concrete spatiotemporal recognition of an object. In this view, objects are initially referred to through demonstration. The connection between word and object is then perpetuated through a community of speakers that ultimately reaches back to the object, person, or thing itself (Kripke). Consequently, a word is defined historically through a community of speakers reaching back to the original referent(s), rather than being defined by a certain set of properties.

This debate relates to the distinction between individuals and classes (Ghiselin 1975; Hull 1978). An individual is a historically bounded entity, with no particular properties that can define it. An individual in this sense must ultimately be referred to through ostensive definition. Type specimens take the function of ostensive definition in scientific taxonomy, and as such a taxonomic name is ultimately defined by reference to a type specimen, not by any particular property (Hull). Ghiselin and others (e.g., Hull) have successfully argued that biological species are "individuals" in this sense. By contrast, a "class" is any object that conforms to a certain set of essential properties. Any one member of a class is interchangeable with any other member as long as they share those properties. Elements of the periodic table are examples of a class, since to be a specific atomic element the object must always have a certain atomic number.

A biological taxon does not have essential properties, and ultimately must be defined through ostension. This view is the one adopted in this article, although there remains debate on the topic (e.g., Ereshefsky 2007). Dupré (1981) echoes the sentiment expressed here in emphasizing that characters can never be a full definition of a taxon, since a taxon can never be circumscribed by necessary and sufficient conditions. The reason for this is that taxa are evolutionary lineages. As such, a taxon is defined as the assemblage of all things that arise from a common ancestor, even though none of those things necessarily share any one character in common (Hull 1978; Grant and Kluge 2004). Hennig (1966) realized this too, acknowledging that "holomorphology" could only be used as a proxy for evolutionary descent.

However, it is very clear from the vast literature describing the world's biodiversity that it is possible to use characteristics to describe biological taxa and, through the use of identification keys, identify an unknown organism. It is important to note that characters used for identification are not being used for the definition of a taxon. As indicated by the arguments above, identifiable characteristics of an organism are incidental to it as a result of its status as an "individual." A character of any nature is never an abstract, essential, defining property of an organism in the way that the atomic number of an element is. The most they can be are convenient "handles" by which one can recognize and refer to an organism or taxon. Based on these arguments, we recognize that characters in an identification key or taxonomic description are technical instruments of vocabulary selected for their ability to describe and differentiate biological taxa (cf. "description" and "diagnosis"; Hull 1978). Phylogenetic characters serve a very different purpose, yet they are also technical instruments, selected in this case for their ability to discover phylogenetic relationships. The nature of characters, then, depends on pragmatic and utilitarian considerations as well as on theory. Their definition must always relate to their function.

A confusion may arise in discerning between characters used for various purposes, such as for phylogenetic analyses, identification keys, classification of ecological communities, and so on. The confusion results because the use of characters in one context, say phylogenetic analysis, may often make use of similar characteristics as other contexts (e.g., description, diagnosis, ecology). In general, characters of organisms can function in different ways and serve many purposes in scientific research. Ultimately, the selection of various types of characters for different purposes operates by different criteria and thus characters are not really synonymous when similar characters are used for different purposes. For example, phylogenetic characters represent a particular subset of all possible attributes or features of organisms, while traits used in an identification key represent a different, though sometimes overlapping, subset of all possible attributes or features of organisms.

A common misconception about the delimitation of phylogenetic characters—that they represent *any* distinguishing feature or attribute of an organism—can thus be laid to rest.

There is no theory-neutral way to individuate characters at any level from DNA sequence data to morphology. It is true that any particular aspect of an organism's morphology, anatomy, chemistry, distribution, genetics, or any of its other peculiarities may be examined as a potential phylogenetic character. However, many characteristics of organisms, such as their morphological features, distribution, trophic level, and so on, might be useful in examining the organism in the context of its ecology or behavior, yet not be useful as phylogenetic characters. The point is that the determination of a phylogenetic character must always relate to the purpose of the reconstruction of phylogeny. That is, phylogenetic characters represent a subset of attributes, features, properties of an organism that are chosen for the specific purpose of revealing the evolutionary history of organisms. Thus, in order to clarify the problem of character delimitation for phylogenetics, we must first clearly understand what it is we want these particular types of characters to do.

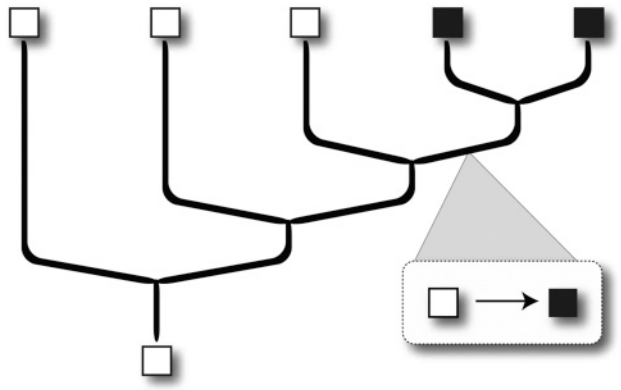

**Figure 1.**
As first explained clearly by Hennig (1966), evolution occurs through descent with modification along lineages that split occasionally. These evolutionary modifications can serve as evidence for the existence of a lineage in the future. As in this figure, the evolutionary modification is represented by the change from a light color to a dark one. The purpose of delimiting phylogenetic characters is to discover modifications such as this that can serve as evidence of evolutionary relatedness.

## What Do We Use Phylogenetic Characters For?

Much of the literature on cladistic methodology places considerably more emphasis on analysis than exploration, taking as the starting point that one has a data matrix, and discussing how to analyze it (Thiele 1993). Character coding represents the link between observation and analysis and greatly influences the results, but has nevertheless received little attention (Pleijel 1995).

In general, phylogenetic analysis consists of two main phases: character analysis and phylogenetic tree building (Neff 1986). There appears to be a common view among phylogeneticists that the character analysis phase has received much less attention as compared to the tree building step (Thiele 1993; Pleijel 1995; Mishler 2005). Additionally, in the literature on phylogenetic character analysis there is little consensus on methods and approaches to code phylogenetic characters. Even the definition of the word "character" remains an open question (Colless 1985; Grant and Kluge 2004), despite many attempts to clearly define the word. By contrast, some degree of consensus has been reached on a limited number of methods of phylogenetic inference (e.g., parsimony, maximum likelihood, Bayesian methods). Phylogenetic inference almost always uses computer algorithms, and there are only a handful of different computer algorithms available for this purpose—e.g., PHYLIP (Felsenstein 1989), PAUP (Swofford 2003), TNT (Goloboff et al. 2000). However, no such comparable homogeneity of methods yet exists for the process of character coding, since characters come in a vast array of forms, whether they are morphological, molecular, or of any other type.

Because of the lack of consensus regarding character coding in phylogenetic systematics, much of the published literature remains fragmentary on the topic, addressing only parts of the problem. Frequently only the theory and philosophical basis of phylogenetic characters (e.g., Kluge 2003; Richards 2003; Grant and Kluge 2004; Fitzhugh 2006), or the actual method of coding the characters (e.g., Wiens 1995) are covered in any one paper that deals with this topic. Obviously, both of these two complementary aspects (theory/method) of character coding are required for a full resolution of this problem. We need to understand the relation that characters have to phylogenetic theory and we need to have some practical method so that character delimitation is a repeatable process. The first criterion reflects the need for a clarification of the theoretical component to the character definition; it is an *ontological* criterion. The second aspect represents a procedural requirement to maintain objectivity; it is an *epistemological* criterion. The central problem for character coding comes from the tension between the procedural requirement of repeatability and the theoretical requirement of biological meaningfulness. Of course, the tension between theory and practice underlies any scientific endeavor.

The four desired criteria for classifications discussed above can help relieve this tension, particularly the emphasis placed above on the fourth one (function in theories about process). What theory needs to guide the specific problem of delimiting characters and character states in phylogenetic systematics? What is the fundamental underlying process model that underlies phylogenetic reconstruction? As first explained clearly by Willi Hennig (1966), it is descent with modification along lineages that split occasionally, allowing the modifications to serve as evidence for the existence of a lineage in the future (Figure 1). Evolution occurs via changes along branches, where a prior condition in a feature changes to a posterior condition. After lineage splitting, the presence of the

posterior (derived, or *apomorphic*) condition can serve as a marker for the past lineage in which it arose. It is these markers that the phylogeneticist seeks to infer from the relative recency of common ancestry of the members of a study group.

Ontologically, these changes along lineages underlie the important concept of *homology*—the general definition of which is two or more features that have evolved via common descent from an ancestor that had that feature (Roth 1988). There are several subtypes of homology, however, some of which involve features duplicated within one organism (paralogy and serial homology). The subtype of homology we are most concerned with in this article is *phylogenetic homology*, or the sharing of homologous traits between organisms. There are two kinds of phylogenetic homology: transformational homology and taxic homology. *Transformational homology* is the diachronic relationship that occurs along a lineage between the prior and the posterior conditions of a feature ("events," according to Kluge 2007). *Taxic homology* is a synchronic relationship among two or more lineages that share the same condition.

Epistemologically, these markers serve as models for phylogenetic character definition. The process of character analysis involves searching for sets of conditions that appear to be transformational homologs of each other. Each independent set of postulated transformational homologs is a character; each condition within the set is a character state. There are extensive rules of inference that govern initial hypotheses of homology, which we will cover in another paper (Harris and Mishler in preparation), but for now it is just important to establish that each character and its states are evaluated by a host of empirical criteria that do not involve any particular phylogeny.

These individual character hypotheses are then combined in the form of a table—a matrix that can subsequently be analyzed to infer patterns of phylogenetic relationship, using a parsimony, maximum likelihood, or other criteria, a full description of which is beyond the scope of this article (see Wiley et al. 1991; Kitching et al. 1998; Felsenstein 2004; Mishler 2009). A joint solution of the matrix is sought—what phylogenetic tree best fits all the separate hypotheses of homology? When a "best fit" tree is chosen, the character states that are congruent in their distribution with the tree are judged to be homologies, those that are not are judged to be homoplasies. Thus a second test of homology (congruence) is applied to the initial hypotheses of homology that were made before the matrix was assembled (Patterson 1982).

## Taxonomic Breadth and the Mutual Constitution of OTU and Character

As the above section illustrates, the function of the phylogenetic character is to serve as evidence of common evolutionary history. This function does not rely on any particular scale of evolutionary lineage: phylogenetic characters are used as evidence of common evolutionary history at any level of inclusiveness in the tree of life. Phylogenetic characters are taken to serve as putative evidence for the existence of monophyletic groups, whether that monophyletic group includes all living things or a single individual organism, a group of cells, or a gene family. To do this, the character in question must vary—there must be at least two states (Figure 1). Certain traits may vary at one level of phylogenetic analysis, but be invariant at another. Consequently, the delimitation of characters is always relative to the breadth of the particular phylogenetic study being undertaken. For example, the "presence of chlorophyll a & b" would be a useless character for a phylogenetic study of the genus *Quercus* (= oak), but may be useful in a phylogenetic study of major lineages of eukaryotic organisms. Furthermore, a certain trait of an organism may appear to be evidence for common evolutionary history at one level in the evolutionary tree, but a result of convergent evolution at another (e.g., the presence of wings would be considered the result of convergent evolution at the scale of all amniotes, but might serve as a homology within a smaller group such as bats and their close relatives). At the molecular level, the best comparative alignment of a gene region depends on the phylogenetic breadth of the organisms being compared. Phylogenetic characters are always related to the phylogenetic question that they provide evidence for and vice versa. Character analysis cannot be absolute. Any attempt at automation of the character analysis step and standardization of character terms (e.g., Pullan et al. 2005) must take this fact into account.

We refer to a terminal unit in the cladogram (usually represented as a row in the character matrix) as an operational taxonomic unit (OTU). This is a general term that can encompass any level of inclusiveness. An OTU could represent anything from a broadly defined group of organisms (e.g., a hypothesized monophyletic group representing a taxonomic family) to the organism at a snapshot of its life (i.e., a semaphoront *sensu* Hennig 1966), to a particular gene within one genome. An important corollary to the above point that characters need to have at least two states is that in general the character should be variable *between* OTUs but invariant *within* those OTUs. That is, to be useful in phylogenetic reconstruction, a character should not be polymorphic within an OTU. However, the actual boundaries of the OTU are not known *a priori*, they must be inferred. An OTU represents a hypothesis of a historical individual, a spatiotemporally restricted entity. But the spatiotemporal boundaries of the OTU need to be tested. The phylogenetic characters are used for this purpose.

Therefore, characters and OTUs are related by being mutually tested by one another. In the process of assembling a data set during character analysis, the OTUs and character states are assembled iteratively. The discovery of a new character may well split what was considered one OTU into two, or

rejection of a character may lump two previously considered OTUs into one. Likewise consistency of variation of a potential character within and between already well-supported OTUs, as described above, is one criterion for accepting it as a phylogenetic character. Epistemological details of this process are given in Harris and Mishler (in preparation). For the purpose of this article it is enough to recognize that there is considerable reciprocal illumination between the concept of character and the concept of OTU. A character is chosen partly because it is invariant within an OTU and the OTUs are recognized because they are homogeneous for all the character states currently known (Mishler 2005).

## Summary

The delimitation of phylogenetic characters is a process that fragments the continuity of the organism into comparable attributes, features, and properties that can function as potential indicators of evolutionary history (i.e., homologs) within the context of a specified breadth of phylogenetic study. Taxa are individuals. Two main conclusions result from these points. First, when a phylogeneticist attempts to delimit phylogenetic characters, those characters cannot function as essential properties or elements of an exact definition of a lineage or taxon. Rather, phylogenetic characters are specialized technical words or other aspects (e.g., base pairs of DNA) selected on the basis of their ability to reveal phylogenetic relationships. Second, characters are always relative to the taxonomic breadth of the study being undertaken and the terminal units in the analysis. Characters serve as putative evidence of shared evolutionary history. Characters should ideally not vary within an OTU, yet vary between them. As such, phylogenetic characters will be related to the total group of organisms being considered in a given study as well as the terminal units (OTUs) whose evolutionary relationships are in question. In contrast to the "real characters" sought after by Wilkins (1668), phylogenetic characters are not absolute; characters are always relative to their purpose, phylogenetic breadth, and the operational taxonomic units in the phylogeny.

## References

Colless DH (1985) On "character" and related terms. Systematic Zoology 34: 229–233.

Dupré J (1981) Natural kinds and biological taxa. Philosophical Review 90: 66–90.

Ereshefsky M (2007) Foundational issues concerning taxa and taxon names. Systematic Biology 56: 295–301.

Felsenstein J (1989) PHYLIP—Phylogeny inference package, Version 3.2. Cladistics 5: 164–166.

Felsenstein J (2004) Inferring Phylogenies. Sunderland: Sinauer.

Fitzhugh K (2006) The philosophical basis of character coding for the inference of phylogenetic hypotheses. Zoologica Scripta 35: 261–286.

Ghiselin MT (1975) A radical solution to the species problem. Systematic Zoology 23: 536–544.

Goloboff P, Farris S, Nixon K (2000) TNT (Tree Analysis Using New Technology). (BETA) Published by the authors, Tucumán, Argentina.

Grant T, Kluge AG (2004) Transformation series as an ideographic character concept. Cladistics 20: 23–31.

Harris ESJ, Mishler BD (in preparation) The practice of phylogenetic character delimitation.

Hegel GWF (1977) The Phenomenology of Spirit (Miller AV, trans). Oxford: Oxford University Press.

Hennig W (1966) Phylogenetic Systematics. Chicago: University of Illinois Press.

Hull DL (1978) A matter of individuality. Philosophy of Science 45: 335–360.

Kitching IJ, Forey PL, Humphries CJ, Williams DM (1998) Cladistics: The Theory and Practice of Parsimony Analysis. Oxford: Oxford University Press.

Kluge AG (2003) The repugnant and the mature in phylogenetic inference: Atemporal similarity and historical identity. Cladistics 19: 356–368.

Kluge AG (2007) Completing the neo-Darwinian synthesis with an event criterion. Cladistics 23: 613–633.

Kripke S (1980) Naming and Necessity. Cambridge, MA: Harvard University Press.

Lewontin RC (2001) Foreword. In: The Character Concept in Evolutionary Biology (Wagner GP, ed), xvii–xxiii. San Diego: Academic Press.

Mishler BD (2005) The logic of the data matrix in phylogenetic analysis. In: Parsimony, Phylogeny, and Genomics (Albert VA, ed), 57–70. Oxford: Oxford University Press.

Mishler BD (2009) Three centuries of paradigm changes in biological classification: Is the end in sight? Taxon 58: 61–67.

Neff NA (1986) A rational basis for a priori character weighting. Systematic Zoology 35: 102–109.

Patterson C (1982) Morphological characters and homology. In: Problems of Phylogenetic Reconstruction (Joysey KA, Friday AE, eds), 21–74. New York: Academic Press.

Plato (1997) Complete Works (Cooper JM, Hutchinson DS, eds). Indianapolis: Hackett.

Pleijel F (1995) On character coding for phylogeny reconstruction. Cladistics 11: 309–315.

Pullan MR, Armstrong KE, Paterson T, Cannon A, Kennedy JB, Watson MF, McDonald S, Raguenaud C (2005) The Prometheus description model: An examination of the taxonomic description-building process and its representation. Taxon 54: 751–765.

Richards R (2003) Character individuation in phylogenetic inference. Philosophy of Science 70: 264–279.

Roth VL (1988) The biological basis of homology. In: Ontogeny and Systematics (Humphries CJ, ed), 1–26. New York: Columbia University Press.

Schwartz SP (1977) Introduction. In: Naming, Necessity, and Natural Kinds (Schwartz SP, ed), 13–41. Ithaca: Cornell University Press.

Swofford DL (2003) PAUP: Phylogenetic Analysis Using Parsimony and Other Methods, Version 4. Sunderland: Sinauer.

Thiele K (1993) The holy grail of the perfect character: The cladistic treatment of morphometric data. Cladistics 9: 275–304.

Wiens JJ (1995) Polymorphic characters in phylogenetic systematics. Systematic Biology 44: 482–500.

Wiley EO, Siegel-Causey D, Brooks DR, Funk VA (1991) The Compleat Cladist: A Primer of Phylogenetic Procedures. Lawrence: University of Kansas Museum of Natural History.

Wilkins J (1668) An Essay Towards a Real Character, and a Philosophical Language. London: Gellibrand and John Martyn.