1 Data

Evidence

The fundamental task of systematics is to explain biological variation by inferring the phylogenetic relationships among organisms and the unique transformation events that link them (Hennig 1966).

Inference of particular functions, adaptations, mechanisms, and constraints, and the many other processes that shaped the evolution of each group, can be informed by the results of phylogenetic analysis. However, evolutionary analysis requires additional assumptions and tests that are external to systematics (for example, Farris 1983; Grandcolas and D'Haese 2003; Grant and Kluge 2003).

Operationally, systematics proceeds by gathering data (observations) from organisms and coding them into evidence to test competing phylogenetic scenarios.

In principle, any observation of a set of creatures has the potential to provide evidence of historical kinship. However, the most objectively critical evidence is derived from those features that are heritable and intrinsic to organisms because they reflect the biological continuity between ancestor and descendant (Hennig 1966). Differences in each of these features can be traced to specific and unique transformations on a cladogram and these transformations allow us to assay the relative

merits of alternative historical explanations. All observations are consistent with all scenarios, but not to an equal extent (see Sober 1983). It is this inequality that drives systematic analysis, forms the basis of our tests, and elevates general historical statements to testable scientific hypotheses.

A distinction is made between data and evidence. Evidence implies an organized set of observations that can be used to test hypotheses. A collection of nucleic acid sequences or statements about an anatomical feature (for example, biramous appendage) have little meaning or value in and of themselves. However, when organized into putatively homologous features (for example, 18S rRNA or abdominal appendages) these observations (data) demand analysis. Coding transforms observation into evidence (characters) and allows hypotheses to be weighed quantitatively.

Value of Evidence

At least initially, we must operate under the principle that all evidence is possessed of equal value in discriminating among phylogenetic scenarios. That is, information from all sources and collected by any means may have value in testing historical hypotheses: the total evidence principle (Kluge 1989). This is not to say that all information is equally discriminating, but this cannot be known prior to systematic analysis.

A corollary of this notion is that phylogenetic characters (that is, coded observations) do not form logical classes on the basis of their ability to differentiate among hypotheses. They can be divided and sorted into all variety of functional, structural, or observational classes, but these have no bearing on the evidentiary content of the characters themselves. This leads to a notion of complete catholicism with respect to sources and types of data—provided that they derive from independently heritable transformations. Anatomical, behavioral, and genomic features are all, on the face of them, informative. There is no reason to segregate character variants into classes or partitions. There are only those features that can objectively distinguish between hypotheses and those that cannot.

Sources of Evidence

Traditionally, the data used in phylogenetic analysis have been categorized as either morphological or molecular. Morphological data have originated principally from studies in comparative morphology and ethology (although the latter are not strictly morphological). Molecular data have originated from studies in molecular biology. According to this categorization, morphological data consist of anatomical features

and can include behavior, while molecular data consists of nucleic acids, proteins, and genomes.

This distinction has given rise to a number of controversies regarding the suitability of each kind of data for phylogenetic analysis. In addition, each type of data is regarded as requiring different analytical treatment. This difference in treatment principally concerns how comparable characters are identified. On the one hand, morphological characters have been treated as a matter of observation, while molecular characters must be inferred. Phylogenetic analysis is better served by distinguishing between phenotypic and genotypic data.

Phenotypic and genotypic data

Phenotypic data result from the structural and functional characteristics that express an organism's genotype along with the organism's response to its environment. Phenotypic data therefore include morphology and behavior as well as many molecular characters, such as amino acid sequences or pheromone profiles. Data collected on phenotypes are derived from structurally and developmentally complex features, which enables testing of each hypothesis of homology in isolation.

However, observed variation may be the result of either heritable transformations or environmental effects. Failure to distinguish between the two causes of variation may confound attempts to infer phylogenetic relationships.

On the other hand, genotypic evidence represents an organism's genetic material and is therefore directly and entirely transmitted from parent to offspring, thus eliminating any concern for the heritability of observed variation. Although this is a clear strength of genotypic data, the peculiarities of these features give rise to novel analytical problems. All instances of each of the four possible nucleotides are physically indistinguishable, regardless of their historical origins: any nucleotide can substitute directly for any other (there are no intermediate states) and any nucleotide can be inserted or deleted. It is therefore impossible to test hypotheses of nucleotide homology in isolation. Only the test of character congruence can be applied to them.

Until the time when whole genomes are available and can be analyzed appropriately and the genetic basis for particular phenotypic variants is known and can be traced to unique transformations, combined analysis of both sources of evidence provides the strongest test of phylogenetic hypotheses.

Homology

At the level of evidence, cladograms imply statements of homology. Alternative cladograms might have alternative optimal homology statements and content. At a basic level, we differentiate among cladograms on their ability to embody the potentially conflicting homology statements of diverse sets of characters.

Features are homologous when their origins can be traced to a unique transformation on the branch of a cladogram leading to their most recent common ancestor. There can be no notion of homology without reference to a cladogram (albeit implicitly) and no choice among cladograms without statements of homology.

This definition of homology makes no reference to "primary" or "secondary" homology (de Pinna 1991). In fact, the perspective here rejects this distinction entirely. De Pinna (1991: 372) based his distinction on the view that homology assessment necessarily requires hypothesis "generation and legitimation" in separate steps, the former rooted in notions of similarity and the latter based on the simultaneous test of character congruence.

All possible hypotheses of homology are defined logically as a function of the number of heritable parts identified for each terminal (just as all possible cladograms are defined logically as a function of the number of terminals; Felsenstein 1978), so no special procedure is required for hypothesis "generation." Likewise, although homology assessment often involves a two-stage procedure of first submitting each hypothesis of homology to a round of separate tests and then submitting the surviving, constrained set of hypotheses to the test of character congruence (that is, "static" homology assessment), this separation is neither a methodological nor epistemological necessity.

POY embodies the concept of dynamic homology (Wheeler 2001a, b) in which the test of character congruence is applied to the entire, unconstrained set of hypotheses of homology, thereby allowing entire transformation series to be discovered on the basis of a single optimality criterion. That is, dynamic homology employs the same procedure to discover both the character (in the traditional sense) and the character-state transformations within the character. Since the same optimality criterion is employed in both cladogram assessment and homology assessment, the globally optimal explanation of the observed variation is achieved by the minimum-cost cladogram-plus-homology-scheme combination.

In the same way that each cladogram has a (potentially) unique set of optimal character origins, each cladogram may have a unique set of optimal correspondences among observed features. Unless these corre-

spondences are unrestricted and allowed to be optimized together with transformations, biased and conditional results may be obtained. Such bias may come from the assumptions of the investigator and his or her notions of the appropriateness of comparison, and conditioned on the hypotheses most in agreement with the preconceived correspondences of "primary" homology.

Dynamic homology is a powerful conceptual approach to the study of highly simplified data types, such as DNA and amino acid sequences or simple morphological structures like annelid segments, where structural or developmental evidence that could allow a defensible choice among competing hypotheses of homology is either nonexistent or unavailable.

Suggested Reading

Farris, J. S. 1983. The logical basis of phylogenetic analysis. *In* N. I. Platnick and V. A. Funk (editors), Advances in Cladistics: 277–302. New York: Columbia University Press.

Hennig, W. 1966. Phylogenetic Systematics. Urbana: University of Illinois Press. 263 pp.

Sober, E. 1983. Parsimony methods in systematics. *In* N. I. Platnick and V. A. Funk (editors), Advances in Cladistics: 37–47. New York: Columbia University Press.

Wheeler, W. C. 2001. Homology and the optimization of DNA sequence data. Cladistics 17: S3–S11.

2 Cladograms

Cladograms, Trees, and Tree-Shaped Objects

Independent of evidence, a cladogram is a branching diagram that depicts the hypothesized phylogenetic (as opposed to ontogenetic or tokogenetic) relationships among terminal taxa. They are Steiner or Wagner trees in much of the systematics literature, meaning that observed taxa (often called operational taxonomic units or OTUs) are confined to the tips (leaves) and are not placed at inner nodes. This is preferred over Prim networks because

- it allows more parsimonious solutions by optimizing novel character combinations to inner nodes as hypothetical taxonomic units (HTUs) (Farris 1970),
- and it avoids the problematic assumption that some observed taxa are directly ancestral to other observed taxa (Platnick 1977).

Cladograms can either be undirected (unrooted) networks or directed (rooted) trees (Farris 1970). Only in the latter case can historical statements be made.

The computer science literature uses "tree" in a slightly different context. A tree is a directed, acyclic graph —in more familiar terms, a rooted branching diagram without reticulation. This literature also

uses "network" to mean an undirected cyclic graph—an unrooted branching diagram with reticulation.

In most cases here, we refer to cladograms. We are, however, drawing on a large wealth of algorithmic and tree manipulation discussion from the computational literature, so we will use these terms more broadly, to the point where they are largely interchangeable. "Network" does not refer to reticulated graphs in this book, although we use the term only rarely and usually describe cladograms as directed or undirected.

Associated with an optimized body of evidence, a directed cladogram becomes a summary statement of phylogeny and homology, representing the historical relationships among both the terminal taxa and the individual characters. Cladograms do not confer information about why or how particular transformations occurred between ancestor–descendant pairs. Also, cladograms need not contain specific information about ancestor–descendant character transformations. In many cases we can determine how costly a cladogram is without ever having to determine the precise ancestral reconstructions required by a tree—for example, by only performing down-pass optimization.

The number of possible cladograms is solely a function of the number of terminals. For n terminals, the number of binary (bifurcated, fully resolved) undirected cladograms is given by (Felsenstein 1978)

$$\frac{(2n-5)!}{2^{n-3}(n-3)!} \qquad \text{(Eq 2.1)}$$

and for directed cladograms by

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \qquad \text{(Eq 2.2)}$$

The number of possible cladograms therefore increases explosively as taxa are added to an analysis. For a given matrix of static homology statements, finding the optimal cladogram is an NP-complete problem (Garey and Johnson 1977; Garey et al. 1977). Heuristic solutions are therefore required to analyze data sets composed of more than about 20 taxa.

Choice among cladograms is mediated by optimizing the evidence on cladograms and calculating the minimum cost in terms of weighted transformations required to explain the observed variation in light of background knowledge. That is, cladograms differ in their ability to explain conflicting evidence in a single historical scenario. Optimality criteria measure this ability quantitatively and permit the relative evalua-

tion of cladograms. Cladograms are ranked and optimal solutions identified through these values.

POY is mainly concerned with parsimony as an optimality criterion, valuing, as it does, simplicity. The cladogram that minimizes transformations to explain the observed variation is the simplest, maximizes evidential congruence, and has greatest explanatory power. There is also, however, some ability to evaluate cladograms comparatively in terms of their likelihood scores, where the cladograms that maximize the likelihood are preferred.

Cladograms and trees are branching diagrams that depict the historical relationships among taxa as inferred from critical evidence and have associated optimality values. That is, they participate in hypothesis testing through the optimality values and are directly supported by evidence. Although they are most clearly exemplified by most parsimonious and maximum likelihood trees, they include the results of minimum-evolution, including neighbor joining (Saitou and Nei 1987) and minimum percent-standard-deviation (Fitch and Margoliash 1967). In each of these examples, there is a direct and unequivocal relationship between a body of evidence and the tree or cladogram.

These are contrasted by what we refer to as tree-shaped objects, which are branching diagrams that look like and are often interpreted as if they were cladograms or trees, but have altogether different purposes and empirical bases. These primarily include diagrams meant to summarize the results of explicit phylogenetic analysis, most notably the strict consensus, MrBayes trees (Huelsenbeck and Ronquist 2003), parsimony jackknife trees (Farris et al. 1996), and supertrees (for example, Binindа-Emonds et al. 2002). The strict consensus representation is objectively interpretable as a summary of the clades that are unambiguously supported by the evidence, and as such is a valuable tool in systematics. MrBayes and parsimony jackknife trees are majority rule summaries of the frequency of clades in a sample of trees. Supertrees are intended to depict the results of analyses involving overlapping sets of terminals.

None of these different kinds of summaries maximizes an underlying optimality criterion. It is widely recognized that the strict consensus representation is a suboptimal explanation of the evidence—that is, it requires more steps than any one of the fundamental cladograms (for example, Kluge 1989), that supported groups can have a lower resampling frequency than unsupported groups (Goloboff et al. 2003a), and that these depictions should not be interpreted as cladograms or trees. However, it is frequently overlooked that the MrBayes solution suffers from the same analytical problem as the jackknife. That is, the Monte Carlo Markov Chain algorithm generates a sample of cladograms and is meant to estimate the posterior probabilities of clades, not trees. As a

majority-rule representation, the MrBayes topology may not be the maximum posterior probability cladogram, and the groups it shows as being recovered in high frequency can actually be unsupported by the Bayesian optimality criterion.

With supertrees, the gulf between evidence and summary depictions increases even more. Because they are intended to summarize the results of different analyses, it is possible for the source trees to have been inferred under contradictory optimality criteria—indeed, this is actually seen by some authors as a strength of supertrees. Of course, that alone is not necessarily an inappropriate procedure. However, at least some proponents of supertrees encourage their interpretation as actual hypotheses of phylogeny— that is, as cladograms or trees equipped with optimality values, branch lengths, and support metrics adequate to test competing evolutionary scenarios (for example, Bininda-Emonds et al. 2002). Such interpretations are indefensible, as no scientific test is possible in the absence of a clear link to empirical evidence— that is, an evidence-based optimality criterion. The perceived need for supertree methods is that

- the amount of systematic data available outstrips present computational abilities
- the obstacles to combining evidence from different sources cannot be overcome.

A recurring theme in this book is that data set size affects only the exhaustiveness of analysis, not the ability to analyze it in a logically consistent manner. Likewise, all veritable evidence can be combined in simultaneous (that is, supermatrix) analysis.

Terms and Notation

Navigating efficiently through descriptions of cladograms and their manipulations requires terminology. As with the cladogram/tree discussion, there is a rich lexicon of descriptive terms, illustrated in Figure 2.1 on page 17.

OTU (operational taxonomic unit) Taxon selected for phylogenetic study, generally based on observed or inferred characteristics; terminal taxon; leaf.

HTU (hypothetical taxonomic unit) The inferred internal vertices or nodes of a cladogram. These are often interpreted as ancestors, but are, in reality, abstractions whose features are constructed to maximize the optimality criterion.

Node A vertex on a cladogram. All HTUs and OTUs are nodes.

Branch The path connecting two nodes; edge.

Root The basalmost, ur-node of a cladogram.

Binary Describes a fully dichotomous, resolved node or cladogram.

Polytomy An internal node that is not binary.

Cost The value of the optimality criterion of a cladogram. This equals the number of steps (transformations, changes) or length in equally weighted parsimony analysis or weighted sum of transformations more generally. In maximum likelihood analysis, cladogram cost is reported as the absolute value of the log likelihood score.

Synapomorphy A transformation on a branch that leads to an HTU.

Autapomorphy A transformation on a branch that leads to an OTU.

Homoplasy Nonminimal transformations of a character on a cladogram.

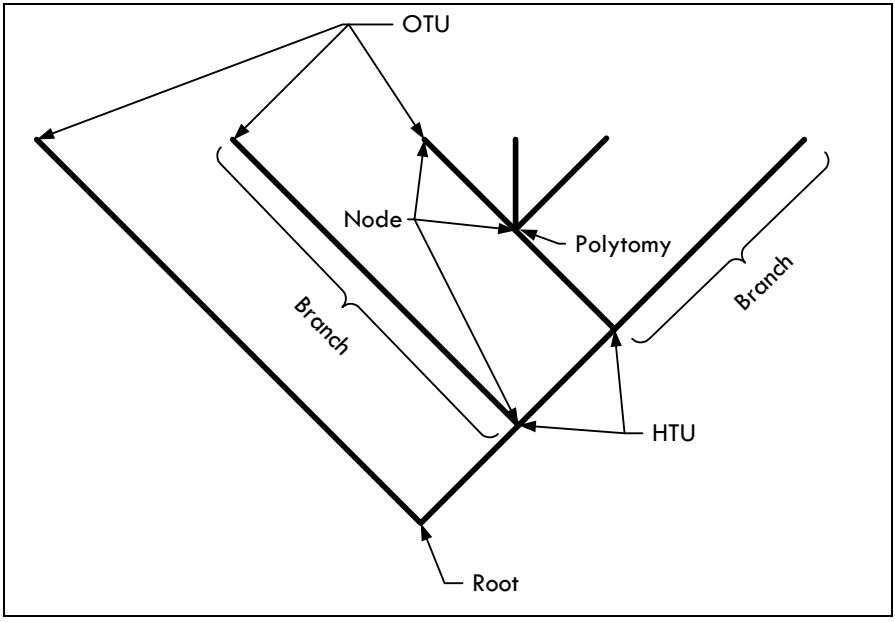

Figure 2.1: Cladogram terms and notation.

Suggested Reading

Farris, J. S., A. G. Kluge, and M. J. Eckhardt. 1970. A numerical approach to phylogenetic systematics. Systematic Zoology 19: 172–189.

Felsenstein, J. 1978. The number of evolutionary trees. Systematic Zoology 27: 27–33.

Goloboff, P. A. 2005. Minority rule supertrees? MRP, compatibility, and minimum flip may display the least frequent groups. Cladistics 21: 282–294.

Platnick, N. I. 1977. Cladograms, phylogenetic trees, and hypothesis testing. Systematic Zoology 26: 438–442.

Prendini, L. 2001. Species or supraspecific taxa as terminals in cladistic analysis? Groundplans versus exemplars revisited. Systematic Biology 50: 290–300.